It's What You Say and How You Say It: Investigating the Effect of Linguistic vs. Behavioral Adaptation in Task-Oriented Chatbots

Lindsey Vanderlyn
University of Stuttgart
vanderly@ims.uni-stuttgart.de

Ngoc Thang Vu
University of Stuttgart
thang.vu@ims.uni-stuttgart.de

Abstract

Given the conflicting expectations users have for how a dialog agent should sound and behave, there is no one-size-fits-all option for dialog system design. Therefore, adaptation is critical to ensure successful and enjoyable interactions. However, it is not yet clear what the effects of behavioral (interaction style) vs. linguistic adaptation (how the agent says this) are in terms of dialog success and user perception. In this work, we implement three different types of task-oriented dialog agents which can each vary their level of formality. We evaluate subjective and objective metrics of dialog success as well as user perceptions through a user study, comparing the collected data to that of Vanderlyn et al. (2024), where users interacted with the same three types of agents without linguistic adaptation. From this, we draw insights into which subjective and objective aspects of success and user perception are influenced by each type of adaptation. We additionally publish all code, user surveys, and dialog interaction logs.

1 Introduction

As dialog agents become more and more embedded in user's daily lives, research has begun to shift from simply improving their performance to also considering hedonic aspects related to user experience (De Souza et al., 2024). It is no longer sufficient to only obtain high objective performance on test data, but rather, it is necessary for systems to also be enjoyable for users and successful in the wild. To this end, researchers have started looking at what expectations, or mental models, users have about such systems, as these can have a large influence on how successful and enjoyable user interactions are (Kulesza et al., 2012; Bansal et al., 2019; Kim and Lim, 2019; Weitz et al., 2021).

However, each user brings their own personality and previous experiences which influence these expectations. This leads to different users having different expectations for how a dialog agent should behave and how they, in turn, can interact with the agent. For example, users with more domain experience might expect to input a specific question to an agent and get a direct answer, whereas users new to a domain might expect to choose a general topic and be asked follow-up questions to narrow down their information need (Väth et al., 2023). Beyond the agent's behavior, the linguistic style in which it communicates also plays a large role in how users perceive said agent (Wang et al., 2008; Ruane et al., 2021) and interact with it (Janarthanam and Lemon, 2014; Mishra et al., 2022). In particular, Goetz et al. (2003); Li et al. (2017) have shown the importance of an agent's style matching the situation in which it is deployed. However, even here, users have different and even opposite expectations for what is appropriate (Vanderlyn et al., 2024).

To address these differences, growing attention has been given to developing adaptive dialog agents (Motger et al., 2022; Schlimbach et al., 2022; Ait Baha et al., 2023). These adaptations generally consider one of two aspects: 1) either the agent's behavior (Chen and Pu, 2012; Shi and Yu, 2018; Shi et al., 2021; Wambsganss et al., 2021), which controls what information content is outputted to the user each turn or the interaction style, or 2) the agent's linguistic style (Mairesse and Walker, 2010; Yang et al., 2018; Ma et al., 2020; Firdaus et al., 2023), i.e., how that information is presented or formulated. However, while these two approaches could be complementary – e.g., Väth et al. (2023) adapt agent behavior, increasing or reducing the number of questions asked based on user domain familiarity, and Janarthanam and Lemon (2014) vary the use of technical language based on the same – to the best of our knowledge, no work has looked at the effects of linguistic adaptation in different types of dialog system or directly compared the effects of each adaptation. We, therefore, seek to understand the individual and combined effects of linguistic

Proceedings of the 31st International Conference on Computational Linguistics, pages 6120–6149
January 19–24, 2025. ©2025 Association for Computational Linguistics

and behavioral adaptation on user interactions and user perceptions of task-oriented dialog agents.

Concretely, we investigate the following questions:

RQ1: What are the effects of adapting language formality in different types of task-oriented dialog systems?

- **RQ1.1:** How does it effect user interactions/ success?
- **RQ1.2:** How does it affect user perceptions?

RQ2: What effect does linguistic adaptation have compared to behavioral adaptation?

- **RQ2.1:** Can these types of adaptation be combined?
- **RQ2.2:** What aspects of the interaction are affected by each type of adaptation?

To answer these questions, we implement three different types of dialog agents for the domain of business travel, each of which can adapt the formality of its linguistic output based on user preference or not. We then perform a user study where participants each conduct three dialogs, rating the quality of each dialog as well as their overall perception of the agent. Finally, we compare the data we collect to that collected by Vanderlyn et al. (2024), where users interacted with the same three types of dialog agent without linguistic adaptation.

Our main contributions are: 1) Demonstrating that adapting formality can significantly improve objective and subjective measures of task-oriented dialog performance. 2) Showing that the effect of the linguistic adaptation varies depending on the type of dialog system used. 3) Showing that behavioral and linguistic adaptation can effectively be combined to improve performance over baselines or single-adaptations. 4) Providing insights into the relative advantages and disadvantages of each type of adaptation And 5) Releasing [1] our code and a new version of the REIMBURSE Dialog Mental Models (RDMM) dataset with twice as many dialogs and mental model annotations.

2 Related Work

2.1 Linguistic Adaptation

One common goal of linguistic adaptation is creating personality adaptive chatbots (Walker et al., 1997; Ait Baha et al., 2023), which can be applied to any domain. This can be done, e.g., by adjusting linguistic markers in their output to mimic different personalities (Hu et al., 2018). However, determining how an agent should adapt is challenging as some studies have found users like the agent to match their personality (Ahmad et al., 2020a), while others found the opposite (Liew and Tan, 2016) or a preference for only extroverted agents (Ahmad et al., 2020b; Hanna and Richards, 2015).

Other approaches consider creating more empathetic agents (Ma et al., 2020) or creating agents which can adapt the politeness of their output, e.g., using the user's emotional state as a signal. Wang et al. (2008) perform a Wizard-of-Oz study for a pedagogical agent, and find that appropriate agent politeness can increase student learning outcomes. Mishra et al. (2022); Firdaus et al. (2023) additionally found that adapting the politeness of an agent to user emotions and personality led to higher success and shorter dialogs when tested in simulation. However, while it is clearer how politeness adaptation could be applied, most previous work has only evaluated against simulated users.

In contrast, Janarthanam and Lemon (2014) take a domain-specific approach, varying the use of technical language on the user's domain familiarity. Through user test, they found that adapting the language generation led to increased success and decreased dialog lengths. Taking inspiration from this, we look to see if a more general framework can also have a similar effect, e.g., formality, which also includes aspects of language complexity.

2.2 Behavioral Adaptation

One way to adapt agent behavior is to solicit explicit user feedback during an interaction, which can, e.g., be used to give users more relevant recommendations or retrieve more relevant skills (Chen and Pu, 2012; Narducci et al., 2018; Shi et al., 2021). However, such explicit feedback places an additional burden on the user.

To this end, other approaches look into implicit adaptation of an agents behavior. For example Ritschel and André (2018) used the presence of smiles to update the content of jokes an agent told. Shi and Yu (2018) tracked user sentiment as a reward signal for their Reinforcement Learning (RL) agent, and find that it improves the dialog success and reduces dialog length in a bus information search compared to a static policy. However, social signals like these are not always available. To this end, Wambsganss et al. (2021) look into adapting the advice of an argument tutor based on the quality of written student arguments. Väth et al. (2023), also adapt agent behavior based on only

[1] code and data: `removed_for_anonymity.com`

text, however here they use only the initial user utterance. Based on the style of question users ask, an agent interpolates between detailed, multi-turn dialog interactions aimed at users new to a domain and single-turn FAQ style interactions for more experienced users. However, despite the improvements these types of adaptation bring, to the best of our knowledge, combining behavioral and linguistic adaptation has not yet been investigated.

3 Implementation

For this study, we implement three different types of dialog agents: one behaviorally adaptive agent and two static agents, following the same implementation as Vanderlyn et al. (2024). For each of these, we also implement four different linguistic templates which can be chosen by the user.

3.1 Baseline Agents

For the behaviorally static agents, we implement a handcrafted (HDC) dialog system and an FAQ system, which represent two of the most common interaction styles for information seeking systems. For both, we use the REIMBURSE-EN dialog tree (Väth et al., 2024) to define the agents' behavior. This tree was designed by subject-experts for the domain of business travel; nodes represent system utterances and edges the possible user responses.

For the HDC agent, we implement a policy which follows the dialog tree node-by-node, outputting the text of the current node at every step. If the user is asked a question, we calculate the similarity (Reimers and Gurevych, 2019) between their response and the expected answers at that node. The answer with the highest similarity to the user's input is selected and the agent travels to the associated follow-up node (equivalent to ex. A in Figure 1). Thus, the interaction follows a paradigm where the system asks questions and the user answers them until they reach their information goal. This style of interaction is particularly useful if the user cannot formulate a concrete question or if the exact details of their situation need to be clarified.

For the FAQ agent, we use the nodes of the dialog tree as documents for the FAQ retrieval. We calculate the similarity (Reimers and Gurevych, 2019) of the user input to all nodes, immediately returning the one most similar to the user question (equivalent to ex. B in Figure 1). In contrast to the HDC agent, interactions with the FAQ agent then follow a paradigm of users inputting a specific question and receiving a single-turn answer. This is particularly useful if the user has a simple, concrete question which does not need further clarification.

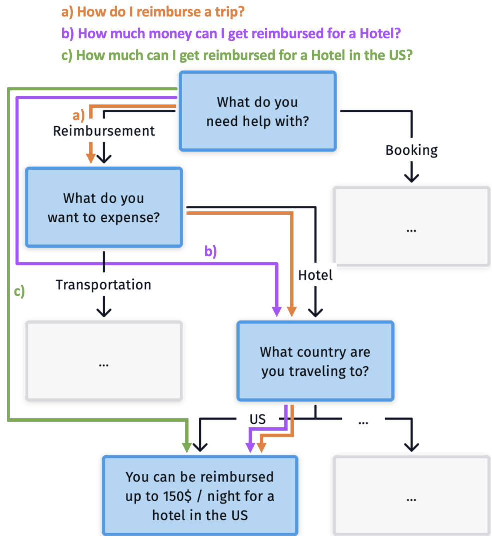


Figure 1: Example of CTS agent behavior (Väth et al., 2023). The agent adapts its behavior based on user expectations implicitly encoded in their input.

