Understanding the Role of Mental Models in User Interaction with an Adaptive Dialog Agent

Lindsey Vanderlyn **Dirk Väth** **Ngoc Thang Vu**
University of Stuttgart
Institute of Natural Language Processing
vanderly@ims.uni-stuttgart.de

Abstract

Mental models play an important role in whether user interactions with intelligent systems, such as dialog agents, are successful. Adaptive dialog systems present the opportunity to align a dialog agent's behavior with heterogeneous user expectations. However, there has been little research into what mental models users form when interacting with a task-oriented dialog system, how these models affect users' interactions, or what role system adaptation can play in this process. This can make it challenging to avoid damage to human-AI partnership. In this work, we collect a new publicly available dataset for exploring user mental models of information seeking dialog systems. We demonstrate that users have a variety of conflicting mental models about such systems, the validity of which directly impacts the success and perception of their interactions. Furthermore, we show that adapting a dialog agent's behavior to better align with users' mental models, even when done implicitly, can improve dialog efficiency, success, and user perception of the interaction. This shows that implicit adaptation can be beneficial for task-oriented dialog systems, so long as developers understand the mental models of their users.

1 Introduction

Adapting dialog agents' behavior to users has long been an area of interest in dialog research. To this end, researchers have explored various strategies, e.g., language style (Ma et al., 2020), sense of humor (Ritschel and André, 2018), recommendations (He et al., 2023), etc. The basic assumption underlying this research is that such changes to a dialog agent's behavior will lead to more successful or enjoyable interactions with users. However, this is not guaranteed. Bansal et al. (2019b) and Kim and Lim (2019) experimented with human-AI and human-dialog system interaction respectively, and found that adaptation can actually damage human-agent partnership when done in a way which does not align with users' expectations. Therefore, in order to develop successful adaptive dialog agents, it is important to first have a good understanding of what these user expectations are.

One way of understanding these is to measure user mental models. Mental models refer to a person's cognitive representation for how or why they believe a complex system (e.g., a dialog system) works (Johnson-Laird, 1980; Halasz and Moran, 1983; Norman, 2014). Rather than trying to process all details of such systems at once, users will create a simplified representation of the system in their mind (Clements, 2004) which they can use to describe, explain, and predict that system. In the case of dialog systems, these models can be arbitrarily simple, e.g., *"the dialog system can recognize keywords"*, or complex, e.g., *"I would expect the chatbot to be able to answer simple questions, where they can retrieve the answers from my account information"* – depending on the user. These expectations and abstractions are shaped both through interaction with the system and through previous experience (Cho, 2018; Rutjes et al., 2019). Research in human-computer interaction and human-centered AI has found that users' mental models of a system play an important role in predicting how they will interact with it (Kim and Lim, 2019) and that accurate mental models, or those which better match the true capabilities of the system they are interacting with, generally lead to more successful interactions (Kulesza et al., 2012; Bansal et al., 2019a; Grimes et al., 2021).

Previous research into mental models of task-oriented dialog systems generally focuses either on single turn agents, like personal assistants (Luger and Sellen, 2016; Cho, 2018; Tenhundfeld et al., 2022), or on collaborative game settings (Gero et al., 2020; Weitz et al., 2021). However, these domains only represent a small subset of dialog interactions. Additionally, they do not consider the potential of adaptation to influence such interac-

Findings of the Association for Computational Linguistics: NAACL 2025, pages 989–1015

tions. While there have been a limited number of studies looking at mental models in other domains (Wang et al., 2021; Brachman et al., 2023), to our knowledge, the only research exploring users' mental models around adaptive dialog agents was performed by Kim and Lim (2019). In their work, the researchers focused specifically on the scenario of users actively trying to teach an adaptive agent, with the assumption that an implicitly adaptive agent could be poorly accepted by users. However, as such a teacher role imposes an additional cognitive load on users, the goal of this paper is to explore how users perceive and react to implicitly adaptive, task-oriented dialog agents. Concretely, we investigate the following research questions:

RQ1: What role do users' mental models play in task-oriented dialog?

- **RQ1.1:** What mental models do users have about task-oriented dialog systems prior to interaction?
- **RQ1.2:** How do these prior models affect interaction with dialog systems?

RQ2: What role does adaptation have on user mental models and interaction?

- **RQ2.1:** How does interaction with an implicitly adaptive agent change user mental models compared to non-adaptive baselines?
- **RQ2.2:** How does implicit adaptation affect success, trust, and usability?

To answer these questions, we implement three different types of task-oriented dialog system in the domain of business travel. We then recruit 66 participants to take part in a user evaluation. Each user interacts with either an implicitly adaptive dialog agent based on (Väth et al., 2024), or one of two non-adaptive baselines, representing both extremes of the adaptive agent's behavioral spectrum. We probe users for their mental models before and after they interact with the dialog system, analyzing how these affected their interaction, how their mental models were updated through the interaction, and how successful each interaction was.

Our main contributions are: 1) Analyzing what mental models users have about information-seeking dialog systems and demonstrating these to be contradictory. 2) Identifying aspects of dialog interaction which were significantly impacted by user mental models across three types of dialog agent, highlighting the need for adaptive systems. 3) Demonstrating that implicit adaptation can be done without significantly impacting mental models. 4) Showing that implicit adaptation in line with users' mental models significantly increases dialog success and user perception compared to static baselines. 5) Creating a new, publicly available dataset[1] for studying mental models consisting of collected dialogs augmented with self-reported mental models (pre- & post interaction) and logs of each user's interactions.

2 Related Work

2.1 Adaptive Dialog Systems

Research into adaptive dialog systems aims to align their behavior to users in order to improve the interaction experience. Adaptation can occur either on the text-level or with respect to the agent's underlying behavior. Text-level adaptation includes changing the an agent's linguistic style, e.g., adjusting the chatbot's utterances to match a user's emotional state (Ma et al., 2020), personality (Yang et al., 2018; Firdaus et al., 2023), or adapting the complexity of language to a user's domain familiarity (Janarthanam and Lemon, 2014). Behavioral adaptation may rely on additional social cues, e.g., laughter (Ritschel and André, 2018), may require users to intentionally fine-tune the dialog system's behavior (Chen and Pu, 2012; Narducci et al., 2018), or may implicitly intuit cues for adaptation from the users' behavior (Väth et al., 2023).

However, trying to implicitly adapt an intelligent agent's behavior without a good understanding of users' mental models can lead to a mismatch in user expectations and lead to less successful human-agent interactions (Weld et al., 2003; Bansal et al., 2019b) and confusion (Zimmerman et al., 2007). At the same time, asking users to directly provide feedback during the interaction interrupts the experience, which can make the dialog less natural. Furthermore, asking for such feedback places additional cognitive load on the user, which is contrary to the goal of intelligent systems, namely to reduce a user's cognitive load (Höök, 2000).

2.2 Mental Models

Mental models are simplified abstractions humans use to describe, explain and predict systems which are too complex to understand all details of (Johnson-Laird, 1980, 1983). These models are shaped both through interactions with the system and previous experiences with similar systems (Cho, 2018; Rutjes et al., 2019). For example, it

[1] GPL-3.0: www.removed-for-anonymity.github.com

is unrealistic that users will know all details of a conversational AI agent. Rather, users will form simplified models about an agent, which they can use to interact with it, updating these models, e.g., when they encounter unexpected behavior (Einhorn and Hogarth, 1986; Mueller et al., 2019).

Understanding user mental models is important as these influence how users will interact with a given system (Kim and Lim, 2019). Representations of user mental models can take different forms depending on the methods used to elicit them, e.g., qualitative or quantitative (Andrews et al., 2023). Although there are many ways to elicit mental models (Cooke et al., 2000), we focus here on surveys and structured interviews. In structured interviews, users are asked a fixed set of free-form questions and conclusions are drawn based on trends in their answers. Surveys on the other hand, generally ask quantifiable questions with fixed answers, e.g., Likert scales. These methods can be combined to provide complimentary insights. In both cases, questions must be chosen carefully as the act of asking questions can influence users to update their mental models (Rowe and Cooke, 1995).

2.3 Mental Models of Dialog Agents

With the increasing popularity of dialog agents, there has been a corresponding need to understand user's mental models, and how they influence interactions with such systems. To date, the bulk of this research has focused on either single-turn, virtual assistants (Luger and Sellen, 2016; Zamora, 2017; Cho, 2018; Tenhundfeld et al., 2022) or user interactions in cooperative games (Gero et al., 2020; Weitz et al., 2021; Zhu and Villareale, 2021). These studies have provided valuable insights into, e.g., the breadth of mental models users form about such systems (Cho, 2018) and how previous experience can shape user expectations and outcomes (Weitz et al., 2021). However, they represent only a small subset of types of dialog systems (Zhang et al., 2020; Valizadeh and Parde, 2022) and do not consider the role of adaptation.

Wang et al. (2021) in contrast, chose to explore mental models around how students interacted with a virtual teaching assistant over a period of two months. During this study, they found that it is possible to link linguistic information from students' inputs with their self-reported mental models over the course of the interaction. This finding supports our choice of adaptive agent, which infers a user's preferred interaction style from their input text.

To the best of our knowledge, Kim and Lim (2019) are the only study which explores user mental models around an adaptive dialog agent. They investigated a co-adapting agent, where the user explicitly tried to teach new behavior to a simulated personal assistant in a Wizard-of-Oz study. They found that users formed one of two main mental models of the agent, either viewing it on a technical or a social level, which greatly influenced how/what they attempted to teach it. However, the mental models probed were intrinsically linked to the user's role as teacher. We seek to expand on this research and explore user mental models in a more standard adaptive setting, where the dialog system implicitly updates its behavior without requiring users to take on an additional co-adapting role.

3 Dialog System Implementations

We choose an information seeking setting, as this represents a common use case of dialog agents. Furthermore, we choose the domain of business travel reimbursement as it is currently difficult to implement LLMs for this type of sensitive domain, due to the financial risks of hallucination.

