How Emotional Expressiveness Affects Trust Formation in a Conversational Decision Support System

L. Zhang

TU Delft

How Emotional Expressiveness Affects Trust Formation in a Conversational Decision Support System

by

L. Zhang

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Thursday July 6, 2023 at 10:30 AM.

Student number:	5567726	
Project duration:	September 1, 2022 – July 6, 2023	
Thesis committee:	Dr. Ujwal Gadiraju	TU Delft, Supervisor
	Dr. Luciano Cavalcante Siebert	TU Delft
	Gaole He	TU Delft, Daily Supervisor

An electronic version of this thesis is available at http://repository.tudelft.nl/.

TU Delft

Preface

This thesis represents the culmination of my journey towards the completion of my MSc in Computer Science at the Faculty of Electrical Engineering, Mathematics and Computer Science (EEMCS) at the Delft University of Technology. I went through a 10-month period filled with challenges, learning, growth, and immense gratification. I am filled with gratitude as I present this document, my final work, which is the result of rigorous academic pursuit and continuous dedication.

First and foremost, I would like to express my deepest gratitude to my esteemed professor, Ujwal Gadiraju, who gave me a lot of freedom and insightful guidance throughout this process. Besides, my daily supervisor Gaole He has been a beacon of support, providing me with academic knowledge, as well as writing tutoring.

This thesis marks not an end, but a checkpoint in my lifelong journey of learning.

L. Zhang
Delft, July 2023

Abstract

Trust is a fundamental component in human-AI relationships, serving as a critical element of user acceptance and satisfaction, particularly within the realm of Decision Support Systems (DSS). The technological advances in conversational user interfaces (CUIs) such as ChatGPT and digital assistants (*e.g.,* Alexa) allow laypeople to interact with DSS without knowing the mechanisms behind them. Extensive research has explored the benefits of CUIs and strives to improve their usability and adoption rate. However, while interacting with such CUIs, how to facilitate proper user trust for decision support is still under-explored. To address the research gap, we aim to test the impact of emotional expressiveness in CUIs on building user trust.

To analyze the impact of emotional expressiveness in CUIs to build user trust and whether voice-based CUI is more efficient in building user trust compared to text-based CUI. We implemented a conversational interface with varying emotional expressiveness that can serve six conditions: two text-based and four voice-based. Text-based CUIs are differentiated by lexical expressiveness. Voice-based CUIs are varying in both lexical expressiveness and prosodic expressiveness. Regardless of the modality and emotional expressiveness, each CUI serves as an interactive medium for users with the DSS, which supports them to find a suitable house given a scenario.

Through an empirical study (N = 151), the experimental results are insufficient to conclude the impact of prosodic expressiveness and lexical expressiveness on user trust and usability in CUIs. In addition, we did not find any statistically significant difference between text-based and voice-based CUIs in trust or perceived usability.

Our findings can potentially be explained by the uncanniness effect [46]: initially, increased emotional expressiveness in a CUI could positively influence user trust, but over time this could turn into a negative impact. These results offer a potential way to explain the complex dynamics of trust in conversational DSS and some implications in CUI design within the context of DSS. Our findings can benefit the future design and development of conversational agents-based DSS by considering emotional expressiveness.

List of Figures

List of Tables

Contents

1 Introduction

Trust serves as a pivotal component within the dynamic landscape of human-computer interaction. It governs user acceptance, influences their reliance on technology, and shapes the willingness to engage with it [15]. Trust becomes even more critical [30, 73], particularly in scenarios where human operators rely on these systems to navigate complex decisions, often with limited access to underlying data or amidst uncertainties. The need to understand factors that influence trust formation in DSS is paramount. In this work, we focus on decision support systems (DSS) — tools that aid in making informed and efficient decisions.

In the expanding digital landscape, artificial intelligence (AI) is experiencing rapid growth, particularly in the field of conversational AI. This includes chatbots and intelligent virtual assistants, with the global market projected to rise at a 22% compound annual growth rate (CAGR) from 2020 to 2025, reaching $ 14 billion [1]. The adoption rate of conversational digital assistants is predicted to double within the next two to five years [10]. One standout success is ChatGPT, which achieved 100 million active monthly users just two months post-launch, earning it the title of “fastest-growing consumer application in history” [33]. While a substantial 90% of surveyed participants were familiar with voice-enabled devices and 72% had used a voice assistant [51], their usage is predominantly confined to basic tasks (such as asking questions or checking the weather). However, when it comes to more complex tasks, trust manifests as an obvious barrier: only half of the respondents have engaged in purchases via their voice assistant. Extensive research [55, 57] has illuminated the benefits of conversational interfaces. For instance, [55] indicates that these interfaces enhance user engagement and working experiences. Similarly, another research [57] reveals their impact on improving the memorability of consumed information. However, trust formation in voice-based decision systems on critical tasks is still under-explored.

To address the research gap, we need to understand the factors that affect user trust in DSS and the best ways to promote trust formation. In this work, we mainly focus on the prosodic and lexical emotional expressiveness, and the modality of CUIs. Emotional expressiveness plays an important role in human-human interaction [36, 2, 27]. For example, Diel *et al.* [13] considered it as a technique to manipulate human likeness. Additionally, Zhu *et al.* [76] found that increased emotional expressiveness in a voice assistant yielded more positive perceptions, including increased engagement, human likeness, and likability. Besides emotional expressiveness, the modality of CUIs is also found to be of great impact. Rheu *et al.* [58] identified 5 critical design factors which influence trust towards conversational agents, including voice characteristics, communication style and etc. Based on existing work, we would like to go one step further and investigate the impact of the emotional expressiveness of CUIs on trust formation in DSS.

In our work, we examined the role that emotional expressiveness plays in the trust formation of CUI. Specifically, we examined the impact of integrating emotionally expressive words and prosodic patterns into our CUIs’ communicative style and voice characteristics. We utilize emotionally expressive words such as “incredible” and “wonderful”. Additionally, we increase the pitch range of speech to increase prosodic expressiveness. Both mechanisms are applied to convey enhanced emotional expressiveness in the conversational interface. In the experiment, each participant needs to identify appropriate housing within certain predefined constraints. The reason for adopting the housing selection task was

informed by the ongoing housing crisis prevalent in numerous nations [42]. Meanwhile, this task is considered as a critical decision-making task where a conversational agent can be useful [24]. We adopted the dataset from [24], which offered real-life housing options and associated scenarios. The primary expectation was for the participants to correctly identify a house meeting all the constraints outlined in their given scenario, which requires calibrated trust in the AI system.

We did not find any significant impact of prosodic expressiveness and lexical expressiveness on user trust and usability in CUIs. In addition, we did not find any statistically significant difference between text-based and voice-based CUIs in trust or perceived usability. Our findings can potentially be explained by the uncanniness effect [46]: initially, increased emotional expressiveness in a CUI could positively influence user trust, but over time this could turn into a negative impact. Many studies [53, 54, 69] have explored the relation between emotional expressiveness and the potential uncanniness effect. The above studies manipulate visual facial expressions to convey emotions. In contrast, our work conveys emotional expressiveness in speech, rather than facial cues.