3.2 Behaviourally Adaptive Agent

Here, we implement a Conversational Tree Search (CTS) agent (Väth et al., 2024), using the same training parameters as Vanderlyn et al. (2024). We use the same dialog tree to define the agent's general behavior, however, in comparison to the static agents described above, CTS agent learns to adapt its behavior based on the initial user input. We train this agent on questions from users with different interaction expectations, e.g, inputting a concrete question and getting a concrete answer or inputting a vague question and being asked follow-up questions/ given more background information until a concrete information need can be identified. The agent then uses RL to adapt its behavior based on a user's input text, outputting or skipping over nodes in the graph as appropriate to answer to the user question. The extreme ends of the CTS agent's behavior can thus be modeled as a non-adaptive handcrafted dialog system (asking every node in the tree) or as an FAQ system (directly giving an answer), with the CTS agent able to adaptively model the full spectrum of behavior between. An illustration of this can be seen in Figure 1, where different user inputs lead to different agent behavior.

All flight reservations must be made through the university's designated Partner Travel Agency, which can be reached at +123456789. Booking through any other agency is not permitted. Additional details are available on the business travel website, where you will also find the contact information for the Partner Travel Agency.

Flights must be booked through the university's Partner Travel Agency (Tel. +123456789). It is not possible to book through an alternative agency. You can find more information on the business travel website. Contact information for the partner travel agency can be found here.

You need to book your flights through the university's Partner Travel Agency. Just give them a call at +123456789. You can't use any other agency for this. For more details, check out the business travel website. All the contact info for the travel agency is there too.

Just a heads-up, you gotta book your flights through the university's Partner Travel Agency. Give them a call at +123456789. You can't use any other agency, though. For more details, check out the business travel website. All the contact info for the travel agency is right there. Happy travels! 😊✈️

Figure 2: Example of the four formality style template, the most formal at the top and least formal at the bottom.

3.3 Linguistic Templates

To design the adaptive templates, we took a combined empirical and theoretical approach. As the behavioral adaptation looks at supporting different levels of user domain-familiarity, we wanted the linguistic adaptation to complement this. To this end, we analyzed the user mental models collected in the RDMM dataset (Vanderlyn et al., 2024), to see which aspects of linguistic style were important to users. We found that the level of formality and the level of detail a chatbot should use came up most frequently. We chose to adapt formality, as there is a clearer theoretical framework.

Following the framework of Heylighen and Dewaele (1999), we tested different LLM prompts which could be used to automatically change the formality level of the original dialog agent utterances. In the end, we settled on three prompts (Appendix E), two to generate less formal utterances than the RDMM baseline and one to generate more formal ones. We limit the adaptive templates to four different levels of formality, as internal testing showed further levels of granularity were difficult to differentiate. The ensuing templates were then manually edited to ensure corresponding nodes had the same information content regardless of linguistic style. Finally we verified that the templates accurately represented the intended formality levels through a pilot study. Examples of each style can be seen in Figure 2 and in Appendix F

3.4 Datasets

REIMBURSE-En All agents were trained/ built using the REIMBURSE-En dataset (Väth et al., 2024), which consists of a dialog tree for the domain of business travel, as well as user questions and user answers (in response to system questions). For dataset statistics, see Appendix G.

	FAQ	HDC	CTS
# Dialogs	61	66	61
# Successful dialogs	35	29	47
# Turns/dialog (Avg.)	2.3	13.3	7.4
# User Surveys	21	22	20
# Dialogs	65	62	64
# Successful dialogs	41	33	57
# Turns/dialog (Avg.)	2.7	12.5	6.4
# User Surveys	22	22	22

Table 1: Original corpus statistics for the RDMM dataset, used as a baseline in this experiment (top) and our extensions for interactions with linguistically adaptive agents (bottom).

RDMM To analyse the effect of linguistic adaptation, we use the data collected in the RDMM dataset by Vanderlyn et al. (2024) as a baseline, extending this dataset with the interaction and survey data from our experiment (Table 1). The dataset consists of dialogs conducted with one of three types of dialog agent as well as surveys detailing each user's mental models of a chatbot before and after the interaction, perceived trust, reliability, and usability ratings after the interaction, and free-form feedback to their experience. Each dialog is labeled with 1) the type of dialog agent, 2) the user's assigned information goal, 3) the number of turns, 4) the end condition (success or failure), and 5) user ratings for perceived length and quality of answer.

We extend this dataset with 192 new dialogs with linguistically adaptive dialog agents, 66 new pre-surveys, and 66 new post surveys, following the same survey method as the original dataset. Example dialogs can be seen in Appendix F and the survey questions in Appendix I.

4 Pilot Study

To verify perception of the templates, we performed a pilot study. We selected ten representative nodes from the dialog graph and recruited 30 participants from the crowdsourcing platform Prolific. Each participant was asked to answer a single, 5-point Likert item about what formality they would expect from a chatbot for business travel, ranging from 1='very formal' to 5='very casual'. They were then randomly assigned one of the dialog nodes and asked to provide their subjective impression for each of the language templates (shown in a random order). As formality is generally considered to be a sliding scale (Heylighen and Dewaele, 1999), we asked participants to then rank the templates by their perceived formality. Finally, we asked them to select which (single) template they considered most appropriate for a dialog agent.

Qualitatively, we find user perceptions lined up well with our expectations. For examples the most formal template was generally seen as “formal", “rigid", “professional", or “precise", while the least formal template was generally seen as “casual", “informal", “playful", or “simple". Furthermore, we verified that user perception of the templates correlated well with the intended order of formality ($r(110)$=0.66, p<0.001). Finally, we found that all templates were chosen as the most appropriate by between a minimum of 4 and maximum of 13 users, highlighting that different users have different expectations and perceptions of formality and reinforcing the need for adaptation.

However, we additionally found that the Likert scale did not match users’ actual preferences for linguistic style. That is, users’ stated formality preference did not correlate well with the formality of the template they considered most appropriate for a chatbot. This is likely because each user seemed to have a different idea of what the words “formal” vs. “casual” means. To avoid this problem in the main study, we instead provided examples of each style rather than asking users to share their preference on an abstract scale.

5 Main Study

In order to compare our results to Vanderlyn et al. (2024), who studied behavioral but not linguistic adaptation, we follow a parallel study design.

5.1 Study Design

Each participant was randomly assigned to interact with either the behaviorally adaptive CTS agent or one of the two static agents. All agents used adaptive linguistic templates. To avoid potential crossover effects from multiple dialog systems and to ensure our results are comparable with previous work, we use a between-subjects design.

Before the interaction, participants were asked to complete a pre-survey, providing demographic information and their experience with chatbot and business travel. Additionally, they were given an example each of the four linguistic styles and asked to choose which one they considered most appropriate for a chatbot about business travel. We used this choice to select the formality of the interaction, which was then kept static through all three dialogs. We chose to use this type of explicit cue for the adaptation as the interactions are relatively short and not all users provide long enough texts to accurately classify the formality level of their input. In this way, we can focus purely on the effect of the adaptation without being affected by the accuracy of a classifier.

During the chat interaction, participants were asked to conduct three dialogs with the system, each with a different pre-defined goal. They were given no instructions on how they should interact. For the first dialog, each participant was randomly assigned a general information goal they wanted answered, e.g., *“you want to find out information on how to book a business trip.”* These represented users new to a domain and were intended to help familiarize participants with the topic of business travel. For the second dialog, participants were randomly assigned a simple, concrete goal. These had a specific answer in the dialog graph, but did not require information about a user’s specific case to answer, e.g., *“You want to know you if can get reimbursed for a taxi”*. In the last dialog, users were randomly assigned a complex question as their goal. These also had a specific answer, but required details about the user’s exact situation to answer, e.g., *“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, participants were asked to complete a post-survey about their mental model of the dialog agent and their perception of the interaction.

5.2 Participants

We recruited 66 participants from the crowdsourcing platform Prolific, split equally between the three conditions. Each participant’s primary language was English. Each participant conducted three dialogs, however 7 were removed due to technical errors resulting in a total of 191 dialogs.

Of the participants, 30 were male, 35 were female, and 1 person identified as other. Their ages ranged from 20 to 59. On average, participants had some familiarity with dialog systems (M=3.48 on a 5-point Likert item) and limited familiarity with business travel (M=2.1 on a 5-point Likert item). 12 users chose formality level one, 10 chose two, 20 chose three, and 24 chose formality level four. There were no statistically significant differences in the distributions of gender, age, previous experience, or preferred user formality level between the conditions.

6 Evaluation Methods & Results

To evaluate the study, we perform both quantitative and qualitative analysis, looking at objective and subjective indicators of dialog success. The full survey can be seen in Appendix I

6.1 Objective and Subjective Dialog Metrics

To objectively evaluate dialog interactions, we record the length of each interaction (the total number of dialog turns) and the objective success of each dialog (whether the user reached the node associated with their goal). To evaluate how users perceived each dialog, we ask them to answer two Likert items after each conversation, rating 1) how well their question was answered and 2) how appropriate the dialog length felt. Results can be seen in Table 2.

Model	# Turns	Success	Perceived Length	Answer Satisfaction
CTS	7.38	77.05	**2.92**	2.87
CTS+*ling_ad*	**6.44**	**89.06***	**3.08**	**3.16***
FAQ	**2.26**	57.38	2.28	2.61
FQA+*ling_ad*	2.68	**63.08**	**2.65***	**2.78**
HDC	13.32	43.94	**3.08**	2.41
HDC+*ling_ad*	**12.53**	**54.84**	2.73	**2.63**

Table 2: Average objective and subjective performance metrics per dialog system. Perceived length is measured on a 5-point scale (1=much too short, 5=much too long), perceived quality on a 4-point scale (1=question not at all answered, 5=completely answered). Entries with * show a significant difference (p<0.05) compared to the linguistically non-adaptive version.

6.2 Mental Models

In order to measure users' mental models of the agent they interacted with, we ask them to complete a survey after the three dialog interactions. As it is difficult to measure mental models without also influencing them (Rowe and Cooke, 1995), we take two complementary approaches. In the first approach, we asked users to answer a series of eight Likert questions about what abilities they thought the chatbot has. The first four entries related to user expectations of what input the dialog agent could understand and the second four to what type of output it could generate. The results of these questions can be seen in Appendix B.

We then asked users free-response questions based on the retrospective technique proposed by Hoffman et al. (2018), asking them about their expectations and perceptions of the system's abilities. To analyze the free-responses, we used the standard content analysis technique proposed by (Hsieh and Shannon, 2005), where utterances are annotated with a fixed set of labels generated from the collected data. We show the frequencies of mental models most affected by adaptation and in Figure 3 and the full results in Appendix B.