In information search, users often have either one of two main strategies for finding information: either navigating step by step to maintain control over context or jumping directly to an answer. These strategies are referred to as "orienteering" and "teleporting" (Teevan et al., 2004). Based on this, we choose an adaptive agent identify the users information search strategy expectation from their initial input and interpolate between these two behaviors. For the non-adaptive baselines, we use a handcrafted dialog system, which outputs all available information (a pure orienteering approach), and an FAQ system, which outputs only a single answer (a pure teleporting approach). In addition to representing the extreme ends of the adaptive agent's behavior, these baselines also represent two of common types of dialog agents employed in sensitive domains (Kearns et al., 2019; Abd-alrazaq et al., 2019; Liao et al., 2020). All agents are in the domain of business travel using the REIMBURSE-En dataset (Väth et al., 2024).

3.1 Adaptive Dialog System

For our adaptive agent (ADA), we use a state-of-the-art LLM-based, controllable dialog architecture proposed by Väth et al. (2024) with gpt-4o-mini. For ADA, subject-area experts create a dialog graph which defines the possible system output texts, avoiding hallucination. In this graph, nodes represent system outputs and edges the set of possible user intents at that node. Given the initial user utterance, the agent uses an LLM to classify whether the user expects a specific answer to a question, or if they expect more guidance to narrow down their information need. For specific questions, the agent utilizes reasoning abilities of the LLM to predict target nodes in the graph, and uses efficient graph navigation algorithms to reach them, only asking follow-up questions needed to clarify the user's information need. For users that need more guidance, the agent will guide them node-by-node through different topics in the graph, using the LLM to determine their next intent at each node. Examples of different agent behavior can be seen in Figure 1.

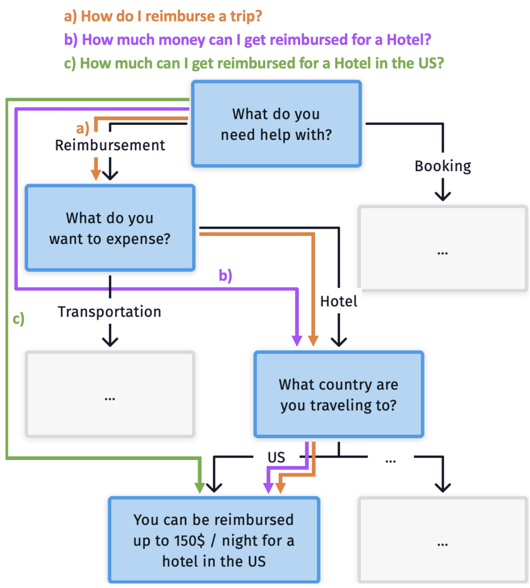


Figure 1: Example of the types of dialog interactions supported by the adaptive agent (Väth et al., 2023). Interaction a) is also the same as the HDC agent, Interaction c) is the same as for the FAQ agent.

This approach utilizes the language understanding abilities of LLMs, while at the same time allowing for stricter control of the output, enabling us to vary only the amount of information/guidance the user is given without changing the text of the outputted nodes. In this way, we can be certain that changes in behavior and perception are a result of the amount of information/guidance the user is given, rather than differences in the generated output. For more details on the agent, see Appendix A. Our code and data a publicly available[1].

3.2 Baseline: Handcrafted Dialog System

The handcrafted dialog agent (HDC) uses the same dialog tree as the adaptive agent. In contrast to the adaptive agent, however, the HDC agent cannot adapt the amount of information/guidance it gives to the user. Instead, it always outputs all nodes along the current path. If the user is asked a question, their response is then compared to the set of pre-defined intents associated with that node using a state-of-the-art similarity model (Reimers and Gurevych, 2019). The agent then moves to the neighboring node connected to the most similar answer and outputs its content to the user. This can be seen as path A in Figure 1.

3.3 Baseline: FAQ System

The FAQ system, again uses the same dialog graph, however in contrast to the other two agents, the FAQ agent only supports single turn interactions. For this agent, we embed the initial user utterance and all node texts using the same similarity model as in the HDC agent (Reimers and Gurevych, 2019). The FAQ agent then directly returns the node text most similar to the user's question. This can be seen as path B in Figure 1.

4 Dataset: RDMM

We collect REIMBURSE Dialog Mental Models (*RDMM*), a new dialog corpus consisting of real user interaction logs with three types of dialog system, surveys of each user's mental model of a chatbot before and after the interaction, and content analysis annotations (subsection 7.1). All three systems were implemented with the REIMBURSE-En dialog graph (Väth et al., 2024), for the domain of business travel. Details on the graph can be found in Appendix B. Our new dataset is publicly available[1].

For our corpus, we collect 180 dialogs from humans interacting with either the HDC baseline, the FAQ baseline, or the adaptive agent. Each dialog covers one of 14 different user information goals taken from the dialog graph. Below is an example dialog between a user and the adaptive agent:

GOAL: You want to know how the reimbursement process works for a research semester. You plan to bring your family with you.

USER: my family are coming with me, how do i get reimbursed for research semester

	FAQ	HDC	Adaptive
# Dialogs	62	56	62
# Successful dialogs	35	26	50
Avg. # Turns/dialog	2.26	15.03	10.61
Avg. # Words in Initial Input	10.21	9.78	11.81
Avg. # Words / Utterance	10.29	5.76	8.96

Table 1: Corpus statistics for collected dialogs

SYSTEM: How long will your research semester last?
USER: 1 month
SYSTEM: Your business trip must be planned as though you were traveling alone (without your family). Provide receipts [...]

As the dialog interaction style between agents was quite different, dialogs with each system varied, e.g., in terms of dialog length and user utterance length. Dialog statistics are shown in Table 1.

Each dialog in the corpus is also labeled with 1) the type of dialog system, 2) the information goal, 3) dialog length, 4) the end condition (success or failure), and 5) subjective user ratings for dialog length and quality of answer.

We further provide annotations of 1) each user's mental model of a dialog system before the interaction, 2) their mental model of the system after the interaction, and 3) their usability and trust ratings after the interaction. Additional dialog examples can be seen in Appendix D and examples of mental model annotations can be seen in Table 2.

5 User Study

In this section, we provide an overview of our experimental objectives, describe the study design, and describe the study participants.

6 Overview of Experimental Objectives

Our research objectives are divided into two main questions, as can be seen in Figure 2: 1) What is the role of mental models in information seeking dialogs? And 2) What effects does adaptation have compared to static dialog agents.

To explore the first research questions, we asked participants about their mental models prior to the interaction. We then asked them to interact with one of three different types of dialog agents and observe differences in how users with different mental models interact with each type of system.

To investigate the second research questions, we asked users to fill out a survey after the dialog interaction providing information about their mental model of the agent they interacted with. We compare the responses to users' pre-interaction mental models and look to see what differences there are between the different types of dialog agents. Finally we look at dialog interaction metrics and user experience metrics to see what differences exist between the different types of dialog agents.

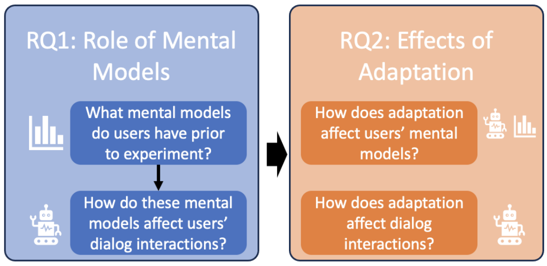


Figure 2: Overview of research objectives and study design. The graph symbol represents questions answered via a survey, the robot symbol represents questions answered after interacting with a dialog agent.

6.1 Study Design

We chose a between-subject design, as we did not want to influence users' mental models by exposing them to multiple dialog systems. Participants were asked to complete a pre-survey, providing their domain familiarity and what mental models they had of dialog systems in general prior to the interaction. They were then randomly assigned to interact with one of the dialog agents over the course of three dialogs. Participants were given no instruction on how they should interact with the dialog system except that they should find an answer to a randomly assigned goal.

Participants were first assigned an "open" goal, representing general domain topics. For example: *"you want to find out information on how to book a business trip."* We chose these goals to help familiarize participants with the domain.

In the second dialog, participants received an "easy" goal. These were more specific information needs, but did not require information about a user's exact situation to be answered, e.g., *"You want to know if you can get reimbursed for a taxi"*.

For the last dialog, participants were assigned a "hard" goal, which required personalized details about the user's planned trip in order to answer. *"You want to know how much money you can get reimbursed for accommodation on your trip to France. You plan to stay with your brother."*

Finally, users took a post-survey, providing an updated mental model of their assigned dialog system, and their impressions of the interaction.

6.2 Participants

Pilot Study To validate our experimental design, we recruited 9 participants between the ages of 20 to 49 from the crowd-sourcing website Prolific. Participants were paid at a rate of 10.50 £/hr, in line with minimum wage. Based on user feedback, we were able to fix technical errors in the study implementation and verify that the time needed by participants was in line with our estimates.

Main Study After performing power analysis, we recruited 66 English speaking crowdsource workers using the same payment rate. Three participants were removed due to technical errors, resulting in a total of 63 participants across the three groups (Adaptive: 22, FAQ: 21, HDC: 20), and 180 dialogs.

Of the participants, 24 were male, 39 were female. Their ages ranged from 20 to 69. On average, participants had some familiarity with dialog systems (3.6 on a 5-point Likert item) and some familiarity with business travel (2.4 on a 5-point Likert item), indicating that the majority of participants had been on one or more business trips. There were no statistically significant differences in the distributions of gender, age, or previous experience with travel between the three conditions.

7 Evaluation Methods

To understand their backgrounds, we asked participants to provide information about their age, gender, experience with dialog systems, and experience with business travel. All surveys can be found in Appendix G.

7.1 Mental Models

Here, we seek to understand what mental models users have about information seeking dialog agents. That is, what expectations users have for how they can interact with such an agent, what the agent can answer, how it can answer, etc. As it is inherently difficult to measure mental models without also influencing users to change them (Rowe and Cooke, 1995), we take two complementary approaches.