The main contribution is that we provide a preliminary understanding of the impact of an agent's level of emotional expressiveness and modality on user trust and usability in AI-assisted decision-making.

1.1. Research Question

Our study investigates the impact of lexical and prosodic expressiveness of CUIs on trust formation through an empirical user study. We predict that enhancing the lexical and prosodic expressiveness will facilitate user trust and perceived usability. We also infer that voice-based CUI may have an advantage over text-based CUI when it is closer to human communication [49].

To address the aforementioned research gap, we aim to find answers to the following two research questions:

RQ1: How do prosodic expressiveness and lexical expressiveness of a conversational agent influence user trust and usability?

RQ2: How does a voice-based CUI differ from a text-based CUI in regulating user trust, and usability in AI-assisted decision-making?

1.2. Outline

The remainder of this manuscript is organized as follows. We first presented related literature to provide more context and background knowledge in section 2. Then we introduce the study design and experimental setup in section 3 and section 4, respectively. After that, we presented experimental results in section 5, and further discussed the findings and implications in section 6. Finally, we conclude and point out future directions in section 7.

2

Related Work

2.1. Trust in Human-Computer Interaction

Definitions and interpretations of trust vary across different contexts. Extensive research has been conducted on interpersonal trust [67, 71], human trust in automation [30, 39, 50], and the combination of both [40]. Our work focuses on measuring trust in a decision support system (DSS), which shares similarities with automation in terms of purpose. Both DSS and automation are used in situations where human operators have limited access to raw data concerning system states, leading to opacity or ambiguity for the human operator [40]. In essence, both DSS and automation frequently assist human operators in making critical decisions. Therefore, we adopt the definition of trust in automation that characterizes it as “the attitude that an agent will help achieve an individual’s goals in a situation characterized by uncertainty and vulnerability” [39].

The Human-Computer Interaction (HCI) community has also shown interest in studying trust in various systems over the years, including automation [39], intelligent systems [16, 31], AI, machine learning, and robotics [65]. [30] have developed a comprehensive model that incorporates trust variables. According to their model, trust in automation comprises three key elements: learned trust, situational trust, and dispositional trust. In our study, we focus on ’learned trust,’ which encompasses initial trust (formed based on the first impression of the system) and dynamic trust (which may evolve during interactions with the system).

2.2. Decision Support System

A decision support system (DSS) is designed to provide valuable insights into technological and management problems, assisting individual decision-makers in their reflective processes [35]. By analyzing problem variables, a DSS aids humans in making informed decisions.

DSS finds application in various scenarios. In the healthcare domain, clinical decision support systems (CDSS) have emerged as crucial tools for improving healthcare outcomes and reducing avoidable medical adverse events [75]. [32] explore how a DSS can leverage the wisdom of crowds to build knowledge bases, addressing the high cost of filling underlying information bases. Similarly, in the marketing domain, the utilization of a marketing decision support system (MDSS) has been shown to enhance the effectiveness of marketing decision-makers by helping them identify key decision factors, leading to better judgments [72].

A recent study by [24] compares the efficacy of conversational interfaces and traditional interfaces. Their research reveals that the conversational interface significantly enhances user trust and satisfaction in an online housing recommendation system compared to the traditional web interface. This study, however, identifies a research gap pertaining to trust formation in voice-based conversational interfaces within the context of DSS, which serves as the focal point of our research.

2.3. Role of Interfaces in Shaping User Trust

In our ever-evolving digital world, where reliance on technology is a fact of life, user trust is paramount. As highlighted in [20], interfaces that prioritize trust in their design can significantly enhance relation-

ships between users and digital agents. A recent example is ChatGPT, which achieved 100 million active monthly users just two months post-launch, earning it the title of “fastest-growing consumer application in history” [33]. Despite their growing prominence, the trustworthiness of CUIs remains a crucial concern, underlined by a range of user studies. The perceptions of trust in CUIs can manifest in various forms and can be influenced by a variety of factors. Inappropriate design choices of interfaces may result in distrust. For instance, Distrust may stem from perceived computational inadequacies or suspicions of malicious intent [61, 64]. Recognizing these divergent facets of trust and comprehending how they are shaped by interfaces can provide crucial insights for building user trust.

To further understand the role of interfaces in shaping user trust, [58] reveal five critical design factors influencing trust towards conversational agents: social intelligence of the agent, voice characteristics and communication style, the agent’s appearance, non-verbal communication, and performance quality. Similarly, [66] identified perceived usefulness, ease of use, and trust as critical elements for Chinese older adults’ adoption of Voice User Interfaces (VUIs). Moreover, [70] highlights that initial impressions and system accuracy are critical.[24] reveal that a text-based conversational interface is more efficient on building user trust and perceived usability compared to a conventional web-based interface in an online housing recommendation system. For autonomous vehicles, interfaces that simulate human characteristics, such as conversational interfaces, have proven effective in increasing people’s trust [59]. Additionally, using virtual assistants in explainable AI encounters increased trust in intelligent systems [74]. However, alongside building trust, avoiding over-trust is vital. This requires ethical design in the development of conversational agents, and being capable of implementing trust calibration techniques. The technique should reduce user trust in the agent when appropriate [14].

The work above explores the potential factors that interface affects user trust. In our research, while designing the interface, we focus on achieving a balance between perceived intelligence and usefulness to facilitate proper trust. On one side of the spectrum, an intelligent system should be able to respond accurately to user requests. Thus, we provide users some freedom to interact with the chatbot, thus cultivating the perception of an intelligent AI. Conversely, given that we require participants to follow certain procedures to find suitable accommodation, we must also emphasize guidance to prevent unintended behaviors. This is done to enhance perceived usefulness, although it may risk reducing the perceived intelligence of the interface. Our challenge was to design a system to help participants find the house efficiently while making them feel that the system was intelligent.

2.4. Role of Emotion in Human-Computer Interaction

Communication between humans often involves the expression of emotions through vocal and nonverbal cues, serving various purposes in interpersonal interactions [6, 21]. Emotional expressiveness can be considered as a form of expressive style [22], and its influence on human-human interaction (HHI) has been studied extensively [48, 4, 26].

However, the role of emotion in human-chatbot communication is not well understood. The most common emotion dimensions are activation (how ready one is to act), evaluation (how positive or negative, how much one likes or dislikes something) and power (or dominance/submission), which are also known as Arousal, Valence, and Control [62].

One challenge in human-chatbot interaction is how to convey emotion through a computer system. This is also related to affective computing(AC), which deals with the recognition and simulation of human affects (i.e. emotion) [7]. Recent research [76] examined the effect of emotional vocalization in its text-to-speech(TTS) output by varying prosodic and lexical expressiveness. Similarly, [63] found that the activation level of emotional states can be reliably communicated by increasing the overall pitch in TTS. Additionally, [3] found that the convincingness and intensity of emotion are strongly related to trustworthiness. Taken together, these findings imply that people experience emotion from computers in a manner that is comparable to how they perceive emotion from people. This idea is in accordance with the [47], which believes that when computers exhibit human-like traits, users adapt social behaviors from human-human interaction to computers.