6.3 Trust, Reliability, and Usability

To evaluate trust and reliability, we used the Trust in Automation (TiA) questionnaire (Körber, 2018) and for usability, the UMUX scale (Finstad, 2010). Results can be seen in Appendix C.

To gain a more granular insights we also asked users to answer free-form questions about which aspects of the interaction they liked or disliked and performed content analysis. The frequencies at which codes appeared can be seen in Figure 4 for the codes which were most influenced by adaptation, with full results in Appendix C.

7 Discussion

We first look at the effects of linguistic adaptation for each type of agent individually before comparing the overall effects of both types of adaptation. To do so, we make use of generalized linear mixed models (GLMMs) (see Appendix D), to account for both the fixed effects of the adaptations as well as any random effects that might have been present.

7.1 RQ1: Effect of Linguistic Adaptation in Different Dialog Systems

First, we consider the objective and subject effects of linguistic adaptation per dialog system.

7.1.1 RQ1.1: Objective Measures

When looking at the GLMM analysis in Appendix D, we find linguistic adaptation significantly improves objective success (p<0.05). We found no significant interaction effects, indicating that this improvement is not dependent on the type of dialog agent. From the free-response comments, we find adapting formality helped users better understand the dialog agent, which could explain the more successful dialogs. This is in line with the results of Janarthanam and Lemon (2014), who found that varying the use of domain-specific language based on the users' experience with the domain increased success in dialogs. However, here we show that formality, a more general framework which can be

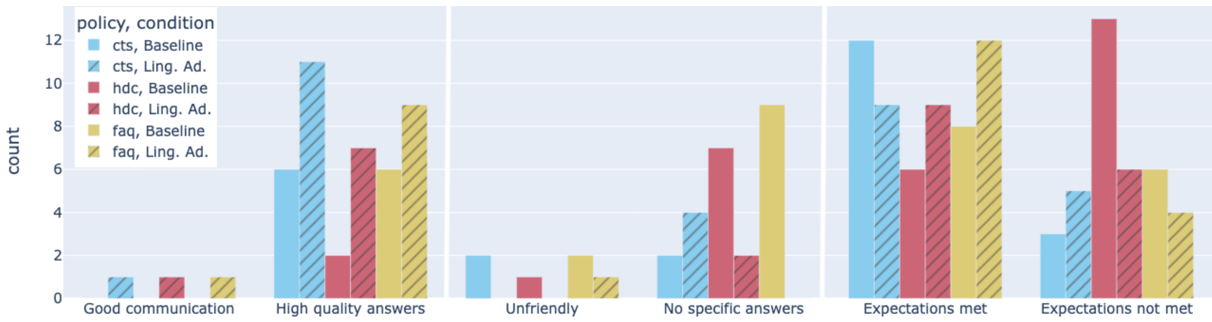


Figure 3: User mental models about different kinds of dialog agents(blue/left=CTS, red/middle=HDC, yellow/right=FAQ; striped=with linguistic adaption, solid=without). Codes on the left side of the figure represent perceived system strengths, codes in the middle system weaknesses, and codes on the right how well users felt their expectations were met.

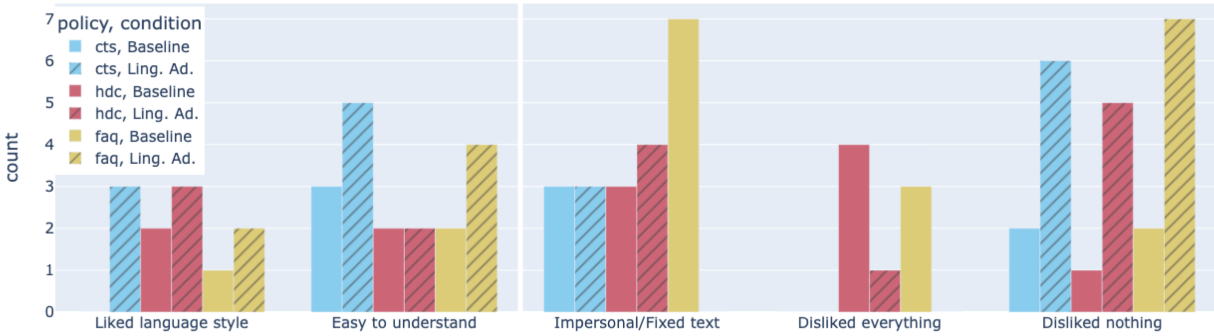


Figure 4: User likes and dislikes of different kinds of dialog agents(blue/left=CTS, red/middle=HDC, yellow/right=FAQ; striped=with linguistic adaption, solid=without). Codes on the left side of the figure represent things users liked and codes on the right user dislikes.

easily applied to any domain, can also achieve such an effect.

7.1.2 RQ1.2: Subjective Measures

We find that regardless of agent, the subjective success and answer quality improved by adapting the formality, while the number of users who disliked any aspect of the interaction decreased (Appendix D). However, in this section we focus on the effects which varied based on the type of dialog agent.

HDC Users interacting with the HDC dialog agent generally had longer, more rigid dialogs. In Figure 4, we see that changing the language style was not enough to mitigate frustrations with rigid texts. In fact, when looking at the dialogs themselves, we find that around 19% were ended early by the user before getting an answer, which could also explain why the HDC condition was the only one where users did not find the linguistically adaptive templates easier to understand: users did not always get far enough to even receive an answer. Despite this, in the GLMM analysis (Appendix D), we see a significant decrease ($p<0.05$) in perceived dialog length. We also see a drop in users who felt that the dialog system could not give specific answers and that overall user expectations were better met (Figure 3), which suggests that even though they viewed the dialogs as quite rigid, adapting the linguistic style helped users to perceive the answers they did get as more specific/relevant to their questions and the overall experience as better able to meet their expectations.

FAQ In the FAQ condition, the GLMM analysis (Appendix D) showed that users perceived dialog length as significantly ($p<0.05$) more appropriate (Table 2) when formality was adapted, compared to the static templates which were viewed as too short. We also see from the Likert responses in Table 4 (Appendix B) that users felt the adaptive agent was significantly ($p<0.05$; Wilcoxon-Mann-Whitney test) more capable of giving specific answers to questions. The same trend emerges in the free-responses (Figure 3), where no users thought the agent incapable of specific responses. Finally, in Figure 4, we see that more users found the agent easy to understand in the adaptive setting, which could contribute to why they viewed the answers as more personalized to their question.

CTS In the CTS condition, users had higher expectations of the dialog system, compared to the baseline, with more users expecting single turn interactions, as are common with generative models. Despite this, we see that linguistic adaptation still improved subjective aspects of the interaction. Figure 3 shows that the linguistically adaptive agent was no longer regarded as unfriendly. Furthermore, Figure 4 shows that more users liked the language style, and more users found the agent easy to understand.

7.2 RQ2 Linguistic vs. Behavioral Adaptation

In this section, we look at individual and combined effects of linguistic and behavioral adaptation using the data we collected and data from the RDMM dataset (Vanderlyn et al., 2024).

7.2.1 RQ2.1: Combining both Types of Adaptation

From the GLMM analysis in Appendix D, we find both linguistic and behavioral adaptation had significant positive effects ($p<0.05$; $p<0.0001$ respectively) on objective and subjective dialog success. Table 2 shows that the combination of both types of adaptation resulted in highest objective performance ($p<0.05$ Barndard Exact test), highest perceived answer quality ($p<0.05$ Wilcoxon-Mann-Whitney test). Figure 4 shows that this combination was found easiest to understand. From this, we see that the best results, both in terms of objective performance and user perception of the answers, come when both the behavior and the linguistic style can adapt to match user expectations.

7.2.2 RQ2.2: Comparison of Adaptation Types

As it may not always be possible to implement both linguistic and behavioral adaptation, we discuss their individual effects in this section.

Behavioral Adaptation In terms of dialog success, we find behavioral adaptation has a larger effect both on objective and perceived success than linguistic adaptation (Appendix D). This suggests that to improve task-success, it may be more important for an agent to match the interaction style a user expects than it is to match the linguistic style. When we look at the dialogs themselves, we find evidence that users of the static FAQ and HDC agents, who had different expectations for how they should interact, often ended the interaction before they received a relevant answer. In contrast, when the interaction style adapted to user expectations, but the language style was kept static, users more often continued the interaction even if they may not always have understood the agent. In terms of user perception, we find that behavioral adaptation can improve how users view the quality of answers as well as the agent's ability to give specific answers (Figure 3) and can increase user satisfaction with the interaction (Figure 4). However, as behavioral adaptation requires changing an agent's entire policy, it may require substantial effort to implement.

Linguistic Adaptation We find that adapting the formality also improved the performance of dialog agents (Appendix D), although to a lesser extent than the behavioral adaptation. When looking at Figure 4, we also see this type of adaptation can improve the perceived quality of communication. For shorter interaction styles (FAQ and CTS), where the user is shown less extraneous information, adapting formality also improved how well users were able to understand the dialog agent. Regardless of dialog system, we also find that the linguistic adaptation led to a significant decrease ($p<0.05$; Barnard Exact test) in users who had something they disliked about the interaction (Figure 4) and a significant increase ($p<0.01$; Barnard Exact test) in the perceived quality of answers (Figure 3). In both cases, these improvements were on par with, or beyond, the effect of only behavioral adaptation. This suggests, that regardless of whether the interaction style matches user expectations, adapting the language style can improve dialog success and user experience. This is particularly interesting as, compared to behavioral adaptation, linguistic adaptation does not necessarily require a new policy, making it potentially easier to implement. In particular, as we show here, tools like LLMs can be leveraged to reduce the effort of adding such adaptations to an existing dialog agent.

8 Conclusions

In this work, we implement three different types of dialog agents able to adapt the formality of their templates based on user preference. We perform a user study to compare the effects of linguistic adaptation across these different types of information seeking dialog agents and investigate the overall effects of linguistic vs. behavioral adaptation, both individually and when combined. We additionally expand the RDMM dataset (Vanderlyn et al., 2024) with the new dialogs and user surveys we collect.

When looking at the effect of linguistic adaptation on different types of dialog agent, we find that the exact effects depend on the type of agent, but that regardless of agent, linguistic adaptation resulted in increases to objective and subjective metrics of dialog success. We thus show that a more general intervention, such as adapting formality, can lead to similar improvements as those from Janarthanam and Lemon (2014), who applied a task-specific adaptation, varying the use of technical language.