The first approach is a series of open-ended questions acting as a stand-in for think-aloud questions one would ask during a laboratory study (Friedman et al., 2018). We asked users about their expectations for both what type of input a information-seeking dialog system can understand and what type of answers it can generate, e.g., *"How would you phrase your input to the chatbot? Is this similar or different to how you would use a search engine or ask a real person?"*.

The second approach was a series of Likert scale items asking users to rate how much they agreed with eight statements on a five-point scale. The first four statements related to their expectations for what type of input a dialog agent could understand and the second four to their expectations for what type of responses they could receive from a dialog system. E.g., *"In general, I think that a chatbot can only give high-level/general answers to questions"*.

Each of the free response questions was analyzed using the standard content analysis technique Hsieh and Shannon (2005). Following this technique, a small subset of user responses are analyzed to identify common trends. These trends are then turned into labels and used to annotate the full corpus of responses. In cases where annotators reached different conclusions, these instances are discussed until the annotators can reach a consensus. In this way, we can quantify what beliefs/expectations, i.e., mental models, users have in a more granular way than is possible through only Likert feedback.

7.2 Dialog Interaction Metrics

To understand the dialog interaction, we look at objective and subjective dialog success and dialog length. We evaluate objective success by checking if the user received the correct answer to their question. For objective length, we count the total number of user and system turns in a dialog. To measure perceived length, we ask participants to answer a single-item Likert scale after each dialog from 1 (much too short) to 5 (much too long). Similarly, we ask for a perceived success rating per dialog with a single-item Likert scale from 1 (question not at all answered) to 4 (question completely answered).

7.3 User Experience Metrics

After the interaction, we ask users to fill out a second survey based on the retrospective technique proposed by Hoffman et al. (2018), asking them what aspects of the experience they found positive or negative. We again process these answers using content analysis.

To measure the perceived trust and reliability, we use the reliability and trust subscales from the Trust in Automation (TiA) questionnaire (Körber, 2018). To measure perceived usability, we use the four item Universal Measure of User Experience (UMUX) (Finstad, 2010). Both are validated questionnaires, which have been found to correlate well with the respective underlying constructs.

8 Results & Discussion

In the following section, we evaluate users as having a certain mental model if they rated that expectation with a value of 4 (agree) or higher in the pre-/post-survey.

8.1 RQ1.1: Mental Models

Prior to the interaction, we investigate what expectations/ beliefs users had of task-oriented agents in general, both quantitatively and qualitatively.

Quantitative Analysis In Figure 3, we see user expectations for what type of input a dialog system can understand and for how a dialog agent can interact with them. To understand how users thought they could interact with a dialog system, we asked them whether they thought they could 1) use natural language, 2) only use keywords, 3) only get an answer if they put in a specific question, and 4) ask a general question which the system would then help them clarify. To understand what expectations users had about a dialog agent's capabilities, we asked them whether they thought the system could provide 1) only general answers, 2) answers specific to a user's situation, 3) an immediate answer, and 4) an answer only after asking multiple follow-up questions.

The results show that *there is no one expectation shared between all users*, neither for how the user should formulate their input nor for how the system should behave. Even in cases where most users expected a certain behavior, there was still a sizable minority who did not.

Qualitative Analysis To get a more granular understanding of users' expectations, we also performed qualitative analysis, looking for trends in user free-response answers. Most of these common expectations fell into one of five categories (see Table 2): expectations about 1) the style of their input text, 2) the content of their input text 3) the style of the system's answer text 4) the content of the system's answer text, and 5) about the general dialog interaction.

However, within these categories, we see the same trend as in the quantitative analysis: *users*

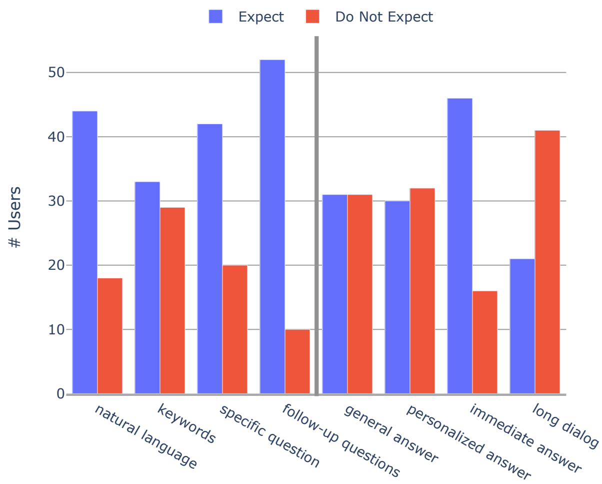


Figure 3: Distribution of mental models for what type of input a dialog system can understand (left 4 columns) and how it can respond (right 4 columns).

have very different, and in many cases contradictory, expectations for how they should interact with a dialog agent, and how the agent, in turn, should behave. The contradictory nature of these mental models further supports the need for reliable adaptive dialog systems.

8.2 RQ1.2: Effect of Mental Models on Interaction

In this section, we investigate how users' initial mental models influence their interactions with different types of dialog systems. We compare differences in objective dialog success, as well as subjective dialog length and subjective success, investigating how each metric was influenced by the presence/absence of a mental model. In this section, we use the quantitative representations of users' mental models shown in Figure 3.

For each expectation, Table 3 lists the dialog systems for which there was a significant difference ($p<0.05$) in a given metric between users who had that mental model (pre-interaction) and those who did not. We used Barnard exact tests to test significance of objective dialog success and Wilcoxon-Mann-Whitney tests for the subjective measures.

We find that nearly all user expectations had an effect on either objective success or on how users perceived the interaction. Which mental models had which effects, largely depended on the style of the dialog agent[2]. In particular, we find that the greatest impacts of mental models occurred in cases where a mismatch existed between user

[2]Tables for the exact differences in metrics for each dialog system can be seen in Appendix C

Table 2: Main and sub-categories resulting from content analysis. For every sub-category (highlighted in bold), an example of participants' free-form feedback is given. Every example response is from a different participant.

A1	I would phrase it much like asking a real person (**like a person**)
A2	I would be more specific and direct with a chatbox than I would with a real person (**precise language**)
A3	I would use key terms, not necessarily full sentences. (**keywords**)
A4	I would use the chatbot like a search engine (**like a search engine**)
B1	If the question is simple and commonly asked (**simple Questions**)
C1	I would expect to get very general information from a chatbot that could be found on a companies website. (**generic information**)
C2	They should be able to provide prices, they should be able to provide the best options for you as a customer. (**Personalized**)
C3	and 100% correct in the information I am being given. (**correct/accurate**)
C4	Technically correct but perhaps not for the context - I would expect to check facts. (**questionable accuracy**)
D1	I would expect the level of information to be detailed and to a high level of knowledge (**highly detailed**)
D2	I would expect a clear, precise answer (**concise**)
D3	with sources I can verify myself if I doubt the accuracy of the information. (**with sources**)
D4	I would expect to get a similar style to what I might get from a real person nowadays. (**casual/friendly**)
D5	I would expect a formal answer (**stilted/formal**)
E1	to save me time from having to browse through all the terms, rules and contracts myself. (**fast interaction**)
E2	and that it would be correct as long as I used the correct terms to ask. (**quality dependent on question**)
E3	Unable to always answer my question. Sometimes go round in circles trying to get the information required (**unreliable**)

A = User input style, **B** = User input content, **C** = Dialog agent output content, **D** = Dialog agent output style, **E** = Interaction

Mental models	Success	Subj. Length	Subj. Success
Natural Language	Adaptive		
Only Keywords			
Only Specific Questions			
Vague Questions + Clarification			HDC
Only General Answers	Adaptive		
Personalized Answers	FAQ	FAQ	HDC
Single-Turn Interaction			
Multi-Turn Interaction		Adaptive, FAQ	

Table 3: Significant effects ($p < 0.05$) of pre-interaction mental models on objective and subjective interaction metrics per dialog system. As no mental models affected objective length, this column was omitted.

Mental models	Adaptive	FAQ	HDC
Natural Language	-0.14	0.05	**-1.15****
Only Keywords	-0.33	-0.05	0.60
Only Specific Questions	0.00	-0.29	0.00
Vague Questions + Clarification	-0.57	**-2.19*****	0.10
Only General Answers	-0.14	**1.05****	-0.10
Personalized Answers	-0.38	-0.57	-0.55
Single-Turn Interaction	-0.38	0.38	-0.65
Multi-Turn Interaction	0.71	**-1.48****	0.60

Table 4: Change in mental models before and after the interaction for each system. $*p < 0.05$; $**p < 0.01$; $***p < 0.001$.

expectations and dialog system behavior. For example, the FAQ system can only output a single general answer. Users of the FAQ system who expected personalized answers were less successful and perceived the dialogs as much shorter compared to users who did not. On the other hand, the HDC system cannot give a direct answer, but will continue asking follow-up questions until a personalized answer is reached. Users of the HDC system who expected a personalized answer or to be able to ask a vague question and have help from the system to clarify their intent, generally viewed their questions as more completely answered than those that did not.

These results are line with work by Kulesza et al. (2012) and Bansal et al. (2019a), who also found that user-AI partnerships are more successful when user-mental models match system capabilities.

8.3 RQ2.1: Effect of Adaptation on Mental Models

When comparing users' mental models before and after interaction (see Table 4), we find no significant changes for users of the adaptive system. In contrast, both HDC and FAQ users experienced significant changes in their mental models of the dialog agent after the interaction. FAQ users found that, in contrast to their initial belief, the dialog agent they interacted with could not handle vague input nor could it support multi-turn interactions or give more than general answers. HDC users needed to revise their belief that the dialog agent they interacted with could handle natural language input. As no such changes took place for users of the adaptive system, this suggests *the adaptive agent is able to implicitly adapt in a way that remains in line with user expectations*, using only the users' input to infer their expectations.

8.4 RQ2.2: Effect of Adaptation on Usability, Trust, & Task Success

In Table 5, we see that the adaptive agent is significantly more successful than either the FAQ or HDC system (p<0.05; Barnard Exact test), while at the same time, requiring significantly (p<0.05; t-test) fewer turns than the HDC policy. Compared to the FAQ system, where dialogs were rated as too short, the adaptive system was subjectively perceived to have a more appropriate dialog length ($p<0.05$; Wilcoxon-Mann-Whitney test).