However, previous work on whether a computer’s emotional expressiveness(e.g., emotional prosody and words) is positively received by a user during the interaction has shown mixed results. Some studies show that emotional expressiveness shown by a system leads to a positive reaction from users [9, 28]. Moreover, some work [11, 76] highlight the importance of increasing prosodic expressiveness in TTS synthesis to further improve the positive perception of the system. However, The presence of

misaligned cues (*i.e.*, the human-like component of emotion in a less anthropomorphized computer) may cause negative effects. Specifically, studies find that some systems exhibits a high degree of emotional expression [41, 68], leading to an uncanniness response [45]. A considerable amount of research has been conducted on the relation between human likeness to the effects of the uncanny valley and recognizes the multidimensionality of human likeness. Despite these insights, the exact nature of the relationship between human likeness and the uncanny valley curve remains unclear. [13] summarized that there are 10 stimulus creation techniques to change human likeness, including Emotion manipulation and 9 other categories. The most relevant technique to our research from these categories is emotion expressiveness. A majority of studies [53, 54, 69], have employed this method to manipulate visual facial expressions to convey emotions. In contrast, a recent investigation has revealed a direct link between prosodic expressiveness and human likeness [76] in a voice assistant. Similarly, Our work focuses on emotional expressiveness in speech, rather than facial cues. In summary, due to potential uncanniness efforts, participants might perceive emotional expressiveness negatively as we increase the emotional expressiveness in the chatbot.

Additionally, it is uncertain how lexical expressiveness and prosodic expressiveness interact. Previous research [49] in human-human interaction (HHI) suggests that emotional tone of voice made it easier for emotional words to be understood in a way that matched their emotions. Additionally, the interaction effect can be influenced by the experimental designs (between participants vs. within subjects). In [76], in between-subject design, both prosodic and lexical expressiveness shows impact on emotional expressiveness. Conversely, in within-subject design, only prosodic expressiveness shows impact on emotional expressiveness. One explanation is that the within subject allow participants to better detect the effect.

In the current study, we manipulated emotional expressions to a speech-based chatbot system across prosodic and lexical expressiveness to determine the effect of such manipulation on user trust. Here, we may predict that the effect of including emotional prosody may be further enhanced by the addition of emotionally expressive words. And they can be more effective together than building trust alone.

2.5. Conversational Crowdsourcing

Numerous studies have discussed the benefits of Conversational User Interfaces (CUI) for crowdsourcing [34, 38, 56]. For instance, [55] identified that an apt conversational style can considerably increase worker engagement. Furthermore, [43] demonstrated that conversational interfaces result in improved worker satisfaction without negatively impacting task completion time or the quality of work. Additionally, [29] developed Crowd Tasker, a platform utilizing a digital voice assistant for crowdsourcing tasks, thereby significantly reducing the time and effort required for task initiation while offering workers increased flexibility compared to a conventional web interface. In addition, [5] has utilized conversational agents to extract knowledge from crowd work to construct a knowledge base. Given these benefits of using conversational interfaces for crowdsourcing, we also apply the crowdsourcing platform to conduct our experiment.

3

Method and Hypotheses

In this section, we describe the house recommendation task and present our hypotheses.

3.1. House Recommendation Task

Participants are asked to imagine themselves as students, specifically for the purpose of a user study on house selection. In this task, participants will be presented with a scenario that describes the housing restrictions faced by a student involved in a simulated experiment. The conversational agent expects participants to engage in a conversation regarding their housing restrictions. Consequently, the interaction will be used to assess user behavior in a subsequent section. The choice of house selection as the task is inspired by previous research [24] that explored trust-related aspects. Detailed interaction between users and CUI is illustrated using an activity diagram(cf. Figure 7.1 in Appendix).

Quality Control Along with participation limits and incentives, we implemented additional quality control measures. As an added quality check, participants were presented with one warm-up question which validate the task ID assigned to them. Moreover, additional attention check questions, which asked participants to select a predetermined answer, were incorporated into both the pre-task and post-task questionnaires. The result that fails to answer any attention check will be considered invalid data input and removed. Despite these efforts, complete success cannot be guaranteed. Consequently, all participant responses will be personally reviewed, based on exported transcripts.

3.2. Dataset and Scenarios

We adopted the housing database and scenarios from [24], which aggregates housing options from reputable online platforms such as housinganywhere.com and kamernet.nl. Both of them are shared [1] for the benefit of the community. We retained 45 houses, removing 5 that were no longer available online. The selected properties each fulfill the following criteria:

1. **Housing type:** Studio, Apartment, and Room.
2. **Duration:** The length of contract.
3. **Rent:** Rental cost in euros.
4. **Supermarket proximity:** Whether or not the property is close to the supermarket.
5. **Registration availability:** Whether or not a tenant can register their address with the municipality.
6. **Commute time:** The commute time to university.

[1] https://drive.google.com/drive/folders/17nXjG6pr6shkCSl0DG3nb2qbl0perqkN?usp=sharing

We have 10 distinct task scenarios, each representing a student with specific housing preferences. Each scenario included four constraints, ensuring a consistent level of complexity across scenarios. An example is: "**Ahmed is starting his bachelor in Delft and is looking for an apartment. He is a bit lazy so he likes to be close to a supermarket and close to uni (commute less than 10min). He wants a contract of at least 24 months** ". In this scenario, users need to input 4 preferences: house type (*i.e.,* apartment), Supermarket proximity (*i.e.,* Yes), Duration (*i.e.,* 24 months), Commute time (*i.e.,* 10 minutes).

We utilized MongoDB Altas [2] for data collection and storage. This encompassed all the user behaviors, While user interactions in the tasks are stored as transcripts by the Voiceflow platform.

3.3. A Decision Support System for Housing

We developed a recommendation system for housing as part of our decision support system. We considered the impact of the system's accuracy on user trust and behavior. The system was designed to operate in two conditions: high accuracy and low accuracy. In the high accuracy condition, when the user correctly inputted all constraints, the system would recommend one correct house that fully satisfied those constraints. In contrast, in the low accuracy condition, the system would recommend one correct house along with five randomly chosen incorrect houses from the list of available options. It's worth noting that participants interacted with the decision support system through either two modalities (Voice vs Text). Both interfaces have the same interaction process, and the system allows easy switching of modality.

3.3.1. Interface

When users click the link on prolific to participate in the task, they first see a home page. On the home page, a simple user instruction (cf. Figure 3.1) is presented to them so that they have a rough idea of what kind of tasks they will encounter and decide whether to continue.