Comparing both types of adaptation, we find that the best performance comes when the interaction style *and* the linguistic style of an agent match user expectations. Adapting both of these led not only to the highest objective dialog success, but also to the highest subjective dialog success. Finally, when looking only at the relevant effects of each type of adaptation, we found that behavioral adaptation had a greater effect on objective success. However, adapting formality still increased dialog success and had a larger effect on subjective metrics, while being much simpler to implement.

9 Limitations

As this study was relatively small, with only 22 participants in each of the 6 conditions, we are only able to talk about medium to large sized effects from the two types of adaptation. Given the increase in, e.g., usability and answer satisfaction, it is possible that with more participants we would have also found significant effects here. Additionally, due to the number of participants, we were not able to look into effects of each type of language style. Finally, while we tried to recruit a diverse background of participants, the study was conducted with participants who had English as a primary language, which may bias the results.

10 Ethical Considerations and Risks

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 paid users in both the pilot study and the main study at a rate of 10.50£/hr, which was in-line with minimum wage in the country of our research institution.

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

Rangina Ahmad, Dominik Siemon, Daniel Fernau, and Susanne Robra-Bissantz. 2020a. Introducing" raffi": A personality adaptive conversational agent. In *PACIS*, page 28.

Rangina Ahmad, Dominik Siemon, and Susanne Robra-Bissantz. 2020b. Extrabot vs introbot: The influence of linguistic cues on communication satisfaction. In *AMCIS*.

Tarek Ait Baha, Mohamed El Hajji, Youssef Es-Saady, and Hammou Fadili. 2023. The power of personalization: A systematic review of personality-adaptive chatbots. *SN Computer Science*, 4(5):661.

Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S Lasecki, Daniel S Weld, and Eric Horvitz. 2019. 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.

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

Ana Carolina Rossi De Souza, Pamella Augusta De Lima Mariano, Guilherme Corredato Guerino, Ana Paula Chaves, and Natasha Malveira Costa Valentim. 2024. Technologies for hedonic aspects evaluation in text-based chatbots: A systematic mapping study. In *Proceedings of the XXII Brazilian Symposium on Human Factors in Computing Systems*, IHC '23, New York, NY, USA. Association for Computing Machinery.

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.

J. Goetz, S. Kiesler, and A. Powers. 2003. Matching robot appearance and behavior to tasks to improve human-robot cooperation. In *The 12th IEEE International Workshop on Robot and Human Interactive Communication, 2003. Proceedings. ROMAN 2003.*, pages 55–60.

Nader Hanna and Deborah Richards. 2015. The influence of users' personality on the perception of intelligent virtual agents' personality and the trust within a collaborative context. In *Advances in Social Computing and Multiagent Systems*, pages 31–47, Cham. Springer International Publishing.

Francis Heylighen and Jean-Marc Dewaele. 1999. Formality of language: definition, measurement and behavioral determinants. *Interner Bericht, Center "Leo Apostel", Vrije Universiteit Brüssel*, 4(1).

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.

Zhichao Hu, Jean Fox Tree, and Marilyn Walker. 2018. Modeling linguistic and personality adaptation for natural language generation. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 20–31, Melbourne, Australia. Association for Computational Linguistics.

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

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.

Jingyi Li, Michelle X. Zhou, Huahai Yang, and Gloria Mark. 2017. Confiding in and listening to virtual agents: The effect of personality. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces*, IUI '17, page 275–286, New York, NY, USA. Association for Computing Machinery.

Tze Wei Liew and Su-Mae Tan. 2016. Virtual agents with personality: Adaptation of learner-agent personality in a virtual learning environment. In *2016 Eleventh International Conference on Digital Information Management (ICDIM)*, pages 157–162.

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

François Mairesse and Marilyn A Walker. 2010. Towards personality-based user adaptation: psychologically informed stylistic language generation. *User Modeling and User-Adapted Interaction*, 20:227–278.

Kshitij Mishra, Mauajama Firdaus, and Asif Ekbal. 2022. Please be polite: Towards building a politeness adaptive dialogue system for goal-oriented conversations. *Neurocomputing*, 494:242–254.

Quim Motger, Xavier Franch, and Jordi Marco. 2022. Software-based dialogue systems: Survey, taxonomy, and challenges. *ACM Comput. Surv.*, 55(5).

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.

R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

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.

Elayne Ruane, Sinead Farrell, and Anthony Ventresque. 2021. User perception of text-based chatbot personality. In *Chatbot Research and Design*, pages 32–47, Cham. Springer International Publishing.

Ricarda Schlimbach, Heidi Rinn, Daniel Markgraf, and Susanne Robra-Bissantz. 2022. A literature review on pedagogical conversational agent adaptation. In *Pacific Asia Conference on Information Systems*, volume 1.

Chen Shi, Yuxiang Hu, Zengming Zhang, Liang Shao, and Feijun Jiang. 2021. User feedback and ranking in-a-loop: Towards self-adaptive dialogue systems. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 2046–2050, New York, NY, USA. Association for Computing Machinery.

Weiyan Shi and Zhou Yu. 2018. Sentiment adaptive end-to-end dialog systems. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1509–1519, Melbourne, Australia. Association for Computational Linguistics.

Natalya Pya Simon N. Wood and Benjamin Säfken. 2016. Smoothing parameter and model selection for general smooth models. *Journal of the American Statistical Association*, 111(516):1548–1563.

Lindsey Vanderlyn, Dirk Väth, and Ngoc Thang Vu. 2024. Investigating the effect of mental models in user interaction with an adaptive dialog agent. *Preprint*, arXiv:2408.14154.

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.

Marilyn A. Walker, Janet E. Cahn, and Stephen J. Whittaker. 1997. Improvising linguistic style: social and affective bases for agent personality. In *Proceedings of the First International Conference on Autonomous Agents*, AGENTS '97, page 96–105, New York, NY, USA. Association for Computing Machinery.

Thiemo Wambsganss, Tobias Kueng, Matthias Soellner, and Jan Marco Leimeister. 2021. Arguetutor: An adaptive dialog-based learning system for argumentation skills. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA. Association for Computing Machinery.

Ning Wang, W. Lewis Johnson, Richard E. Mayer, Paola Rizzo, Erin Shaw, and Heather Collins. 2008. The politeness effect: Pedagogical agents and learning outcomes. *International Journal of Human-Computer Studies*, 66(2):98–112.

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.

Simon N Wood. 2004. Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association*, 99(467):673–686.

Simon N. Wood. 2010. Fast Stable Restricted Maximum Likelihood and Marginal Likelihood Estimation of Semiparametric Generalized Linear Models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 73(1):3–36.

Simon N Wood. 2017. *Generalized additive models: an introduction with R*. chapman and hall/CRC.

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.

A Adaptive Dialog Agent Implementation

The CTS agent was published under the GPL-3.0 license, which is consistent with our use of it in this paper.

In this section, we describe the CTS task and evaluation objectives as well as the model parameters and training parameters we used in this study.

A.1 Conversational Tree Search Task

Given a dialog tree (e.g. Figure 1) and a user input, the goal of the CTS task is to efficiently traverse this tree to provide an answer to that question or information need (Väth et al., 2023). A RL policy is trained to either output text at the current node (e.g., asking a question or providing additional information as context to the user), or to skip that node and directly move to a neighbouring node..

In order to model different styles of user interaction (general domain exploration vs specific questions), this task framework considers two types of goal setting:

- **Guided Dialog**: where a user has a vague information goal and needs to be guided through the dialog graph, node by node.
- **Free Dialog**: where a user has a concrete question which the system should answer as directly as possible. At each turn the agent may choose to ask a question to clarify the user's intent or skip to the next node on the path to the user's answer.

A.2 Evaluation Objectives

In this paper, we follow the modified evaluation method in (Väth et al., 2024), which was also used by Vanderlyn et al. (2024). In this method, a concrete goal is drawn for users in both guided and free modes, in contrast to the original method (Väth et al., 2023), which only considered turn-wise goals (agent only needs to navigate to the correct follow-up node) for guided-mode. In short, the evaluation objectives used in this paper are:

- **Free Mode**: In free mode, the objective is to maximize both *task success* (reaching a final, pre-drawn goal node) and the *skip ratio* (percentage of nodes in the dialog which are skipped instead of outputted to the user).
- **Guided Dialog**: For guided dialog, the objective is to maximize *task success* while minimizing the *skip ratio*.

A.3 RL Model Parameters and Training Resources

Parameter	Value
Layer type	Linear
Activation (after each layer except in Dialog Mode Classifier Head)	SELU
Shared Layer Neurons (one value / layer)	$8096, 4096, 4096$
Value Function Layer Neurons (one value / layer)	$2048, 1024$
Advantage Function Layer Neurons (one value / layer)	$4096, 2048, 1024$
Dialog Mode Classifier Neurons (one value / layer)	$256, 1$
Dropout (after each layer)	25%

The agent was trained on a single RTX 3090 GPU. In total, we required approximately 840 total hours including parameter tuning and training.

A.4 RL Training Parameters

The following parameters were used to train the CTS agent (chosen through manual tuning) with performance measured against a user simulator:

Parameter	**Value**
Optimizer	Adam
Learning Rate	$1e^{-4}$
λ	0.1
Maximum Training Dialog Turns	$2M$
Max. Gradient Norm	1.0
Batch Size	256
γ	0.99
Exploration fraction of Training Turns	0.99
Exploration Scheme	ϵ-greedy
ϵ start	0.6
ϵ end	0.0
Training frequency (w.r.t. dialog turns)	3
Training start (w.r.t. dialog turns)	1280
DDQN Target Network update frequency (w.r.t. training steps)	15
Q-Value clipping	10.0
Munchausen τ	0.03
Munchausen α	0.9
Munchausen Clipping	-1
Evaluation frequency (w.r.t. dialog turns)	10000
Evaluation dialogs	500

Table 3: RL Training Parameters

B Mental Models

In this section, we present both the Likert item responses and the full results of the content analysis.

B.1 Likert Responses

Table 4 shows user responses to the eight post-survey Likert items they were asked about their mental model of the dialog agent they interacted with. All questions in this table were on a 5-point Likert scale where 5 represented that the user fully agreed with the statement and 1 that they fully disagreed.

Although the linguistic adaptation did alter many of the ways users perceived their conversational partner, we were not able to find significance for these at our current sample-size. The only effect which we found to be significant was in the FAQ condition, where users perceived their dialog partner as significantly ($p<0.01$; Wilcoxon-Mann-Whitney test) better able to give specific answers to their questions.