When taken together, these results suggest that *adapting to the user's interaction style has a positive effect*, not only increasing task success, but also reaching a middle-ground between too short single-turn FAQ and too long handcrafted dialogs. This suggests that the agent was able to successfully adapt to both users who expected a short FAQ/style interaction and users who expected more comprehensive guidance, which resulted in more users being able to reach their goal.

Model	# Turns	Success	Perceived Length	Answer Satisfaction
Adaptive	10.61	**80.65**	**2.89**	**2.94**
FAQ	**2.26**	57.38	2.28	2.61
HDC	15.05	46.43	2.73	2.86

Table 5: Average objective and subjective performance metrics per dialog system. Perceived length is measured on a 5-point scale, where 3 is most appropriate; perceived quality on a 4-point scale.

Model	Trust	Reliability	Usability
Adaptive	**3.38**	**3.2**	**66.67**
FAQ	2.83	2.79	57.73
HDC	3.15	3.06	58.43

Table 6: Averaged trust, reliability, and usability ratings.

When looking at the user trust, reliability, and usability ratings in Table 6, we see the adaptive agent is rated highest in all categories, although these differences are not significant. In particular, the usability results suggest that the burden of answering follow-up questions from the adaptive agent – and thus extending the dialog length– is no higher for users than that of the single-turn dialog scenario in the FAQ setting.

When taken together we find that, in contrast to previous work (Kim and Lim, 2019), *an explicit training period is not necessarily required prior to adaptation, as long as the implicit adaptation is in line with user mental models*.

9 Conclusion

In this work, we investigate what mental models users have about information seeking dialog systems, how these mental models impact interactions with (adaptive) dialog systems, how these mental models are updated through such interaction, and how adaptation affects success and user perceptions. Additionally, we release a novel corpus of dialogs, self-reported mental models, and content analysis annotations, which can be used for studying mental models and their implications for interaction with task-oriented dialog agents.

We find that users form a variety of contradictory mental models about how they can interact with an information seeking dialog agent and about the system's behavior. These mental models significantly impacted how users perceived the dialog systems and how successful they were at using them, highlighting the importance of aligning dialog system functionality with user expectations.

Furthermore, we demonstrate that the implicit adaptation carried out by the adaptive dialog agent did not cause significant changes to users' mental models. Rather, the adaptive agent was able to implicitly align its behavior with user expectations based on their initial input. Adapting in this way significantly improved objective evaluation metrics, i.e. dialog success and dialog length, as well as the perceived length, and perceived answer quality.

In conclusion, we find that, in contrast to the results of Kim and Lim (2019), a teaching period may not be required before proactive adaptation can be beneficial, so long as that adaptation is aligned with users' expectations. This highlights the importance of first understanding user mental models either through direct evaluation or implicitly deriving this information, e.g., from user input to the system.

10 Limitations

While we tried to recruit a diverse background of participants, the study was conducted in English with participants living in English-speaking countries, which may bias the results. Additionally, we explore only one possible axis of adaptation, namely how many follow-up questions/ how much additional information should be asked or skipped before delivering users an answer. However, during our analysis of users' mental models before the interaction, we found that there are multiple axes in which users have contradicting expectations, e.g., level of answer detail, linguistic style of dialog system output, etc. In the future it would be interesting to explore how each of these axes affect the dialog interaction, either individually or in combination.

Additionally, the type of adaptation we explore is directly linked to information search strategies and we only explore this through the single domain of business travel. As our results are not directly linked to the content of the domain, but rather the type of domain, we believe that they will generalize to other (complex) information search domains, although perhaps in simpler domains which do not require personalized answers, the effects would be less dramatic. However, in the future, it would be interesting to see what other types of behavioral adaptation might be more appropriate for dialog agents in other types of conversational interaction, e.g., for recommendation domains, argumentation domains, cooperative domains, etc.

11 Ethical Considerations and Risks

Before starting the study, we performed power analysis to ensure that we would have enough power to detect an effect if one were present, so as to not waste the time of any participants who took part in the study. To ensure that users could give informed consent, we provided a detailed description of the task and research objectives both on the crowdsourcing platform and once they had accepted the task. In respect of participant privacy, we specifically did not collect personally identifying data from any users. To this end, we store all logs and survey responses using an anonymous hash generated based on a given username, rather than with the username itself. In this way, users could log in again if they needed to take a break in the middle of the interaction, but we had no way of directly linking any recorded results to, e.g., users' Prolific account identifiers. To ensure that participants were fairly compensated, we followed best practices recommended by the crowdsourcing platform paying users at 9£/hr, which was in-line with minimum wage in the countries we recruited participants from at the time of the experiment. We additionally used our pilot study to verify that our estimated time was below the median time we selected when advertising the task, meaning most participants had a higher hourly wage.

In terms of risks, the goal of this paper is to lay a ground-work for creating more effective adaptive dialog agents. However, this does have the possible risk of creating chatbots which could also be used to more effectively replace human jobs.

References

Alaa A. Abd-alrazaq, Mohannad Alajlani, Ali Abdallah Alalwan, Bridgette M. Bewick, Peter Gardner, and Mowafa Househ. 2019. An overview of the features of chatbots in mental health: A scoping review. *International Journal of Medical Informatics*, 132:103978.

Robert W Andrews, J Mason Lilly, Divya Srivastava, and Karen M Feigh. 2023. The role of shared mental models in human-ai teams: a theoretical review. *Theoretical Issues in Ergonomics Science*, 24(2):129–175.

Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S Lasecki, Daniel S Weld, and Eric Horvitz. 2019a. Beyond accuracy: The role of mental models in human-ai team performance. In *Proceedings of the AAAI conference on human computation and crowdsourcing*, volume 7, pages 2–11.

Gagan Bansal, Besmira Nushi, Ece Kamar, Daniel S Weld, Walter S Lasecki, and Eric Horvitz. 2019b. Updates in human-ai teams: Understanding and addressing the performance/compatibility tradeoff. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 2429–2437.

Michelle Brachman, Qian Pan, Hyo Jin Do, Casey Dugan, Arunima Chaudhary, James M Johnson, Priyanshu Rai, Tathagata Chakraborti, Thomas Gschwind, Jim A Laredo, et al. 2023. Follow the successful herd: Towards explanations for improved use and mental models of natural language systems. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, pages 220–239.

Li Chen and Pearl Pu. 2012. Critiquing-based recommenders: survey and emerging trends. *User Modeling and User-Adapted Interaction*, 22(1):125–150.

Janghee Cho. 2018. Mental models and home virtual assistants (hvas). In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI EA '18, page 1–6, New York, NY, USA. Association for Computing Machinery.

DH Clements. 2004. Perspective on "the child's thought and geometry". *Classics in mathematics education research*, pages 60–66.

Nancy J Cooke, Eduardo Salas, Janis A Cannon-Bowers, and Renee J Stout. 2000. Measuring team knowledge. *Human factors*, 42(1):151–173.

Hillel J Einhorn and Robin M Hogarth. 1986. Judging probable cause. *Psychological bulletin*, 99(1):3.

Kraig Finstad. 2010. The usability metric for user experience. *Interacting with computers*, 22(5):323–327.

Mauajama Firdaus, Arunav Shandilya, Asif Ekbal, and Pushpak Bhattacharyya. 2023. Being polite: Modeling politeness variation in a personalized dialog agent. *IEEE Transactions on Computational Social Systems*, 10(4):1455–1464.

Scott Friedman, Kenneth Forbus, and Bruce Sherin. 2018. Representing, running, and revising mental models: A computational model. *Cognitive Science*, 42(4):1110–1145.

Katy Ilonka Gero, Zahra Ashktorab, Casey Dugan, Qian Pan, James Johnson, Werner Geyer, Maria Ruiz, Sarah Miller, David R. Millen, Murray Campbell, Sadhana Kumaravel, and Wei Zhang. 2020. Mental models of ai agents in a cooperative game setting. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–12, New York, NY, USA. Association for Computing Machinery.

G. Mark Grimes, Ryan M. Schuetzler, and Justin Scott Giboney. 2021. Mental models and expectation violations in conversational ai interactions. *Decision Support Systems*, 144:113515.

Frank G Halasz and Thomas P Moran. 1983. Mental models and problem solving in using a calculator. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pages 212–216.

Ming He, Jiwen Wang, Tianyu Ding, and Tong Shen. 2023. Conversation and recommendation: knowledge-enhanced personalized dialog system. *Knowledge and Information Systems*, 65(1):261–279.

Robert R. Hoffman, Shane T. Mueller, Gary Klein, and Jordan Litman. 2018. Metrics for explainable AI: challenges and prospects. *CoRR*, abs/1812.04608.

Hsiu-Fang Hsieh and Sarah E Shannon. 2005. Three approaches to qualitative content analysis. *Qualitative health research*, 15(9):1277–1288.

K. Höök. 2000. Steps to take before intelligent user interfaces become real. *Interacting with Computers*, 12(4):409–426.

Srinivasan Janarthanam and Oliver Lemon. 2014. Adaptive generation in dialogue systems using dynamic user modeling. *Computational Linguistics*, 40(4):883–920.

Philip N Johnson-Laird. 1980. Mental models in cognitive science. *Cognitive science*, 4(1):71–115.

PN Johnson-Laird. 1983. *Mental models: Towards a cognitive science of language, inference and consciousness*. Harvard University Press.

William R. Kearns, Nai-Ching Chi, Yong K. Choi, Shih-Yin Lin, Hilaire Thompson, and George Demiris. 2019. A Systematic Review of Health Dialog Systems. *Methods of Information in Medicine*, 58(6):179–193.

Da-jung Kim and Youn-kyung Lim. 2019. Co-performing agent: Design for building user-agent partnership in learning and adaptive services. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–14, New York, NY, USA. Association for Computing Machinery.

Moritz Körber. 2018. Theoretical considerations and development of a questionnaire to measure trust in automation. In *Congress of the International Ergonomics Association*, pages 13–30. Springer.