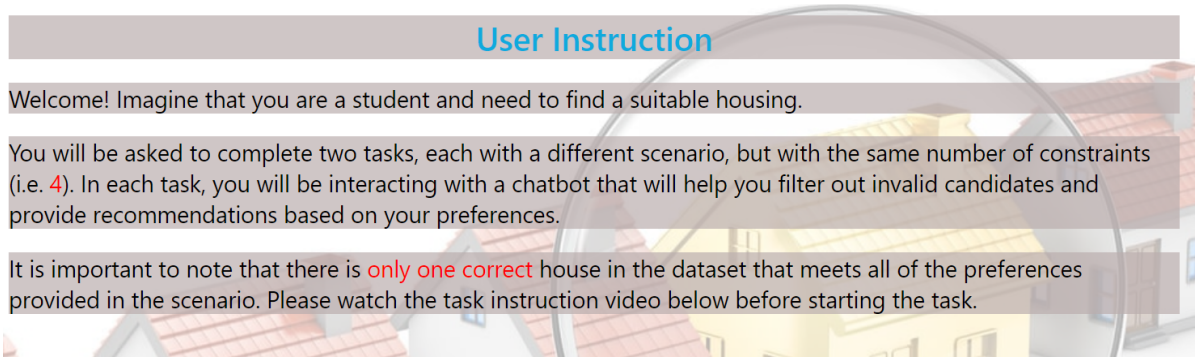

Figure 3.1: Home page instruction.

[2] https://www.mongodb.com/atlas/database

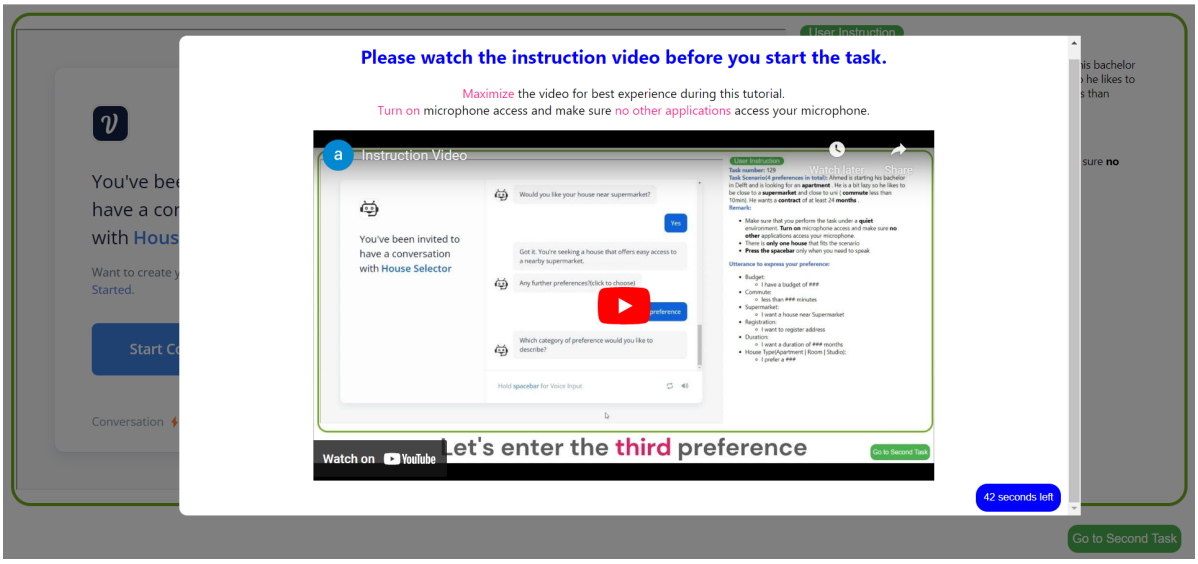

Figure 3.2: The instruction video to show how to interact with CUI. The timer at the bottom right (*i.e.,* 1 min duration) aims to prevent users from skipping the video.

After users choose to continue and finish the pre-task survey. They will see the instruction video page (cf. Figure 3.2). On this page, they need to watch videos that explain the details of the interaction. The rationale behind this is that we found it difficult for people without similar experiences to engage with the CUI during the pilot study.

After reading the instruction video, the task page (cf. Figure 3.3) is presented. The page includes the task number, task scenario (t1), remarks (t2), and utterances to express their preferences (t3). Remark emphasizes they need to perform the task in a quiet environment and allow microphone and speaker access. Utterance is shown here because we notice that it is difficult for CUI to capture the entity of users' expressions in the pilot study. So users can use this template to express their preferences more smoothly. Finally, when can click the "Start Conversation" (t4) button to start the task.

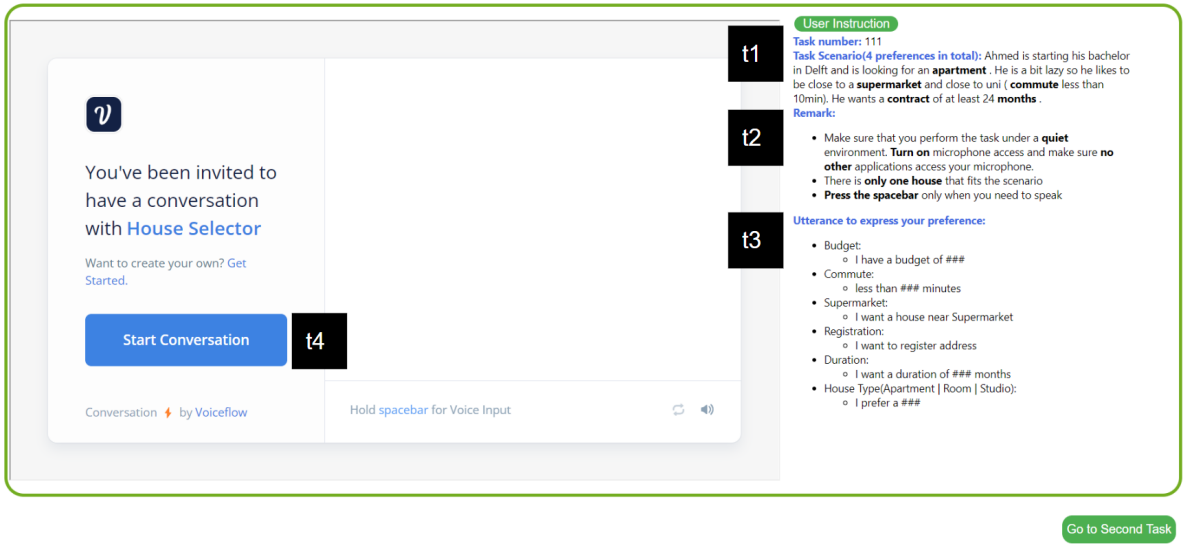

Figure 3.3: The task page which includes Voiceflow interface, task number, task scenario, and utterance to describe preferences.