Mental Model	CTS	CTS+*ling_ad*	FAQ	FAQ+*ling_ad*	HDC	HDC+*ling_ad*
Understands natural language	3.24	**3.59**	3.45	**3.59**	2.45	**2.67**
Understands keywords	**3.71**	3.59	**3.52**	3.41	**4.05**	3.55
Needed a specific question	3.48	**3.91**	3.57	**3.76**	3.83	**3.90**
Could ask follow-up questions if my question was vague	3.81	**3.90**	1.83	**2.43**	**3.14**	2.90
Could only give general answers	**3.62**	3.48	**4.20**	3.71	3.73	**4.00**
Could give personalized answers	3.00	**3.52**	1.90	**3.45****	2.23	**2.45**
Could provide an immediate answer	3.62	**4.09**	4.00	**4.05**	3.09	**3.88**
Needed to ask multiple questions to answer	**3.65**	3.00	2.00	**2.23**	**3.86**	3.45

Table 4: Likert scores for post-interaction mental models users had in each type of dialog system. All questions were on a 5-point scale (1=don't agree at all, 5=completely agree) ** represents a significant difference ($p<0.01$) between the linguistic adaptation and baseline versions of an agent.

B.2 Content Analysis

In Table 5, we show the main and sub-categories of mental models users discussed in their free-form feedback. These models were elicited by asking users what they thought were strengths of the agent (things it could do well) and what they thought were weaknesses (things it struggled to do).

Table 5: Main and sub-categories resulting from the content analysis of user mental models of the agent. 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	It did better than my expectations (**Expectations Met**)
A2	I had higher expectations (**Expectations Not Met**)
B1	clear and concise answers to my questions (**High Quality Answers**)
B2	I received similar generic answers (**General Answers**)
B3	I believe it cannot give specific answers (**No Specific Answers**)
C1	Understanding the more complex/ not straightforward questions (**Could not Understand/Answer**)
C2	Reply to general questions (**Understood Common Questions**)
C3	it could understand some keywords in my messages (**Understood Keywords**)
D1	very fast as well (**Fast interaction**)
D2	it could have asked additional questions to gather more information about the inquiry before answering. (**Interaction too short**)
D2	two many claifying questions (**Interaction too long**)
E1	That chatbot could communicate very well (**Good Communication**)
E2	The chatbot does not do well in being friendly with its language. (**Unfriendly**)

A = Expectation match, **B** = Answer Quality, **C** = System Understanding, **D** = Dialog Interaction, **E** = Communication.

We found that user mental models fell into one of five main categories. Below we describe each category and show the counts of each sub-category code occurrence.

Expectations Match This category contains two codes which simply indicate whether the user felt their expectations were met by the system. The first code is "Expectations met/exceeded" the second code "Expectations not met" was created by merging two proto-codes "Expectations not met" and "Negative expectations met". The coders felt combining these led to a better understanding of which users were able to interact in a way they preferred. Figure 5 shows the frequencies of each code.

Figure 5: Counts from the content analysis for the sub-category occurrences for the category “Expectations Match”

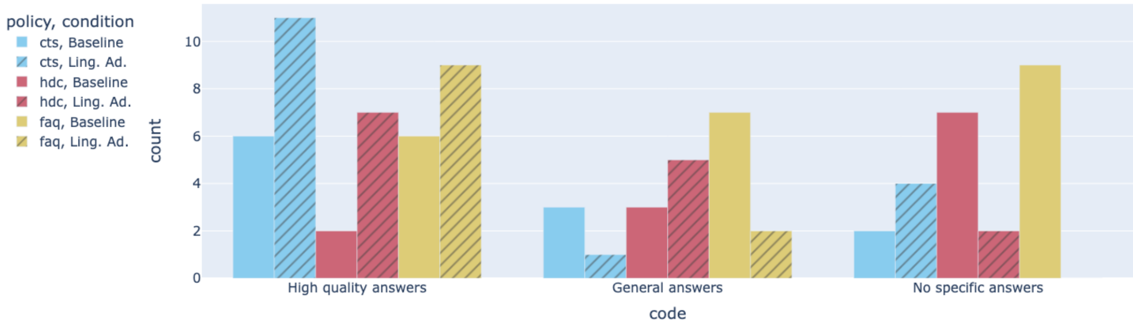


Figure 6: Counts from the content analysis for the sub-category occurrences for the category “Answer Quality”

Answer Quality This category contains all codes relating to how users perceived the system’s ability to answer. The first codes are “General answers” and “No specific answers” referring to users thinking the dialog agent was good at providing general answers and that it could not provide specific ones. The code “High quality answers” was created by combining the proto-codes “Precise,“ “Detailed”, and “Accurate” as the coders considered these all markers of satisfaction with the answers and none occurred frequently enough to be its own category. Figure 6 shows the frequencies of each code.

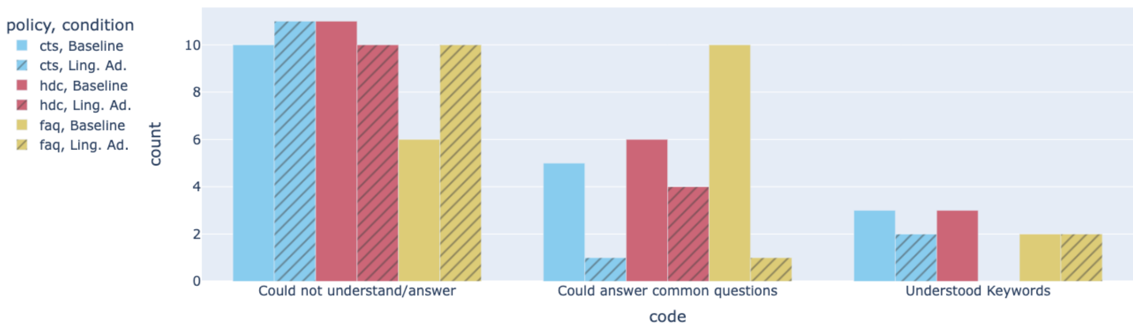


Figure 7: Counts from the content analysis for the sub-category occurrences for the category “System Understanding”

System Understanding This section focuses on codes relating to the user perception of what type of input the dialog system could understand. The codes here are: “Could answer common questions,” “Understood keywords,” and “Could not understand/answer”. Figure 7 shows the frequencies of each code.

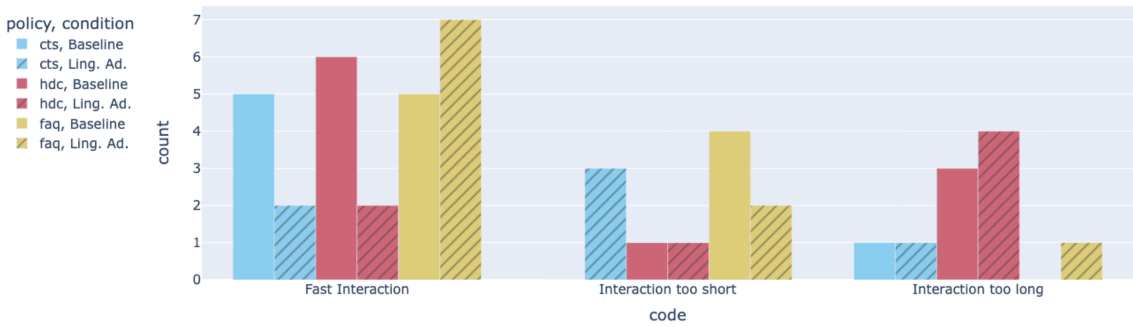


Figure 8: Counts from the content analysis for the sub-category occurrences for the category "Dialog Interaction"

Dialog Interaction This category focuses on codes relating to how the dialog interaction was perceived. Here we have the codes: "Fast interaction," "Interaction too fast," and "Interaction too long." It should be noted that "Fast interaction" was generally seen as a strength, which is why it is considered a separate code from "Interaction too fast." Figure 8 shows the frequencies of each code.

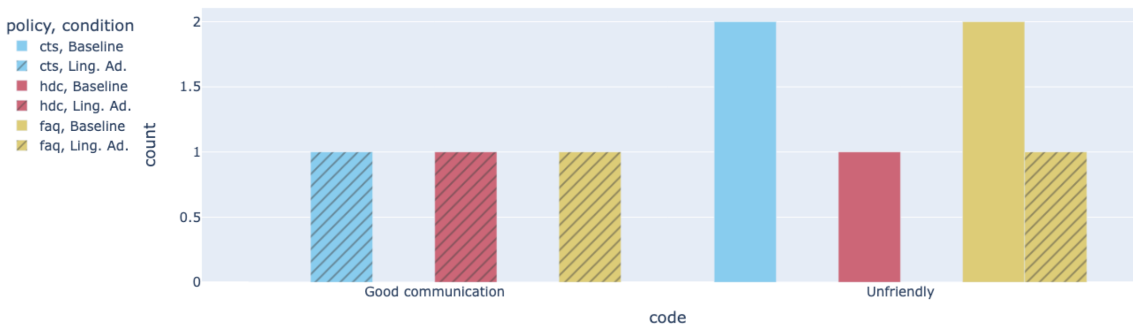


Figure 9: Counts from the content analysis for the sub-category occurrences for the category "Communication"

Communication This category contains code relating to how users perceived the communication abilities of their partner. The two codes here are: "Good communication" and "Unfriendly". Figure 9 shows the frequencies of each code.

C Trust, Reliability, and Usability

C.1 Likert Responses

In Table 6, we show the user responses to the Trust and Reliability subscales of the Trust in Automation questionnaire (Körber, 2018), which consist of six and two questions respectively, each rated on a five point Likert scale (1: strongly disagree to 5: strongly agree). We also show user responses to UMUX questionnaire (Finstad, 2010), which can range from 0 to 100.

Model	Avg. Trust	Avg. Reliability	Avg. Usability
CTS	**3.16**	2.96	63.83
CTS+*ling_ad*	3.05	**3.12**	**65.34**
FAQ	2.83	2.79	57.74
FQA+*ling_ad*	**3.07**	**2.98**	**59.66**
HDC	**2.61**	2.42	36.93
HDC+*ling_ad*	2.55	**2.47**	**41.96**

Table 6: Average user ratings for trust, reliability and usability. No significant differences were found.

C.2 Content Analysis

When analysing user free-responses about what aspect of the interaction they liked or disliked, we found the following codes, shown in Table 7.