Todd Kulesza, Simone Stumpf, Margaret Burnett, and Irwin Kwan. 2012. Tell me more? the effects of mental model soundness on personalizing an intelligent agent. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, page 1–10, New York, NY, USA. Association for Computing Machinery.

Q. Vera Liao, Werner Geyer, Michael Muller, and Yasaman Khazaen. 2020. Conversational Interfaces for Information Search. In Wai Tat Fu and Herre van Oostendorp, editors, *Understanding and Improving Information Search: A Cognitive Approach*, pages 267–287. Springer International Publishing, Cham.

Ewa Luger and Abigail Sellen. 2016. "like having a really bad pa": The gulf between user expectation and experience of conversational agents. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, page 5286–5297, New York, NY, USA. Association for Computing Machinery.

Yukun Ma, Khanh Linh Nguyen, Frank Z. Xing, and Erik Cambria. 2020. A survey on empathetic dialogue systems. *Information Fusion*, 64:50–70.

Shane T Mueller, Robert R Hoffman, William Clancey, Abigail Emrey, and Gary Klein. 2019. Explanation in human-ai systems: A literature meta-review, synopsis of key ideas and publications, and bibliography for explainable ai. *arXiv preprint arXiv:1902.01876*.

Fedelucio Narducci, Marco de Gemmis, Pasquale Lops, and Giovanni Semeraro. 2018. Improving the user experience with a conversational recommender system. In *International Conference of the Italian Association for Artificial Intelligence*, pages 528–538. Springer.

Donald A Norman. 2014. Some observations on mental models. In *Mental models*, pages 15–22. Psychology Press.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. 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 2019, Hong Kong, China, November 3-7, 2019*, pages 3980–3990. Association for Computational Linguistics.

Hannes Ritschel and Elisabeth André. 2018. Shaping a social robot's humor with natural language generation and socially-aware reinforcement learning. In *Proceedings of the workshop on NLG for human–robot interaction*, pages 12–16.

Anna L Rowe and Nancy J Cooke. 1995. Measuring mental models: Choosing the right tools for the job. *Human resource development quarterly*, 6(3):243–255.

Heleen Rutjes, Martijn Willemsen, and Wijnand IJsselsteijn. 2019. Considerations on explainable ai and users' mental models. In *Where is the Human?*

Bridging the Gap Between AI and HCI, United States. Association for Computing Machinery, Inc.

Jaime Teevan, Christine Alvarado, Mark S. Ackerman, and David R. Karger. 2004. The perfect search engine is not enough: A study of orienteering behavior in directed search. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '04, pages 415–422, New York, NY, USA. Association for Computing Machinery.

Nathan L. Tenhundfeld, Hannah M. Barr, Emily H. O'Hear, and Kristin Weger. 2022. Is my siri the same as your siri? an exploration of users' mental model of virtual personal assistants, implications for trust. *IEEE Transactions on Human-Machine Systems*, 52(3):512–521.

Mina Valizadeh and Natalie Parde. 2022. The AI doctor is in: A survey of task-oriented dialogue systems for healthcare applications. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6638–6660, Dublin, Ireland. Association for Computational Linguistics.

Dirk Väth, Lindsey Vanderlyn, and Ngoc Thang Vu. 2023. Conversational tree search: A new hybrid dialog task. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1264–1280, Dubrovnik, Croatia. Association for Computational Linguistics.

Dirk Väth, Lindsey Vanderlyn, and Ngoc Thang Vu. 2024. Towards a zero-data, controllable, adaptive dialog system. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16433–16449, Torino, Italia. ELRA and ICCL.

Dirk Väth, Lindsey Vanderlyn, and Ngoc Thang Vu. 2024. Towards a zero-data, controllable, adaptive dialog system. *Preprint*, arXiv:2403.17582.

Qiaosi Wang, Koustuv Saha, Eric Gregori, David Joyner, and Ashok Goel. 2021. Towards mutual theory of mind in human-ai interaction: How language reflects what students perceive about a virtual teaching assistant. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA. Association for Computing Machinery.

Katharina Weitz, Lindsey Vanderlyn, Ngoc Thang Vu, and Elisabeth André. 2021. "it's our fault!": Insights into users' understanding and interaction with an explanatory collaborative dialog system. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 1–16, Online. Association for Computational Linguistics.

Daniel S. Weld, Corin Anderson, Pedro Domingos, Oren Etzioni, Krzysztof Gajos, Tessa Lau, and Steve Wolfman. 2003. Automatically personalizing user interfaces. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence*, IJCAI'03, page 1613–1619, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Min Yang, Qiang Qu, Kai Lei, Jia Zhu, Zhou Zhao, Xiaojun Chen, and Joshua Z. Huang. 2018. *Investigating Deep Reinforcement Learning Techniques in Personalized Dialogue Generation*, pages 630–638. SIAM.

Jennifer Zamora. 2017. I'm sorry, dave, i'm afraid i can't do that: Chatbot perception and expectations. In *Proceedings of the 5th international conference on human agent interaction*, pages 253–260.

Zheng Zhang, Ryuichi Takanobu, Qi Zhu, MinLie Huang, and XiaoYan Zhu. 2020. Recent advances and challenges in task-oriented dialog systems. *Science China Technological Sciences*, 63(10):2011–2027.

Jichen Zhu and Jennifer Villareale. 2021. Understanding mental models of ai through player-ai interaction. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems (CHI EA'21)*, page 11.

John Zimmerman, Anthony Tomasic, Isaac Simmons, Ian Hargraves, Ken Mohnkern, Jason Cornwell, and Robert Martin McGuire. 2007. Vio: A mixed-initiative approach to learning and automating procedural update tasks. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '07, page 1445–1454, New York, NY, USA. Association for Computing Machinery.

A Adaptive Dialog Agent Implementation

The architecture for the adaptive agent we use was published under the GPL-3.0 license, making our use of it consistent with its intended use.

For all reported results, we used the `gpt-4o-mini-2024-07-18` model.

A.1 Adaptive Agent Policy

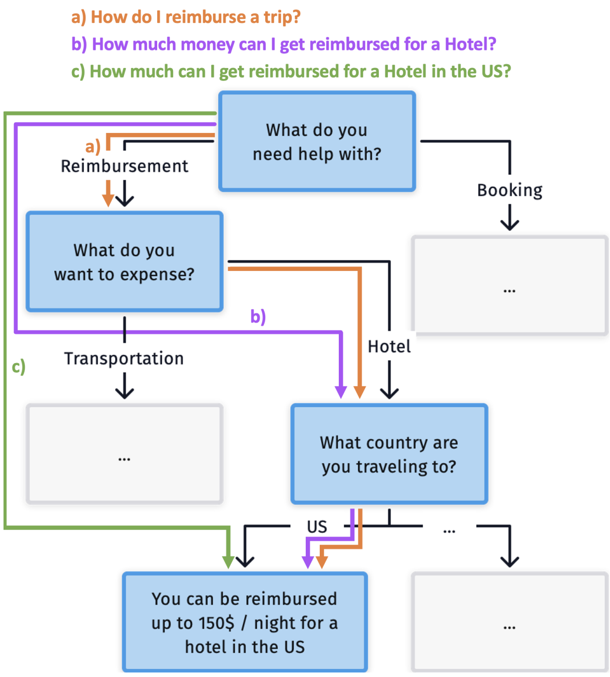


Figure 4: Example of the types of dialog interactions supported by the adaptive agent (Väth et al., 2023). Interaction a) is also the same as the HDC agent, Interaction c) is the same as for the FAQ agent.

The adaptive agent we use is implemented based on work by Väth et al. (2024) who implement an LLM-based agent for the Conversational Tree Search task, a state-of-the-art approach to controllable dialog. The following section is based on their work.

The agent is based on a modular architecture including the following modules: 1) an LLM-based interaction mode classifier, which decides whether the user's initial input is a concrete question or a less grounded input, 2) an LLM-based intent classifier, to classify user intent at every step, 3) a filter module, searching node candidates in the dialog graph that might answer a user's question, 4) a belief state tracker, storing variable values to fill templates, 5) a dialog policy, responsible for walking the dialog graph and choosing to output or skip a node.

A.1.1 Guided Mode

If the user interaction expectation is identified as requiring additional system guidance, the adaptive agent will walk the user node-by-node through the graph, much like the HDC policy. At each step, the agent will use the LLM to identify the user's next intent, matching the user input to the fixed set of intents available at that node. This mode can be seen in interaction style a) in Figure 4.

A.1.2 Question Answering

On the other hand, if the user interaction expectation is identified as wanting a specific answer, the adaptive agent will identify answer candidates (using the filter module) to see if the question can be directly answered (interaction style c) in Figure 4) or if the system should ask follow-up questions to help the user narrow down their information need between the answer candidates (interaction style b) in Figure 4).

Identifying Answer Candidates To identify candidate answers, the agent performs similarity ranking between the user utterance and all nodes in the graph, using an MPNet-based sentence-transformer (Reimers and Gurevych, 2019). The 15 most similar nodes are then passed into the LLM which is tasked with identifying which nodes could actually answer the user's question. Performing the search in a two-step manner reduces the number of input tokens and the search space to be reasoned over, reducing computation time and increasing the accuracy of the results.

Asking Follow-Up Questions To determine which follow-up questions are necessary, the agent looks at the possible answer candidates and identifies the longest shared path that could lead to all answers. From the dialog graph $G = (V, E)$, with nodes V and edges E, the filter module retrieves a list of all possible answer node candidates $V_a \subseteq V$. For each answer node candidate $v_a \in V_a$, the policy calculates all possible paths $P_a := \{(v_1, \ldots, v_n) | v_n = v_a \land \forall v_i, v_{i+1} : (v_i, v_{i+1}) \in E\}$ that lead from the current node $v_1 \in V$ to that answer. The agent then selects the longest shared path prefix which leads to all candidate answer nodes. Shared prefixes are calculated by looking at all paths to the candidate answers: $P_{shared} := \{(v_1, \ldots, v_n) | \forall v_a \in V_a \exists p \in P_a \forall 1 \leq i \leq n : p_i = v_i\}$. Once the longest prefix $argmax_{p \in P_{shared}} |p|$ has been identified, the policy outputs the final node in this sequence to the user. By definition, this node must represent either a necessary clarifying question from the system or is the intended answer to the user's question. If there are still open goals, the user intent is identified based on their next input. The agent then continues to the corresponding neighbor, and a new longest prefix is calculated using the remaining goal nodes.