After users click the "Start Conversation" button, they can start to engage with the CUI. Figure 3.4 provides an overview of the interaction. **First**, it welcomes participants and asks for the task number assigned to them to initiate the chat (a1). The conversational agent waits until the correct task number

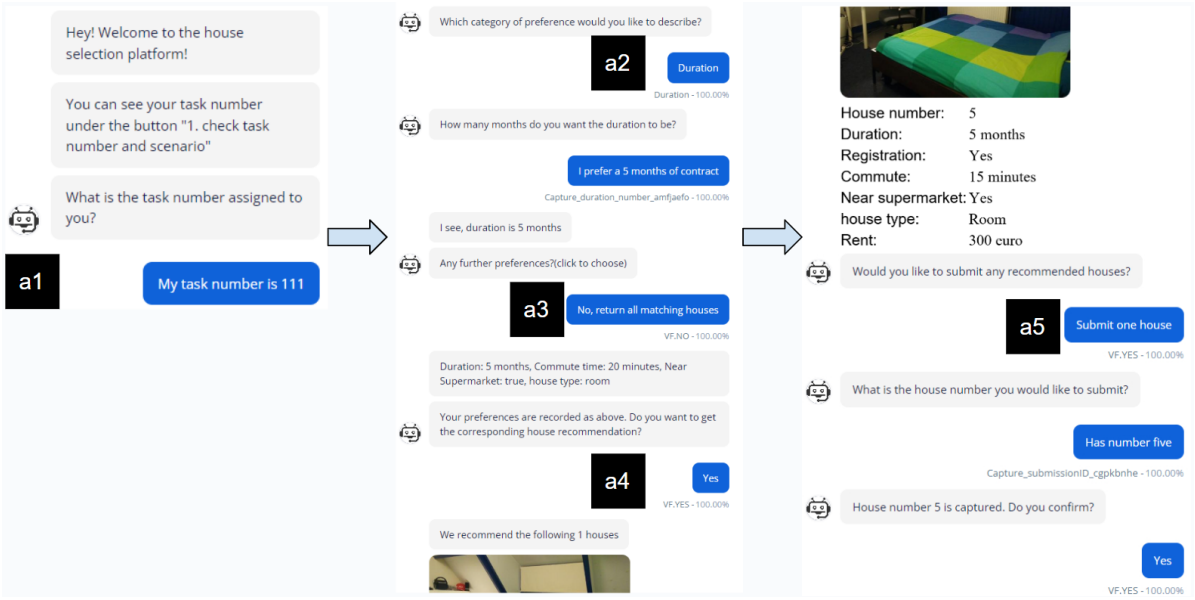

Figure 3.4: Conversational agent interactions.

associated with the scenario is entered, serving as both a warm-up question and an attention check for the participant. **Second**, the discussion continues, and users start to describe their preferences (a2), and this discussion will loop until users indicate they have no further preferences. Here, we designed that users have to input at least 2 preferences to have better recommendations. **Third**, when users have no further preferences to express (a3), then they must decide whether to receive recommendations or modify previously recorded preferences (a4). In a case that they choose to modify preferences, they will be redirected to (a2) to modify their preferences. Otherwise, we move on to the final stage. **Finally**, after the recommendations are present, participants can either submit the suggested housing option or explore all available houses within the system to make a selection(a5). Overall, the conversational agent encourages open-ended discussions. Only when the voice intent cannot be recognized, the option buttons are shown to participants so that they can click to proceed, ensuring ease of engagement. In case participants encounter a mistake brought by inaccurate speech recognition, they also have the option to reset constraints.

3.3.2. Implementation Details

Implementation Overview. This is a full-stack project. We implement a frontend webpage using ReactJS and host it on Netlify. Chatbots are created using Voiceflow and embedded in the webpage. Data and API are maintained with MongoDB Atlas. We can easily make UI and chatbot changes and deploy the changes to our hosted websites within a few seconds. Please find an architecture overview of the components used in the implementation in Figure 3.5.

Netlify. Netlify is a cloud service that hosts the ReactJS website and serverless function. The ReactJS website includes an embedded iframe that allows users to interact with the chatbot. When users engage with the website (*i.e.,* fill in a survey), HTTP requests are sent to the serverless function. The serverless function acts as a bridge and routes these requests to MongoDB Atlas.

Data Service. MongoDB Atlas provides the data service for the application. It includes two main components: the Data API and MongoDB. The Data API [3] endpoint handles MongoDB operations, such as reading or writing data, and acts as an interface for interacting with the MongoDB database hosted by MongoDB Atlas. This enables seamless communication between the serverless function and the database.

[3]https://www.mongodb.com/docs/atlas/api/data-api/

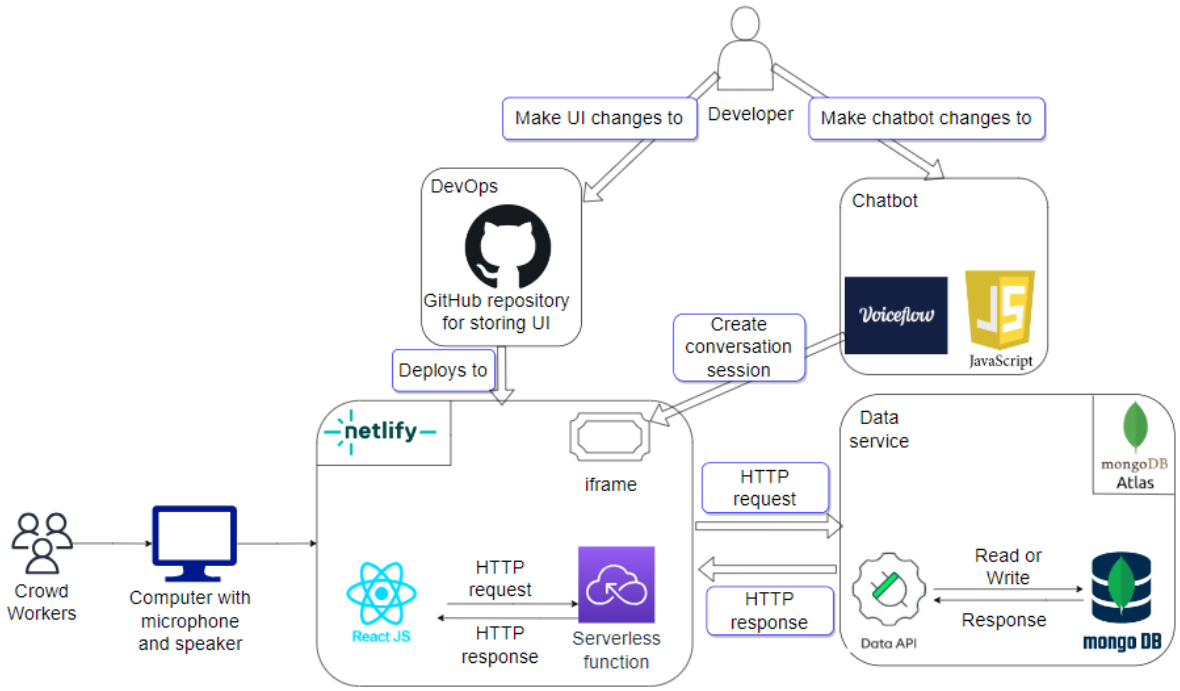

Figure 3.5: Architecture overview of the house selector.