Table 7: Main and sub-categories resulted from content analysis for user experience. 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 didnt like the interactions with the chatbot (**Liked Nothing)**
A2	There is nothing that I did not like about it (**Disliked Nothing**)
B1	The responses were easy to understand (**Easy to Understand**)
B2	Bot felt friendly. (**Liked Language Style**)
B3	The lack of personalisation. (**Impersonal**)
C1	It was easy to use and the experience was seamless (**Easy to Use**)
C2	It gives fast answers (**Fast**)
C3	too many steps to get to the answer I was looking for (**Interaction too Long**)
D1	it gave me precise answers to my specific questions (**Accurate Answers**)
D2	it didnt answer my questions (**Couldn't Understand/Answer**)

A = General Experience, **B** = Communication, **C** = Interaction, **D** = Agent Ability.

We found that user experience feedback fell into one of four main categories. Below we describe each category and show the counts of each sub-category code occurrence.

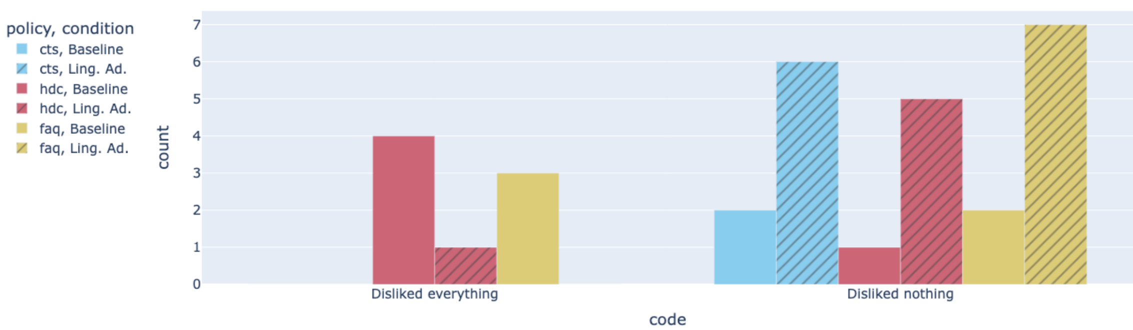


Figure 10: Counts of codes from the user experience content analysis in the category "General Experience"

General Experience This category relates to codes which provide general feedback on the user experience. Here coders found that "Disliked everything" and "Disliked nothing" came up frequently enough to deserve their own codes. Counts of the sub-codes in this category can be seen in Figure 11.

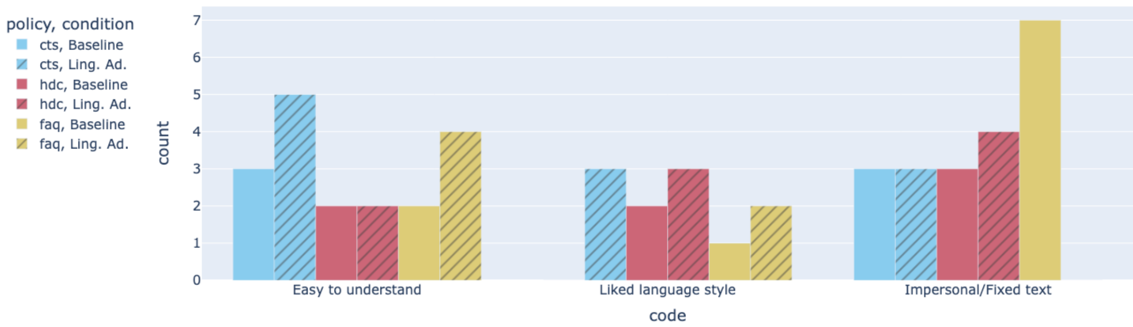


Figure 11: Counts of codes from the user experience content analysis in the category "Communication"

Communication Codes in this category relate to which aspects of the communication users liked or disliked. The first two codes are "Easy to understand" and "Impersonal/Fixed text". The last code, "Liked language style" combines the proto-codes "polite/professional," "robot/don't need a human," and "friendly language style". This code is interesting as although all users stated these as something they liked about the agent and it related to their perception of the language style, these represent different linguistic styles. Counts of the sub-codes in this category can be seen in Figure 11.

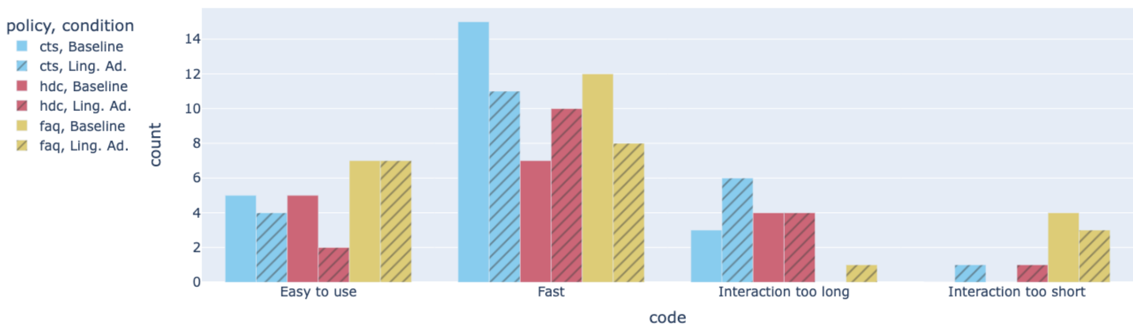


Figure 12: Counts of codes from the user experience content analysis in the category "Interaction"

Interaction Codes in this category represent how users viewed the experience of interacting with the dialog agent. Codes include "Easy to use," "Fast," "Interaction too long," and "Interaction too short." Although there is some overlap with mental models, these were listed separately as the question referred not to what abilities they thought the agent had, but only to what they liked/disliked about the interaction. Counts of the sub-codes in this category can be seen in Figure 12.

Agent Usability Codes in this section represent the aspects of agent capabilities that were mentioned as having a positive or negative impact on user experience. The codes here are: "Accurate" and "Could not understand/answer." Counts of the sub-codes in this category can be seen in Figure 13.

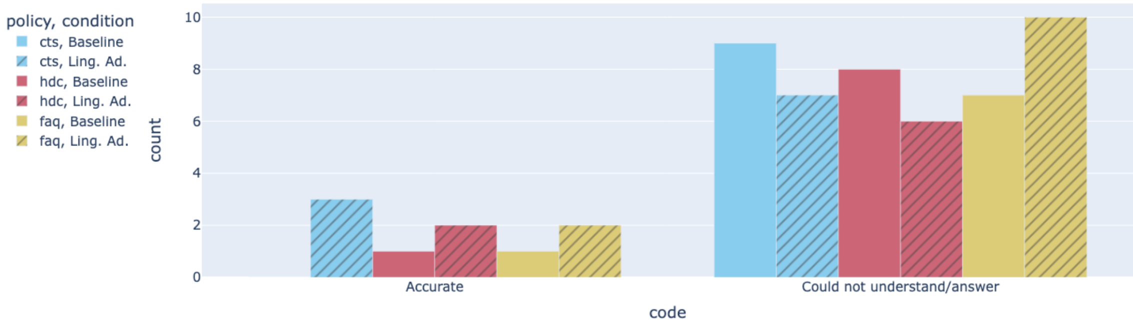


Figure 13: Counts of codes from the user experience content analysis in the category "Agent Ability"

D GLMM Analysis

D.1 Model Specifications

We fit a binomial model (logit link) to include both fixed effects of this study: behavioral and linguistics adaptation for objective success. For the subjective measures, we fit a categorical model using the same fixed effects.

We use R (R Core Team, 2021) and mgcv 1.8-42 (Wood, 2004, 2010; Simon N. Wood and Säfken, 2016; Wood, 2017) to fit all our models.

D.2 Objective Performance

	df	Chi.sq	p-value
dialog system	2	30.28	0.00
linguistic style	1	4.08	0.04

Table 8: Test statistics for behavioral and linguistic adaptation for the objective success model

Table 8 displays the test statistics regarding the dialog success model. We observe a statistically significant and positive effect of linguistic adaptation (beta = 0.46, 95% CI [0.01, 0.90], p = 0.043; Std. beta = 0.46, 95% CI [0.01, 0.90]). Compared to the behaviorally adaptive CTS agent, we also find a statistically significant and negative effect of the FAQ agent (beta = -1.19, 95% CI [-1.79, -0.60], p < .001; Std. beta = -1.19, 95% CI [-1.79, -0.60]) and of the HDC agent (beta = -1.64, 95% CI [-2.22, -1.05], p < .001; Std. beta = -1.64, 95% CI [-2.22, -1.05]).

D.3 Subjective Performance

	df	F	p-value
dialog system	2	7.13	0.00
linguistic style	1	4.91	0.03

Table 9: Test statistics for behavioral and linguistic adaptation for the subjective success model

Table 9 displays the test statistics regarding the subjective success model. We observe a statistically significant and positive effect of linguistic adaptation (beta = 0.41, 95% CI [0.05, 0.78], p = 0.027; Std. beta = 0.41, 95% CI [0.05, 0.78]). Compared to the behaviorally adaptive CTS agent, we also found a statistically significant and negative effect of the static FAQ agent (beta = -0.62, 95% CI [-1.06, -0.17], p = 0.007; Std. beta = -0.62, 95% CI [-1.06, -0.17]) as well as of the static HDC agent (beta = -0.85, 95% CI [-1.30, -0.39], p < .001; Std. beta = -0.85, 95% CI [-1.30, -0.39]).

	df	F	p-value
dialog system	2	16.60	0.00
dialog system:linguistic style	3	4.31	0.01

Table 10: Test statistics for behavioral and linguistic adaptation for the subjective length

D.4 Subjective Length

Table 9 displays the test statistics regarding the subjective length model. Here we find a statistically significant and positive interaction effect of the adaptive linguistic style on dialog system [faq] (beta = 0.81, 95% CI [0.18, 1.44], p = 0.012; Std. beta = 0.81, 95% CI [0.18, 1.44]). We also find, a statistically significant and negative interaction effect of the adaptive linguistic style on dialog system [hdc] (beta = -0.86, 95% CI [-1.59, -0.14], p = 0.020; Std. beta = -0.86, 95% CI [-1.59, -0.14]). Finally, we find a statistically significant and negative effect of dialog system [faq] (beta = -1.40, 95% CI [-2.05, -0.75], p < .001; Std. beta = -1.40, 95% CI [-2.05, -0.75]).