A.2 Choice of LLM

To choose an LLM, we compared GPT-4o-mini, LLAMA, and Gemma-2 as well as testing a Reinforcement Learning (RL)-based version of the agent which used only a small Language Model (Väth et al., 2024). We selected a GPT-based implementation for this paper as the dialog success rate was much higher than the open-source LLMs, when tested in simulation. Since the goal of this paper is not to test the capabilities of a single agent (or LLM), but rather to explore the role of mental models in adaptive dialog, we only used the GPT-based agent with real users. Although we also tested the the RL-based agent with real users, the study outcomes were nearly identical, therefore we only report the results from the GPT-4o-mini-based agent.

B Dialog Graph Statistics

The dialog graph represents the domain used for the collected dialog corpus. It consists of a dialog tree which defines the possible system utterances as well as a set of collected user utterances (user questions and answers to system follow-up questions). This graph can be used directly to train an adaptive CTS agent.

The translated dialog tree retains the same structure as the original German graph, with the same number of user questions and user follow-up answers. As a note, this translation and its distribution is consistent with the original intent of the REIMBURSE (Väth et al., 2023) dataset, which was published under the GPL-3.0 license as our new corpus is also published under this same license.

A summary of the tree and translated user inputs can be seen in Table 7:

	Train	**Test**
# Nodes	123	123
Tree Depth	32	32
Max. Node Degree	14	14
# User Questions	279	17
Avg. User Questions	3.5	2.2
# Ans. Paraphrases	246	162
Avg. Ans. Paraphrases	3.4	2.2

Table 7: Overview of the *REIMBURSE-EN* dialog graph (numbers rounded to one decimal). These numbers are identical to the original German-language version of the dialog graph.

C Effects of Mental Models on Objective and Subjective Dialog Metrics

We chose to analyze the effects of mental models per dialog system, as each dialog agent had different behavior. In each of the tables below, we report differences in objective and subjective metrics for dialog success and dialog length.

C.1 HDC System

Mental models	Δ Length (turns)	Δ Success (%)	Δ Subj. Length	Δ Subj. Success
Natural Language	N/A	N/A	N/A	N/A
Only Keywords	-0.46	-7.69	-0.22	-0.09
Only Specific Questions	1.26	15.63	0.19	0.33
Vague Questions + Clarification	2.07	-9.67	0.25	**1.02***
Only General Answers	-3.71	17.29	-0.46	-0.10
Personalized Answers	-0.45	3.75	0.15	**0.85***
Single-Turn Interaction	2.67	-10.10	0.23	0.61
Multi-Turn Interaction	1.24	22.50	0.02	0.29

Table 8: Effects of pre-interaction mental models on objective and subjective interaction metrics with the HDC system. * marks significant effects ($p < 0.05$).

For the HDC condition, we do not report differences for the mental model 'Natural Language' as only a single user did not believe they could use natural language to interact with the agent. Although there were no significant effects for success, we believe this could have been due to the sample size of our study, and it would be interesting to see if a larger experiment would also reveal difference here, e.g., with the mental model 'Multi-Turn Interaction'. We do, however, see differences in user perception of how successful the interaction was. Users expecting the system to ask clarifying questions or expecting personalized answers were significantly happier with the answers they received than those that did not.

C.2 FAQ System

Mental models	Δ Length (turns)	Δ Success (%)	Δ Subj. Length	Δ Subj. Success
Natural Language	0.18	-2.14	-0.10	-0.23
Only Keywords	-0.20	1.59	-0.01	-0.12
Only Specific Questions	-0.11	1.26	0.17	0.29
Vague Questions + Clarification	-0.30	8.00	0.05	0.02
Only General Answers	-0.23	4.07	-0.09	-0.08
Personalized Answers	-0.11	**-39.00***	**-0.46***	-0.43
Single-Turn Interaction	0.01	3.73	-0.27	-0.18
Multi-Turn Interaction	-0.02	0.42	**0.56***	0.12

Table 9: Effects of pre-interaction mental models on objective and subjective interaction metrics with the FAQ system. * marks significant effects ($p < 0.05$).

In the FAQ condition, we see that users with the mental model 'Personalized Answers' were significantly less successful than users without it. As the domain is relatively complex, and the FAQ system cannot ask clarifying questions, depending on how much detail users put into their initial utterance, the agent might falsely match to a similar (but incorrect) answer.

When looking at the perceived length, users expecting a personalized answer generally found the dialog system shorter than those who did not. Users expecting a multi-turn interaction, interestingly found dialogs longer than those who did not. We hypothesize that this could be as a result of users conflating length of dialog with the length of system utterances, rather than the number of turns as we initially intended. This could explain why the length was perceived as longer, as some of the system utterances were quite long.

C.3 Adaptive System

For the adaptive condition, we do not report any differences in metrics for the expectation 'Vague Users Questions with System Clarifications' as there were only two users who did not expect this. Here, we see that two user mental models affected the success of the interaction 'Natural Language' and 'Only General

Mental models	**Δ Length (turns)**	**Δ Success (%)**	**Δ Subj. Length**	**Δ Subj. Success**
Natural Language	-2.34	**24.53***	0.10	0.04
Only Keywords	4.69	-14.60	-0.40	0.15
Only Specific Questions	0.86	11.87	-0.24	-0.01
Vague Questions + Clarification	N/A	N/A	N/A	N/A
Only General Answers	4.50	**24.76***	-0.07	0.21
Personalized Answers	0.72	-13.46	0.00	0.15
Single-Turn Interaction	0.08	9.63	0.30	0.36
Multi-Turn Interaction	2.46	-9.63	**-0.86***	-0.25

Table 10: Effects of pre-interaction mental models on objective and subjective interaction metrics with the adaptive system. * marks significant effects ($p < 0.05$).

Answers'. Users who used natural language input were able to benefit from the adaptive agents better reasoning capabilities compared to users who only used keywords and were then walked node-by-node through the dialog graph. As can be seen in Table 5, users of the HDC system were significantly less successful. From analyzing the dialogs, it seems like the complex nature of the domain meant users navigating by keywords often struggled to understand domain-specific vocabulary, such as "per diem" and often incorrectly guessed which general category the information they were looking for could be found. In contrast, users who input even a very broad question using natural language were then only asked targeted follow-up questions by the adaptive agent which helped them more successfully navigate the domain.

Finally, we also see that the expectation 'Multi-turn Interaction' also had a significant effect on how long users perceived the dialog to be, although the actual dialogs did not show any significant differences in length.

D Dialog Corpus Examples

The following show full examples of dialogs from the corpus. Here we show two dialogs for each goal type and from each type of system. We show two dialogs for each chosen goal (each from the same type of dialog system) to demonstrate how different users tried to interact with an information seeking dialog agent.

D.1 Open Goal - Adaptive Agent

D.1.1 Ex. 1

DIALOG: USER: 6d91b06047716a0a0b2c2cd86fdc78 | POLICY: ADA | GOAL-TYPE: OPEN

GOAL TEXT: You want to know how to book a hotel

USER: book a trip

SYSTEM: Are you going on an intracity trip or a business trip?

USER: business trip

SYSTEM: Did you get written permission from your supervisor?

USER: yes

SYSTEM: What country are you traveling to?

USER: USA

SYSTEM: What city are you traveling to?

USER: LA

SYSTEM: For all business trips (even for short trips or ones lasting only a single day) outside the country to EU-countries, Iceland, Norway, Liechtenstein or Switzerland, you will need to fill out an A1 Form , which must be brought with you when you leave the country. You will need to fill out a new A1 for each business trip you take. Filling the form out once, does not allow it to apply for subsequent trips you take.

SYSTEM: Please check the current COVID-19 travel warnings travel restrictions from the foreign ministry and the RKI. Business trips to high risk areas or virus variation areas are not generally not allowed. In In extreme cases, authorization can be given by the leadership of Department 4 (Administrative Department).

SYSTEM: COVID-19: Business trips should be reduced to an absolute minimum and are only allowed when they are absolutely necessary. The decision for necessity can be made by the leaders of the administrative department (see Hygiene concept, 2.4).

SYSTEM: Are you planning to extend your trip with private travel?

USER: no

SYSTEM: Which topic do you want to know more about?

USER: accommodation

SYSTEM: If the trip is being paid for by a third party (not from the university) or was offered to be paid for by one, no money can be reimbursed.

SYSTEM: Where do you plan to stay?

USER: hotel

SYSTEM: It is required that the bill is addressed to your employer with your name entered as the guest.

SYSTEM: In the United States you have a budget of up to 90€ (or 120€ in cities with a population of more than 100,000 or generally/seasonally high hotel prices) for accommodations, so long as you present proof of a paid invoice.

SYSTEM: Exceeding the accommodation allowance can be allowed on a case-by-case basis when justified as necessary, e.g., If there were no cheaper hotels near the event location A cheaper hotel farther away would have resulted in higher transportation costs.

SYSTEM: Resort taxes may be applied to trips outside of Germany.

DIALOG END: SUCCESS

DIALOG LENGTH: 16

SUBJECTIVE QUALITY: 2

SUBJECTIVE LENGTH: 3

D.1.2 Ex. 2

DIALOG: USER: eea1e7836583c5dd7406fd719bc973 | POLICY: ADA | GOAL-TYPE: OPEN

GOAL TEXT: You want to know how to book a hotel

USER: I want to know how to book a hotel

SYSTEM: What topic do you have questions about? Transportation or Accomodation?

USER: Book a hotel

SYSTEM: Hotels can be booked using the university's partner travel agency, through a booking portal, or booked directly The contact for the partner travel agency can be found here. The university recommends these hotels.