DevOps. When user make UI changes to the ReactJS and push to GitHub. The Github will continusly integrate the changes and deploys to netlify.

Conversational agent. The Conversational agent is implemented using a low-code platform called "Voiceflow" [4], which utilizes JavaScript. With Voiceflow, developers can modify the chatbot logic and create conversation sessions. These conversation sessions are shared using a sharable link generated by Voiceflow. The chatbot is then embedded within an iframe on the website, allowing crowd workers to engage with it.

Voiceflow allowed us to build an interactive voice experience with various features and capabilities. One of the notable functionalities is the voice effects, including modifying volume, Emphasis, Speech, and Pitch. We adjust the pitch range to manipulate the prosodic expressiveness. In addition to voice effects, Voiceflow provided seamless integration with external APIs and allowed us to incorporate custom JavaScript blocks into the chatbot's logic. This enabled us to access and manipulate external data sources or perform complex operations as needed, extending the functionality of the chatbot beyond its basic capabilities. Another noteworthy feature of Voiceflow is its capability to train entities based on given utterances. Entities are essential for extracting important information from user inputs. To identify the "budget" entity, we trained the chatbot using the utterances such as "I don't have a budget limit", "The rent should be below xxx", and "I have a budget of xxx". This allowed the chatbot to accurately extract the budget keyword and the amount.

The excellent platform made the chatbot more versatile, engaging, and effective in interacting with users, ultimately improving the overall user experience. We would like to share the implementation of our chatbot [5].

3.4. Hypotheses

Our experiment was designed to answer questions surrounding the impact of prosodic and lexical expressiveness of CUI on user trust in AI-assisted decision making. [58]reveal that voice characteristics and communication style can influence trust towards CUI. Apart from that, [76] found that increased prosodic and lexical expressiveness shows a positive influence on the perception of voice assistants

[4]https://www.voiceflow.com/
[5]https://drive.google.com/drive/folders/1GI_l3a4eJ_EfqQomEGN0DEmHpyQ9gNtp?usp=sharing

in terms of human-likeness and emotional expressiveness. Based on existing work, We would like to go one step further and hypothesize that:

(H1a) Emotionally more-expressive prosody has a positive impact on building user trust.
(H1b) Emotionally more-expressive words have a positive impact on building user trust.

Zhu et al. [76] suggest that higher emotional expressiveness leads to higher human-like. Additionally, [8] found that anthropomorphic appearances and human-like conversational styles combined increased user satisfaction with chatbots. Thus, we hypothesize that:

(H1c) Emotionally more-expressive prosody has a positive impact on perceived usability.
(H1d) Emotionally more-expressive words have a positive impact on perceived usability.

Compared to text-based interaction, speech is natural and intuitive. [60] shows that speech interaction exhibits higher perceived efficiency, higher enjoyment, and higher service satisfaction than text-based interaction. Besides, voice-based CUI may have an advantage over text-based CUI as it is closer to human communication [49]. Thus, we hypothesize that:

(H2a) Voice-based conversational interface can build user trust more effectively than a text-based conversational interface.
(H2b) Voice-based conversational interface has higher usability compared to a text-based conversational interface.

4

Study Design

4.1. Experimental Conditions

A controlled crowdsourcing experiment was conducted using a 3×2 between-subject design. Table 4.1 provides an overview of the 6 conditions. For the voice-based chatbot, the independent variables consisted of lexical expressiveness (less-expressive word vs expressive word) and prosodic expressiveness (less-expressive prosody vs expressive prosody), resulting in four experimental conditions (*i.e.,* conditions 1, 2, 3, and 4, cf. Table 7.2 in the Appendix). For the text-based chatbot, lexical expressiveness was the only variable, creating two additional conditions (*i.e.,* conditions 5 and 6). Thus, 6 conditions were established in total (cf. Table 4.1):

	Independent Variables	Emotionally less-expressive Word	Emotionally Expressive Word
Voice-based CUI	Emotionally less-expressive Prosody	Condition 1	Condition 2
	Emotionally Expressive Prosody	Condition 3	Condition 4
Text-based CUI	-	Condition 5	Condition 6

Table 4.1: Experiment conditions across prosodic expressiveness, lexical expressiveness, and modality.

In each condition, participants were given two tasks of equal difficulty (*i.e.,* 4 constraints). The recommender system was programmed to provide correct recommendations for one task and incorrect recommendations for the other. The task ordering was equally divided among participants: 50% completed the accurate system scenario first, followed by the incorrect scenario, while the remaining 50% performed the tasks in reverse ordering. To ensure that each of the 10 scenarios was assigned approximately the same number of times, we implemented scenario counters. This counterbalancing helped to eliminate any potential biased effects.

To manipulate the emotional expressiveness of the conversational agent in each condition, we follow the design of [76]. This work alters the prosodic expressiveness and lexical expressiveness in the acknowledgment component of the system's outputs. As the bot engages in conversations with participants, it is crucial to maintain consistency of emotional expressiveness across different CUIs. To achieve this, we limited the emotion modification to the acknowledgment templates. This approach not only ensures uniformity among participants but also allows for the integration of emotional expressiveness in a more natural manner.

Prosodic expressiveness In general, prosody can be manipulated by changing the rate of speech, volume, and pitch. In our case, we applied SSML emotion tagging [1] which allows programmers to adjust the pitch ("x-low", "low", "medium", "high", and "x-high") of speech using a neural TTS technique [23]. To ensure that any effects we noted could be attributed unequivocally to the emotional expression without

[1] https://developer.amazon.com/en-US/docs/alexa/custom-skills/speech-synthesis-markup-language-ssml-reference.html

sounding too distorted, we created two distinct pitch levels to represent emotionally less-expressive and emotionally expressive prosody:

- **Emotionally less-expressive prosody**: ”low” pitch
- **Emotional expressive prosody:** ”high” pitch

Lexical expressiveness For lexical emotion manipulation, we selected words from the NRC Valence, Arousal, and Dominance (NRC-VAD) Lexicon that were suitable for acknowledging the user's response, thereby enhancing lexical emotional expressiveness [44]. This lexicon provides over 20,000 English words rated on a 0-to-1 scale for valence, arousal, and dominance. In our study, we chose to focus exclusively on the valence dimension of the NRC-VAD Lexicon, as it directly corresponds to the emotional polarity of the terms, signifying whether they are perceived as positive or negative. Our objective was to manipulate the emotional expressiveness of the chatbot, and we believed that the positivity or negativity of a term—its valence—would have the most direct and perceivable impact on this aspect. Hence, we form a collection of words (cf. Table 7.1 in the appendix) to represent emotionally less-expressive and emotionally expressive words:

- **Emotionally less-expressive words:** includes words with lower valence scores (e.g. “noted”, “understood”) and terms commonly used in chatbot interactions(“yea”, “sure”, and “ahhh”,) [9].
- **Emotionally expressive words:** words with valence scores above 0.9, which are perceived as more “positive”, and hence more emotionally expressive.