D.5 Mental Models: Answer Quality

	df	Chi.sq	p-value
agent	2	3.42	0.18
style	1	5.40	0.02

Table 11: Test statistics for behavioral and linguistic adaptation on the mental model "High Quality Answers"

Table 11 shows the test statistics for the presence of the mental model 'High Quality Answers'. Here we find a statistically significant and positive effect from linguistic adaptation (beta= 0.93, 95% CI [0.15, 1.71], p = 0.020; Std. beta = 0.93, 95% CI [0.15, 1.71]). We did not find a statistically significant effect of dialog agent type. From this, we see that varying the formality can increase user perception of the answer quality.

D.6 Mental Models: No Specific Answers

	df	Chi.sq	p-value
agent	2	1.00	0.61
style	1	7.20	0.01

Table 12: Test statistics for behavioral and linguistic adaptation on the mental model "No Specific Answers"

Table 12 shows the test statistics for the presence of the mental modal 'No specific answers', which occurred when users did not think the system was capable of generating answers specific to their exact scenario. We find a statistically significant and negative effect of linguistic adaptation (beta= -1.38, 95% CI [-2.38, -0.37], p = 0.007; Std. beta = -1.38, 95% CI [-2.38, -0.37]). We do not find a significant effect for dialog agent type. From this, we see that by varying formality, perceived the dialog agent as better able to give specific answers.

D.7 Experience: Disliked Nothing

Table 13 shows test results for the presence of the code 'disliked nothing' in content analysis of the usability free-response questions. Here we find that there is a statistically significant and positive effect of linguistic adaptation (beta = 1.49, 95% CI [0.43, 2.56], p = 0.006; Std. beta = 1.49, 95% CI [0.43, 2.56]). We did not find any significant effects for type of dialog system. From this, we see that regardless of dialog system, adapting formality can increase the overall user experience.

	df	Chi.sq	p-value
agent	2	0.65	0.72
style	1	7.56	0.01

Table 13: Test statistics for behavioral and linguistic adaptation on the usability code “Disliked nothing”

E LLM Prompts for Different Formality Levels

For the experiments recorded in this paper, we used gpt-40-mini-2024-07-18.

Formality Level	Prompt
4	{“role": “system", “content": "You are an assistant in the middle of a conversation. Your goal is to rewrite the original text of the current dialog turn in a more formal, professional way. Remember that you are the employer."}, {“role": “user", “content": f“{TEXT}; formal rewrite:"}
3	*BASELINE - used existing templates*
2	{“role": “system", “content": “You are a company assistant in the middle of a conversation. Your goal is to rewrite the complex and formal original text of the current dialog turn in a simple and understandable way. Be friendly and always personal. Remember that you are the employer."}, {“role": “user", “content": f“{TEXT}; personal rewrite:"}
1	{“role": “system", “content": “You are in the middle of a conversation. Your goal is to rewrite the complex and formal original text of the current dialog turn in a simple and understandable way. Be overly friendly and use casual language."}, {“role": “user", “content": f“{text}; Overly friendly Rewrite:"}

Table 14: Prompts used to generate generate different formality levels for system utterances, from the most formal (4) to least formal (1)

After experimenting with several prompts, we found that the most consistent results came from prompting the LLM to take on a certain persona, rather than specifying specific attributes that the different levels should have. The final prompts that we used for this experiment can be seen in Table 14. For formality level 3, we did not use any prompt, but rather directly used the system utterances from the REIMBURSE-EN (Väth et al., 2024) dialog tree.

F Dialog Examples

In this section, we show some example dialogs demonstrating different types of user goals, user interaction styles, and template language styles.

```
USER: cb8d48c57182851b8d62eb2400885f (POLICY: cts) (GOAL-TYPE: EASY)
GOAL TEXT: You want to know if you can get reimbursed if you reserve a seat for yourself on the train
GOAL NODE: 16363756478730906 Seat reservations are allowed for train travel
USER: can i get reimbursed if i reserve a seat for myself on the train?
SYSTEM: (NODE: 16363756478730906) You can totally reserve seats for your train trip! 😊🚆
DIALOG END: SUCCESS
DIALOG LENGTH: 2
SUBJECTIVE QUALITY: 1
SUBJECTIVE LENGTH: 2
```

Figure 14: An example of a dialog with formality level 1, here the user asks a direct question and gets a direct answer.

```
USER: 9c684c2c6abb202ee607c3e38df6d6 (POLICY: cts) (GOAL-TYPE: OPEN)
GOAL TEXT: You want more information about how to plan a research semester.
GOAL NODE: 16387868859695624 Thank you for using the business travel chatbot.Please let us know about your experience by filling
USER: I need some help planning my research semester
SYSTEM: (NODE: 16370487978253782) How long will you be doing your research semester?
USER:  3 weeks
SYSTEM: (NODE: 16353297472197738) Is this as part of a reallocation (changing your official work location)?
USER: No
SYSTEM: (NODE: 16370501863636503) Since there hasn't been any change to your workplace location (reallocation), your research
semester is considered a business trip.
SYSTEM: (NODE: 16353303012891903) Are you traveling by yourself or with your family?
USER: I will be travelling alone
SYSTEM: (NODE: 16370483570980906) If you need to get reimbursed for any business expenses, just fill out <a title="Form for travel
reimbursement" href="https://www.redacted.pdf">this</a> form.
SYSTEM: (NODE: 16363752591243585) What country are you headed to on your trip?
USER:  UK
SYSTEM: (NODE: 16363753123487784) Which city are you heading to?
USER:  Manchester
SYSTEM: (NODE: 16389695846089108) Just a heads-up: for any business trips outside the country to EU countries, Iceland, Norway,
Liechtenstein, or Switzerland, you'll need to fill out an A1 Form and take it with you. This applies even for short trips or day
trips. Remember, you need to fill out a new A1 Form for each trip you take. Filling it out once won't cover future trips.
SYSTEM: (NODE: 16363753477187975) Could you please check the latest COVID-19 travel warnings and restrictions from the foreign
ministry and the RKI? We generally don't allow business trips to high-risk or virus variant areas. However, in extreme cases, you
can get special permission from the leadership of Department 4 (Administrative Department). Thanks!
SYSTEM: (NODE: 16363754346769013) Let's cut down on business trips as much as possible. Only go if it's really necessary. The
leaders in the administrative department will decide if a trip is essential. Thanks!
SYSTEM: (NODE: 16348093549280346) Are you thinking about adding some personal travel time to your trip?
USER:  Yes
SYSTEM: (NODE: 16363754793760356) If you want to add some personal time to your business trip, just make sure it's reasonabe
compared to the length of your work travel. Work comes first! If you have any questions, feel free to reach out to the travel
department.
SYSTEM: (NODE: 16348097487038694) What would you like to learn more about?
USER: Accommodation in the area
SYSTEM: (NODE: 16378350683837525) If someone else (not the university) is paying for the trip, we can't reimburse you.
SYSTEM: (NODE: 16370471522263677) Where are you planning to stay?
USER: Hotel
SYSTEM: (NODE: 16363811629822245) Please make sure the bill is in our company's name, but list you as the guest.
SYSTEM: (NODE: 16365521324065600) In {{ COUNTRY }}, you can spend up to 90€ on accommodations. If you're in a city with more than
100,000 people or where hotel prices are usually high, you can spend up to 120€. Just make sure to keep and show us the paid
invoice.
SYSTEM: (NODE: 16363810595542638) Sometimes, if you need to spend more on a hotel, that's okay if there's a good reason. For
example, if there were no cheaper hotels close to the event, or if staying farther away would have meant spending more on
transportation.
SYSTEM: (NODE: 16365623179506925) You might have to pay resort taxes if you travel outside of Germany.
DIALOG END: SUCCESS
DIALOG LENGTH: 20
SUBJECTIVE QUALITY: 3
SUBJECTIVE LENGTH: 3
```

Figure 15: An example of a dialog with formality level 2, demonstrating a user who asks a vague question and is walked through the entire process of answering that question.

```
USER: 1c12c92ae12dfb635103a0d179258e (POLICY: cts) (GOAL-TYPE: HARD)
GOAL TEXT: You want to know how you can get reimbursed for your flight.<ul><li>You plan to extend your stay with private vacation
before flying back</li></ul>
GOAL NODE: 16363755463439219 Costs which occur as part of the business aspect of a trip are reimbursable. Therefore, you will nee
USER: Change return flight
SYSTEM: (NODE: 16460436532310883) What's going on? What kind of emergency are you dealing with?
USER: Personal reasons
SYSTEM: (NODE: 16460439592347465) If it's your mistake, you won't get your money back.
SYSTEM: (NODE: 16460439966919842) Hey, no worries! If it's not your fault, like if a meeting went long, you can get your money
back as long as you explain why.
SYSTEM: (NODE: 16348071239272239) Want to know more about booking a ride or the extra costs? 😊
USER: Can i get a voucher?
SYSTEM: (NODE: 16478798353898107) Are you thinking about adding some personal travel time to your trip? 😊
USER:  Yes
SYSTEM: (NODE: 16363755463439219) If you spend money on stuff that's just for the business part of your trip, you can get that
money back. So, make sure you get a comparison offer that shows how much the business part of the trip would cost.
DIALOG END: SUCCESS
DIALOG LENGTH: 7
SUBJECTIVE QUALITY: 3
SUBJECTIVE LENGTH: 3
```

Figure 16: An example of a specific goal with language style 1. Here the language use of “other fees” rather than “ancillary costs” helps the user understand which category extending their trip with private travel falls under.

```
USER: 52294614b51ef6bf0986ea3aad6788 (POLICY: cts) (GOAL-TYPE: HARD)
GOAL TEXT: You want to know how you can get reimbursed for your flight.<ul><li>You plan to extend your stay with private vacation
before flying back</li></ul>
GOAL NODE: 16363755463439219 Costs which occur as part of the business aspect of a trip are reimbursable. Therefore, you will nee
USER: I'm planning on travelling for business. Will my flight still be reimbursed by the company if i decide to stay for a couple
more days for a personal vacation?
SYSTEM: (NODE: 16478798353898107) Do you intend to extend your trip for personal travel purposes?
USER:  yes
USER: (POST-NLU) True
SYSTEM: (NODE: 16363755463439219) Expenses incurred as part of the business-related aspects of a trip are eligible for
reimbursement. Consequently, you are required to obtain a comparative offer that calculates the costs exclusively associated with
the business components of the trip.
DIALOG END: SUCCESS
DIALOG LENGTH: 3
SUBJECTIVE QUALITY: 4
SUBJECTIVE LENGTH: 4
```

Figure 17: An example of the same goal as in the previous dialog with formality level 4. Here the user knows exactly what they want, inputting a very precise question and after only one follow-up question is given the answer to their question.