DIALOG END: SUCCESS

DIALOG LENGTH: 5

SUBJECTIVE QUALITY: 3

SUBJECTIVE LENGTH: 3

D.2 Easy Goal - FAQ Agent

D.2.1 Ex. 1

DIALOG: USER: 7addf36ad46517524764cd73b84d5b | POLICY: faq | GOAL-TYPE: EASY

GOAL TEXT: You want to know if you can be reimbursed if you need to book a taxi during your trip.

USER: If I need to book a taxi during my trip, can I be reimbursed? If so, how should I do this?

SYSTEM: A taxi is only reimbursable if there is a business related reason or a pressing personal reason, e.g.: insufficient public transportation, Health reasons, Disability.

DIALOG END: SUCCESS

DIALOG LENGTH: 1

SUBJECTIVE QUALITY: 3

SUBJECTIVE LENGTH: 2

D.2.2 Ex. 2

DIALOG: USER: 266f4afc0f4957543ecd710668a801 | POLICY: faq | GOAL-TYPE: EASY

GOAL TEXT: You want to know if you can be reimbursed if you need to book a taxi during your trip.

USER: hello

SYSTEM: Thank you for using the business travel chatbot.Please let us know about your experience by filling out a quick survey, that way we can improve our performance and your experience.

: RESET

USER: hello i would like some information please

SYSTEM: Please use the country database to inform yourself about possible Risks related to your trip. It is required to confirm that you have checked this information and gotten necessary advice as part of your travel application. Your supervisor will confirm this when signing the form giving you permission to travel.

DIALOG END: FAILURE

DIALOG LENGTH: 1

SUBJECTIVE QUALITY: 1

SUBJECTIVE LENGTH: 1

D.3 Hard Goal - HDC Agent

D.3.1 Ex. 1

DIALOG: USER: ba401ddeae9aef6345d8c3dbddb3bf | POLICY: hdc | GOAL-TYPE: HARD

GOAL TEXT: You want to know how much money you can get for your accommodations. You are traveling to France for your next trip
You plan to stay with your brother in his apartment.

USER: Book a trip

SYSTEM: Are you going on an intracity trip or a business trip?

USER: what is the difference between an intracity business trip and business trip?

SYSTEM: Business trips are business related travel outside of the city. The following are not business trips: Company outings, Trips for a side job, Travel to a continuing education training (if it is done out of private interest)

SYSTEM: Intracity business trips are business related travel within the city.

: RESET

USER: Book a trip

SYSTEM: Are you going on an intracity trip or a business trip?

USER: intracity business trip

SYSTEM: Did you get verbal permission from your supervisor?

USER: yes

SYSTEM: Do you want to know more about booking Transportation or about ancillary costs?

USER: incidental costs

SYSTEM: Which incidental costs ocurred during your trip or do you anticipate occurring

USER: accomodation

SYSTEM: Meals which are part of a conference are reimbursable, as long as the conference is being attended out of business interests.

: RESET

USER: book a trip

SYSTEM: Are you going on an intracity trip or a business trip?

USER: business trip

SYSTEM: Did you get written permission from your supervisor?

USER: yes

SYSTEM: What country are you traveling to?

USER: (PRE-NLU) France

SYSTEM: What city are you traveling to?

USER: Paris

SYSTEM: For all business trips (even for short trips or ones lasting only a single day) outside the country to EU-countries, Iceland, Norway, Liechtenstein or Switzerland, you will need to fill out an A1 Form , which must be brought with you when you leave the country. You will need to fill out a new A1 for each business trip you take. Filling the form out once, does not allow it to apply for subsequent trips you take.

SYSTEM: Please check the current COVID-19 travel warnings travel restrictions from the foreign ministry and the RKI. Business trips to high risk areas or virus variation areas are not generally not allowed. In In extreme cases, authorization can be given by the leadership of Department 4 (Administrative Department).

SYSTEM: COVID-19: Business trips should be reduced to an absolute minimum and are only allowed when they are absolutely necessary. The decision for necessity can be made by the leaders of the administrative department (see Hygiene concept, 2.4).

SYSTEM: Are you planning to extend your trip with private travel?

USER: yes

SYSTEM: (A private extension to a trip must remain in proportion to the length of business-related travel. Business interests must remain the priority. For questions, please contact the travel department.

SYSTEM: Which topic do you want to know more about?

USER: accomodation

SYSTEM: If the trip is being paid for by a third party (not from the university) or was offered to be paid for by one, no money can be reimbursed.

SYSTEM: Where do you plan to stay?

USER: with family

SYSTEM: Outside of Germany, you are entitled to 30€ if you are staying with friends or family.

DIALOG END: SUCCESS

DIALOG LENGTH: 14

SUBJECTIVE QUALITY: 4

SUBJECTIVE LENGTH: 4

D.3.2 Ex. 2

DIALOG: USER: 3531297f0bd32d54dcc23e8e8a50ad | POLICY: hdc | GOAL-TYPE: HARD

GOAL TEXT: You want to know how much money you can get for your accommodations. You are traveling to France for your next trip
You plan to stay with your brother in his apartment.

USER: how much am i entitled to claim for accomodation

SYSTEM: What emergency are you experiencing?

USER: not an emergency, but accommodation costs

SYSTEM: What do I do in an emergency on a business trip? Please call the number listed on the Emergency-Card : +123456789. The Emergency-Card is also available in business card format. For more information, please enquire with the business travel department or your institute. This phone number is for emergencies only! Do not distribute it. Emergency App24/7 Emergency calling, Text chat, Video advice per the Emergency call Business App for iOS and AndroidDownload link: Notfall-AppQR-Code:Registration code: Note: distributing the registration code to third parties is strictly prohibited.

SYSTEM: Do you have any further questions?

USER: yes

SYSTEM: What topic do you have questions about? You can either click on an answer from the suggested topics or enter your own text.

USER: how much can i get back for accommodation costs?

DIALOG END: FAILURE

DIALOG LENGTH: 5

SUBJECTIVE QUALITY: 1

SUBJECTIVE LENGTH: 1

E Data Agreement

Before beginning the experiment, users were provided with a data agreement. Although we did not collect any personally identifying data, we wanted to make sure that users were aware of what they would be asked to do, the purpose of the research, what data we would collect and how the data would be processed.

Impressum

Data Collection Policy

Please consider this information carefully before deciding whether to accept this task.

PURPOSE OF RESEARCH: To understand what expectations people have for task-oriented, text-based conversational agents and how these affect their interaction with such systems.

WHAT YOU WILL DO: You will be assigned to interact with one of three dialog systems. You will pretend that you are going on a business trip and interact with the assigned dialog system to find out answers to three different questions about the company's business travel regulations. Not all dialog systems will be able to deliver a good answer, if after trying, you cannot find an answer, you are free to move on to the next goal.

TIME REQUIRED: Participation will take approximately 15-20 minutes.

RISKS: There are no anticipated risks associated with participating in this study. The effects of participating should be comparable to those you would experience from viewing a computer monitor for 15-20 minutes and using a mouse and keyboard.

LIMITATIONS: This task is suitable for all people who can read from and input text into a computer.

CONFIDENTIALITY: Your participation in this study will remain confidential. Your responses will be assigned a code number. You will be asked to provide your Prolific ID, but this **will not be stored**, but rather converted to an anonymous hashed ID. You will be asked to provide your age and gender and previous experience with chatbots/business travel. Throughout the experiment, we may collect data such as your textual input, and your feedback in form of a questionnaire. The records of this study will be kept private. In any sort of report we make public we will not include any information that will make it possible to identify you. Research records will be kept in a locked file; only the researchers will have access to the records.

PARTICIPATION AND WITHDRAWAL: Your participation in this study is voluntarily, and you may withdraw at any time.

DATA REGULATION: Your data will be processed for the following purposes:

- Analysis of the respondents' evaluations of the dialog and their experience
- Analysis of potential influencing factors for individual behavior of the participants in the interaction with the dialog system
- Scientific publication based on the results of the above analyses

Your data will be processed on the basis of Article 6 paragraph 1 subparagraph 1 letter a GDPR. No personally identifying data will be collected. You are entitled to the following rights (for details see here)

- You have the right to receive information about the data stored about your person.
- Should incorrect personal data be processed, you have the right to correct it.
- Under certain conditions, you can demand the deletion or restriction of the processing as well as object to the processing.
- In general, you have a right to data transferability.
- Furthermore, you have the right of appeal to the [redacted] for Data Protection.

You can revoke your consent for the future at any time. The legality of the data processing carried out on the basis of the consent until revocation is not affected by this.