We have included several examples to illustrate the differences between lexically expressive and less expressive acknowledgment templates using the valence scores we obtained (cf. Table 7.1 in the appendix). The user interaction and perception studies subsequently employed these acknowledgment templates in chatbot interactions (cf. Table 7.2).

4.2. Measures

Trust. To measure participants' trust in the interface used to complete the scenarios, we used subscales Propensity to Trust (TiA-PtT) and Trust in Automation (TiA-Trust) from [37], which consists of 5 questions. The questions were answered using a 5-point Likert scale ranging from 1: Strongly Disagree to 5: Strongly Agree.

Affinity for Technology (ATI). We regard ATI as a moderator variable in our study. To gauge the extent of participants' affinity for technology interaction, we employ the ATI scale developed by [19]. This scale, though not essential to all research topics, will enrich our discussion section. It comprises a 9-item questionnaire with a 6-point Likert scale, measuring the enthusiasm of consumers towards embracing new technological systems.

Usability. To evaluate user usability and their inclination towards adopting the interfaces, we use a simplified model for expedient and straightforward assessment. This mode consists of 15 questions derived from the 'User-Centric Evaluation Framework for Recommender Systems' questionnaire [52]. The questions cover a range of facets, such as Quality of Recommended Items, Interaction Adequacy, Interface Adequacy, Perceived Ease of Use, Perceived Usefulness, Attitudes, and Behavioral Intentions. We obtain a 'usability score' for each response by averaging the scores across these parameters.

User behavior. We scrutinize participant behavior based on three criteria: accuracy, time spent, and the number of houses viewed. We examine the accuracy of the user's submission to determine whether the accuracy level of the decision support system—high or low—affects user behavior, given that each scenario has only one correct answer. To gain insight into how different interfaces might influence user behavior, we also track the time participants spend on active tasks and the number of houses they view.

4.3. Participants.

The study by [43] indicates that conversational interfaces can effectively be used for crowdsourcing microtasks, yielding high worker satisfaction without any negative impact on task execution time or job quality. Prolific, an online crowdsourcing platform, was utilized to recruit participants for user research. Participants were compensated at a rate of £7.50/h upon successful completion of the assignment. The requisite sample size was determined via a power analysis for an Analysis of Covariance (ANCOVA) using G*Power [17]. This analysis was based on the following assumptions: an effect size of 0.25 (indicating a moderate effect), a significance level (α) of 0.05 / 6 = 0.008 (adjusted for the testing of 6 hypotheses to control the family-wise error rate), and a statistical power (1- β) of 0.8 (suggesting an 80% probability of detecting a true effect if it exists). The number of groups (*i.e.,* 6), degrees of freedom (*i.e.,* 5), and the number of covariates (*i.e.,* 2) for each hypothesis were considered in the calculation. Based on these parameters, a total sample size of 305 participants was needed for the experiment. Considering the complexity of engaging with our conversational agent. We use the following pre-screening criteria:

- First language is English.
- Age above 18.
- A minimum approval rate of 90%.
- No prior participation in any other experimental condition.

Participants who met these criteria were permitted to participate in the research. An age limit was implemented due to legal considerations, and the language requirement ensured that participants were capable of comprehending the scenarios and effectively communicating with the bot.

4.4. Procedure

Upon obtaining participant consent, instructions specific to the experimental condition were provided, followed by directing participants to the appropriate interface. An overview of the workflow is depicted in Figure 4.1.

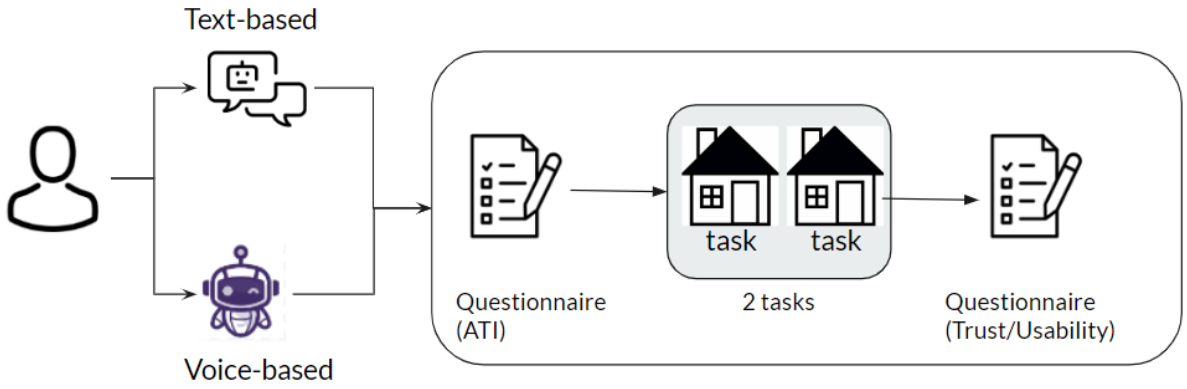

Figure 4.1: Workflow overview of the experiment procedure.

First, participants were evenly distributed across the six conditions (*i.e.,* two conditions are text-based, and four conditions are voice-based). Second Prior to commencing the tasks, participants need to agree on a consent form (cf. Consent Form in the appendix) and responded to a series of pre-task questions regarding their affinity to conversational agents. Subsequently, participants were assigned two tasks, each featuring a housing search scenario of comparable difficulty but varying AI accuracy. For both voice-based and text-based CUI, we prepare respective instruction videos explaining how to interact with the agent. (Check text-based[2] and voice-based[3] videos on Youtube). Upon task completion, participants were directed to a post-task questionnaire evaluating trust and usability in relation to the system's recommendations.

[2]https://youtu.be/PC-_rHG3tsI
[3]https://youtu.be/lmHy6xf1qq4

4.5. Data Preprocessing

We now have the knowledge to conduct the experiment after fully understanding the experimental setup. Data preparation must be done prior to evaluating the retrieved submissions. After all, before the gathered data can be translated into metrics that can be analyzed, it must be manually checked for quality control, evaluated by people for job performance, or changed from raw logs. The procedures followed before the findings are presented are described in this chapter. Overview of data collected:

- Total cost for recruiting crowd workers: £616.67
- Valid Submissions: 151
- median time of completion: 19.09 minutes
- Valid Submissions per group: 27, 26, 22, 24, 25, 27
- ordering for the system with high or low accuracy: 89 participants received the first system with high accuracy, followed by low accuracy. 90 participants received the opposite ordering.
- The tasks completed in each scenario are as follows: 21, 40, 35, 37, 39, 39, 40, 47, 30, 30. While the distribution is perfectly balanced, we ensured that all scenarios had similar difficulties by incorporating four constraints. This approach mitigates any potential bias effects.