G REIMBURSE-EN Dataset Statistics

The diagraph dataset consists of a dialog tree for the domain of business travel as well as user utterances representing either user questions (associated with a specific node in the tree) or user answers (associated with intents defined in the tree). The tree contains 123 nodes with a maximum depth of 32 and a maximum node degree of 14. On average, each node in the tree has 3.5 questions associated with it in the training set and 2.2 questions in the test set. Additionally, each expected answer (or edge) in the graph has 3.4 paraphrases associated with it in the training set and 2.2 in the test set. The statistics can be seen in Table 15.

Dataset	Split	#Nodes	Tree Depth	Max. Node Degree	#User Questions	Avg. User Questions	#Answer Paraphrases	Avg. Answer Paraphrases
REIMBURSE-En	Train	123	32	14	279	3.5	246	3.4
	Test				173	2.2	162	2.2

Table 15: Overview of original *REIMBURSE*, translated *REIMBURSE-En*, and newly created *ONBOARD* and *DIAGNOSE* datasets (numbers rounded to one decimal).

H RDMM Dataset Demographics

The dataset was created by recruiting 63 participants from the USA, UK, Australia, and Canada via the crowdsourcing website Prolific. In the end, there were 20 participants in the CTS group, 21 in the FAQ group, and 22 in the HDC group.

20 participants were male, 42 were female, and 1 person identified as other. Participant ages ranged from 20 to 69. On average, they had some familiarity with dialog systems (3 on a 5-point Likert item) and limited familiarity with business travel (1.9 on a 5-point Likert item). There were no statistically significant differences in the distributions of gender, age, or previous experience between the three conditions.

I User Study Materials

I.1 Data Agreement

At the beginning of the interaction we provided users with the following data agreement. Although we did not collect any personally identifying data, we wanted to ensure they understood what they would be asked to do, the purpose of the research, what data we would collect and how we would process that data.

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 [redacted]

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.

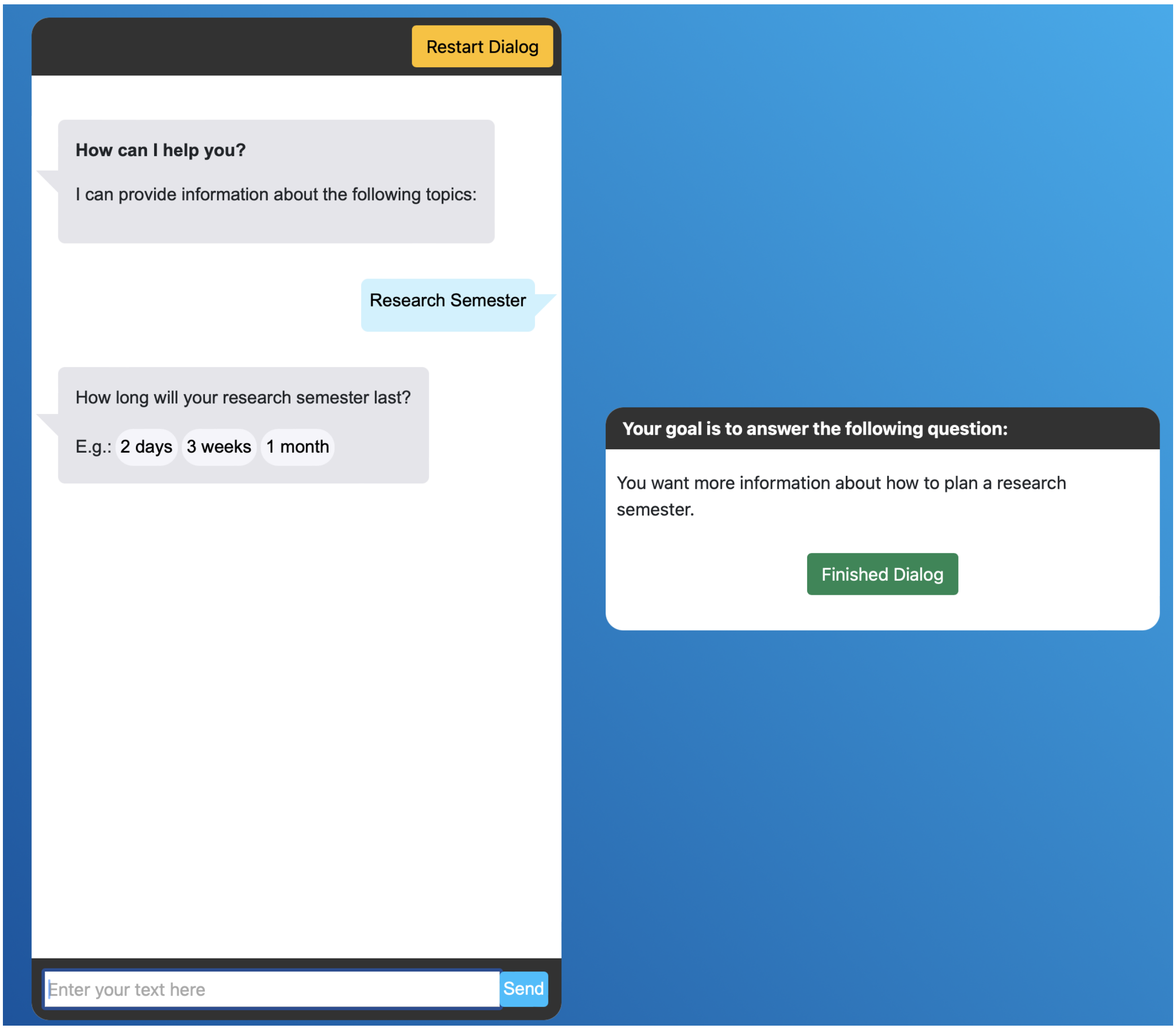
COMPENSATION: Upon completion of this task, you will receive a link to verify your completion with Prolific.

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

I agree

I.2 Dialog Interaction

During the main interaction, users were shown a split-screen view. On the left side, was the chat interface, where they could interact with the agent and see a history of the dialog so far. Here users also had the option to restart a dialog if they were not satisfied with the answer they received. On the right side, users were shown a box describing the information goal they were trying to achieve. Once they felt they had gotten an answer that matched this goal or given up on being able to do so, they could click on the button to move on to the next dialog.



I.3 Pre-Interaction Survey

Below are the questions asked in the pre-survey. We asked users to first provide general demographic information as well as their previous experience with chatbots and with business travel. We then asked users to provide some information about their mental models/expectations of an information seeking chatbot in general, both through the use of Likert questions and free-response questions. Here, we focused on how users expected to be able to interact with a chatbot and how they expected the chatbot to be able to respond to them.

Finally users were given examples of each formality level and asked to choose the one they found most appropriate.

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	○	○	○	○	○

Expectations of Chatbots

Please select the template that you would find most appropriate for a chatbot:

○ All flight reservations must be made through the university's designated Partner Travel Agency, which can be reached at +123456789. Booking through any other agency is not permitted. Additional details are available on the business travel website, where you will also find the contact information for the Partner Travel Agency.

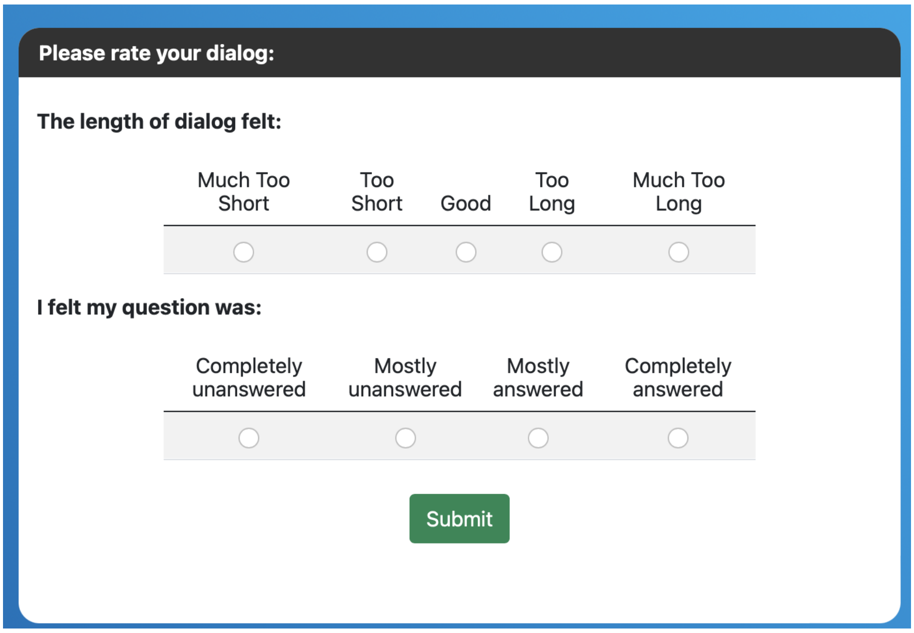
○ Flights must be booked through the university's Partner Travel Agency (Tel. +123456789). It is not possible to book through an alternative agency. You can find more information on the business travel website. Contact information for the partner travel agency can be found here.

○ You need to book your flights through the university's Partner Travel Agency. Just give them a call at +123456789. You can't use any other agency for this. For more details, check out the business travel website. All the contact info for the travel agency is there too.

○ Just a heads-up, you gotta book your flights through the university's Partner Travel Agency. Give them a call at +123456789. You can't use any other agency, though. For more details, check out the business travel website. All the contact info for the travel agency is right there. Happy travels! 😊✈️

I.4 Post-Dialog Survey

Once users indicated that a dialog was over, we then asked them to rate their perception of how well that dialog had gone. To do so, we used two Likert items, one asking for a rating of how long they perceived the dialog to be (5-point scale: 1 = much too short, 3 = good length, 5 = much too long) and how well they felt their question had been answered (4-point scale: 1 = not at all answered, 4 = completely answered).



I.5 Post-Interaction survey

Finally, at the end of the interaction, users were asked to complete the post-survey shown below. Here we again asked users to provide information about their expectations/mental models, this time specifically targeting the agent they had interacted with. To this end, users were asked to discuss how well the agent was able to meet their expectations, and provide insight into which were met and which were not. Users were then asked to fill out the same Likert questions about how they thought they could communicate with the chatbot. Finally they were asked to answer free-response questions describing what they thought the chatbot was good at doing and what they thought it could not do well.

We additionally asked users 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 about what they perceived positively and negatively about the experience.

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?