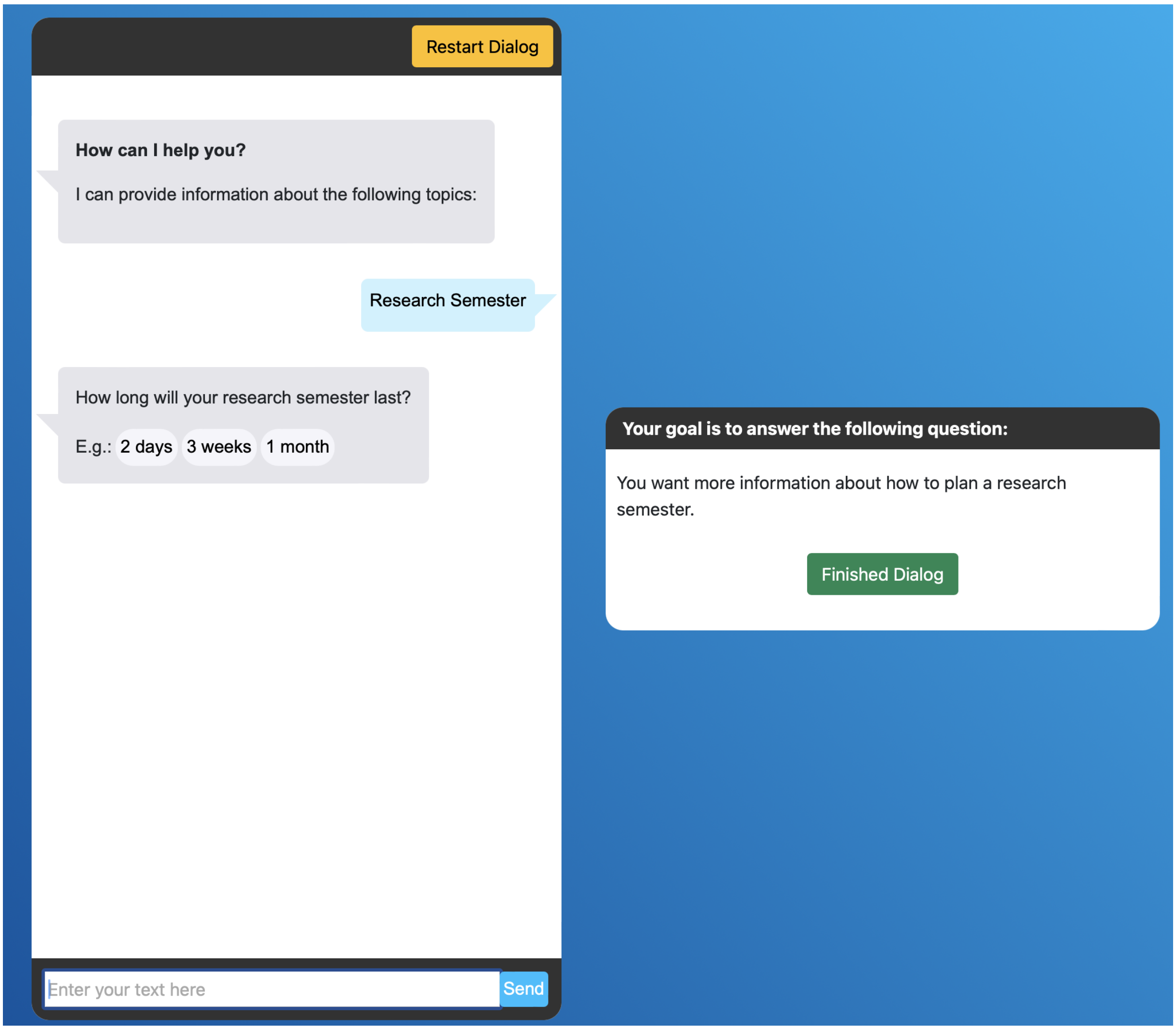
COMPENSATION: Upon completion of this task, you will receive a link to verify your completion with Prolific.

CONTACT: This study is conducted by researchers at [redacted] If you have any questions or concerns about this study, please contact [redacted] at [redacted]

I agree

F Study Instructions

During the interaction, users were provided with the following interface, on the right side they had an information goal for which they should find an answer. On the left side, they had a window with their conversation with the chatbot. Once they felt they had found an answer to their question, they could click on the button underneath the goal to move on to the next dialog.



G Interaction Surveys

G.1 Pre-Interaction Survey

The survey given to users before the interaction can be seen below. Here they were asked general questions about their demographics, previous experience with the domain and chatbots. Users were also asked Likert and free-response questions about their expectations for an information seeking chatbot. In particular about how they expected to be able to input text and and how they expected the chatbot to answer.

Pre-Interaction Survey

Demographic Information

What gender do you identify as?

- ○ Male
- ○ Female
- ○ Other

What is your age?

- ○ Less than 20
- ○ 20 to 29
- ○ 30 to 39
- ○ 40 to 49
- ○ 50 to 59
- ○ 60 to 69
- ○ 70 or older

Previous Experience with Chatbots

- ○ I've never used a chatbot
- ○ I've used a chatbot once
- ○ I've used a chatbot more than once
- ○ I frequently use chatbot(s)
- ○ I use chatbot(s) daily or near daily

Previous Experience with Business Travel

- ○ I've never been on a business trip
- ○ I have been on a business trip once
- ○ I have been on more than one business trip
- ○ I frequently go on business trips
- ○ I am a part of the business travel department at my company

Expectations of Chatbots

The following questions are aimed at understanding what your expectations/previous experiences are for a **business travel chatbot**. Based on your previous knowledge of chatots, please answer them assuming you would be interacting with a chatbot to find out more about business travel regulations at a particular company.

What type of information would you expect to be able to get from a chatbot? In what circumstances would you consider using a chatbot to find out information vs. contacting a real person or reading through company policy documents?

How would you phrase your input to the chatbot? Is this similar or different to how you would use a search engine or ask a real person?

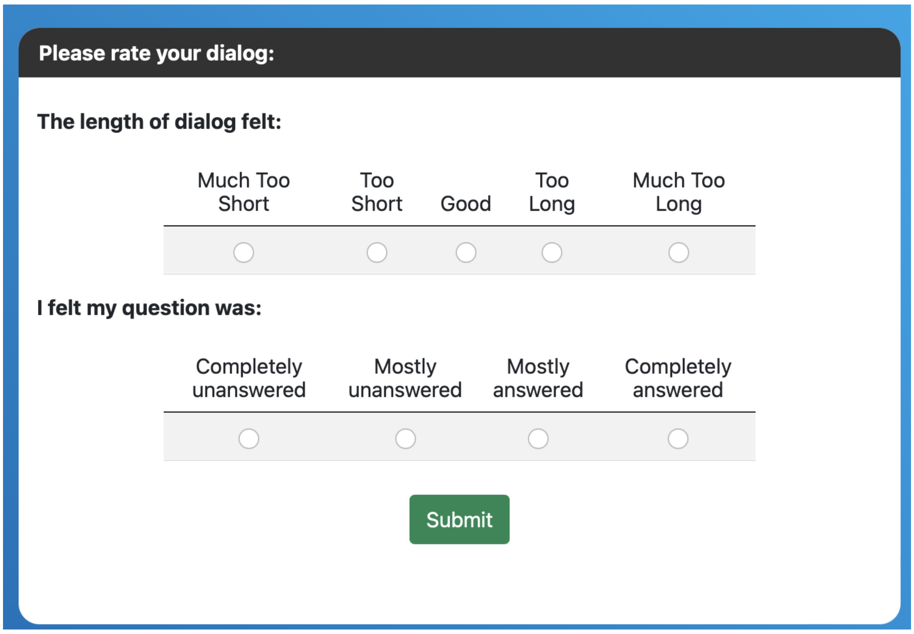
What type or quality of answer would you expect to be able to get from a chatbot, e.g., style, level of detail, correctness, etc.?

Please mark how much you agree with the following statements:

	Strongly Disagree	Disagree	Neither Agree Nor Disagree	Agree	Strongly Agree
In general I think that a chatbot can understand natural language questions, like I would phrase them if I were asking another person	○	○	○	○	○
In general I think that a chatbot can only recognize keywords/or provide fixed options that I can select	○	○	○	○	○
In general I think that in order to get a good answer from a chatbot, I have to ask a very precise question	○	○	○	○	○
In general I think that a chatbot can ask clarifying questions to help me narrow down my problem, e.g., if my original question is vague	○	○	○	○	○
In general I think that a chatbot can only give high-level/general answers to questions	○	○	○	○	○
In general I think that a chatbot can give me a personalized answer specific to my case	○	○	○	○	○
In general I think that a chatbot can provide an immediate answer as a direct response to my question (single turn interaction)	○	○	○	○	○
In general I think that a chatbot would need to ask multiple questions before it is able to give me an answer	○	○	○	○	○

G.2 Post-Dialog Survey

After each interaction, users were asked to rate their perception of the dialog length on a five-point Likert scale and their perception of how well their question was answered on a four-point Likert scale.



G.3 Post-Interaction survey

The survey given to users after interacting with their assigned style of chatbot can be seen below. Users were asked to provide free-form feedback about how well their interactions were met as well as to answer the same Likert questions about their mental model of a chatbot that they had answered prior to the interaction. Users were also asked to fill out a usability questionnaire (Finstad, 2010) and the trust and reliability subscales from the Trust in Automation questionnaire (Körber, 2018) as well as answering free-form questions on their experience and perception of the chatbot.

Post-Interaction Survey

Expectations

How well did the chatbot you interacted with match your expectations for a chatbot? Please describe in which ways your expectations were or were not met.

Chatbot Capabilities

After interacting with the chatbot, please mark how much you agree with the following statements:

	Strongly Disagree	Disagree	Neither Agree Nor Disagree	Agree	Strongly Agree	N/A
The chatbot was able to understand natural language questions, like I would phrase them if I were asking another person	○	○	○	○	○	○
The chatbot was only able to recognize keywords/ input from a fixed set of options I could select	○	○	○	○	○	○
In order to get a good answer, I had to ask a very precise question	○	○	○	○	○	○
The chatbot was able to ask clarifying questions to help me narrow down my problem, e.g., if my original question is vague	○	○	○	○	○	○
The chatbot was only able to give a general answer to my questions	○	○	○	○	○	○
The chatbot was able to give personalized answers specific to my case	○	○	○	○	○	○
The chatbot was able to provide an immediate answer as a direct response to my question (single turn interaction)	○	○	○	○	○	○
The chatbot needed to ask multiple questions before it is able to give me an answer	○	○	○	○	○	○

User Experience

Please mark how much you agree with the following statements:

	Strongly Disagree	Disagree	Neither Agree Nor Disagree	Agree	Strongly Agree
The chatbot was capable of interpreting situations correctly	○	○	○	○	○
The chatbot works reliably	○	○	○	○	○
A malfunction of the chatbot is likely	○	○	○	○	○
The chatbot is capable of handling complex tasks	○	○	○	○	○
The chatbot might make sporadic errors	○	○	○	○	○
I am confident about the chatbot's abilities	○	○	○	○	○
I trust the chatbot	○	○	○	○	○
I can rely on the chatbot	○	○	○	○	○
This chatbot's capabilities met my requirements	○	○	○	○	○
Using this chatbot is a frustrating experience	○	○	○	○	○
This chatbot is easy to use	○	○	○	○	○
I have to spend to long correcting things with this chatbot	○	○	○	○	○

What could the chatbot do well?

What could the chatbot not do well?

What did you like about your interaction with the chatbot?

What did you dislike about your interaction with the chatbot?