Despite quality control, mistakes may happen. Thus, all submissions undergo manual quality checks. We provide the number of rejected submissions and reasons as follows:

- **Pilot study submission:** We had done a pilot study so we can reflect and improve usability based on the results. We need to remove their submissions. 10 submissions are removed.
- **Failed attention checks:** 7 submissions have been rejected. 4 submissions are rejected after a failed attention check. 3 individuals fail both attention tests.
- **Invalid submission:** 6 submissions are rejected. These participants input the incorrect completion code on Prolific (the method used to verify study completion).
- **Outliers:** 4 submissions were excluded. 3 users select all ”strongly disagree” for the TiA-Trust survey, and 1 user selects all ”strongly disagree” for the usability survey.
- **Operation issues:** 17 submissions are removed because the experiment variables are not properly modified.
- **Unexpected time spent:** 26 data sets are removed (each user contributed two data sets due to two tasks). The time spent on each task should fall within a range of 2 to 10 minutes, so we excluded times exceeding 10 minutes and falling below 2 minutes. Regarding the participant who spent more than 7 minutes, we manually checked their logs with the chatbot, and we confirmed that this unexpectedly long time was not due to a system lag.

As for the collected data, the data collection was exported from a MongoDB database containing the submitted arguments and answers from the pre-and post-task questionnaires. The raw data were processed using Jupyter Notebook which allows us easily view the data frame chances in every stage. The dataset containing the original analysis scripts (*i.e.,* Jupyter Notebook files) used to pre-process the data is published on Google Drive [4].

Statistic tests. After data are processed using Python, we need to select appropriate statistic tests. Table 4.2 presents an overview of all the variables considered in our study. Characterizing data types and value types prior to hypothesis testing is crucial as it guides us to find appropriate statistical tests[5].

[4] https://drive.google.com/drive/folders/1_X64LR85tewIk7hxDO10arIJEwqPARoL?usp=sharing
[5] https://timdraws.net/files/StatisticalTestFinder.pdf

Variable Type	Variable Name	Value Type	Value Scale
Emotional Expressiveness (IV)	Prosodic expressiveness Lexical expressiveness	Categorical Categorical	[low, high] [low, high]
Interaction (IV)	Modality	Categorical	[voice, text]
Trust (DV)	TiA-Trust	Likert	5-point, 1: strong distrust, 5: strong trust
Usability (DV)	Usability	Likert	5-point, 1: low, 5: high
User Behavior (DV)	Submission Accuracy Time Spent (mins) Submissions viewing all recommendations	Continuous, Interval Continuous Continuous, Interval	[0.0, 1.0] [0, 10] [0.0, 1.0]
Covariates	ATI TiA-PtT	Likert Likert	6-point. 1: low, 6: high 5-point. 1: strong distrust, 5: strong trust

Table 4.2: Variables considered in our experimental study. “DV” refers to the dependent variable. “IV” refers to the independent variable.

In the initial formulation of our hypothesis, we elected not to incorporate covariates due to two primary factors. Firstly, our preliminary data did not provide substantial evidence to warrant their inclusion. Secondly, our objective was to maintain the focus and simplicity of our hypothesis, thereby minimizing the risk of inflating the Type I error rate by overcomplicating the hypothesis. Table 4.3 shows a overview of statistical tests of hypotheses.

Hypothesis	Conditions	IV	DV	Statistical test
H1a H1b H2a	pairwise comparisons across all conditions (1 through 6)	Prosody Word	TiA-Trust	Two-way ANOVA Post-hoc tukey test
H1c H1d H2b	pairwise comparisons across all conditions (1 through 6)	Prosody Word	Usability	Two-way ANOVA Post-hoc tukey test

Table 4.3: Overview of the statistical tests and variables used to test the hypotheses.

However, as our study progressed and we collected more comprehensive data, the significance of two covariates emerged. Recognizing their potential influence, we elected to incorporate them into our subsequent hypothesis testing. It should be noted that this adaptation in our analysis strategy was motivated by the evolving understanding of our data set and the relevance of these covariates therein. We chose to reserve the discussion of these covariates and user behaviors for the discussion section of our thesis rather than amend our initial hypothesis. Thus, we took extra statistical tests as follows (cf. Table 4.4):

Covariates	IV	DV	Statistical test
PtT_trust ATI	Prosody Word	TiA-Trust	Two-Way ANCOVA
PtT_trust ATI	Prosody Word	Usability	Two-Way ANCOVA
	Prosody Word	Submission Accuracy	Kruskal-Wallis H Test
	Prosody Word	Time Spent	Kruskal-Wallis H Test
	Prosody Word	Submission Viewing All Recommendations	Kruskal-Wallis H Test
	Prosody Word	Number of Viewed Recommendations	Kruskal-Wallis H Test

Table 4.4: Overview of the follow-up statistical tests.

5

Results

In this chapter, we presented the descriptive statistics of the study data, hypothesis test results, and exploratory findings.

5.1. Descriptive Statistics

To ensure the quality of user response, we filter out participants who failed any attention check in our study. Finally, we have 151 valid submissions to test the hypotheses. All participants are balanced across six experimental conditions (cf. Table 5.1)

User Interface		Low word	High word	Total
Text-based interface		27	26	53
Voice-based Interface	Low prosody	22	24	46
	High prosody	25	27	52
Total		74	77	151

Table 5.1: The number of valid submissions per experimental condition.

The gender distribution was fairly balanced (male: 48.3%, female: 51.7%). The mean age of the participants was 36.68 years (*SD* = 13.59), with the youngest participant being 18 years old and the oldest participant being 77 years old.

In table 5.2, we show the descriptive statistics for the trust and usability scores across CUIs, varying in modality and emotional expressiveness.

User Interface	Prosody	Word	TiA-Trust ($M \pm SD$)	Usability ($M \pm SD$)
Voice-based Interface	low	low	3.41 ± 0.69	3.53 ± 0.56
	low	high	3.85 ± 0.51	3.91 ± 0.51
	high	low	3.30 ± 0.90	3.42 ± 0.61
	high	high	3.23 ± 0.72	3.68 ± 0.60
Text-based Interface	-	low	3.64 ± 0.95	3.84 ± 0.56
	-	high	3.57 ± 0.73	3.79 ± 0.52

Table 5.2: TiA-Trust and Usability grouped by interface type, prosodic expressiveness, and lexical expressiveness.

5.2. Hypothesis Tests

To address the hypotheses, we used an alpha level of .05. And the conditions highlighted below can be found in Table 4.1

H1a: *effect of prosodic expressiveness on building user trust.*

To analyze the main effect of prosodic expressiveness (*i.e.,* low pitch and high pitch) on building user trust, we need to compare the result from **Condition 1 to Condition 3**, and **Condition 2 to Condition 4**. A two-way ANOVA test was conducted. The result (cf. Table 5.3) showed significant effects of prosody ($p < 0.05$) on the TiA-Trust. Through a follow-up Tukey HSD tests, no significant difference is found between any pair of groups. This indicates that our experimental results do not provide any support to the impact of prosodic expressiveness on building user trust. Thus, **H1a** is rejected.

H1b: *effect of lexical expressiveness on building user trust.*

To analyze the main effect of lexical expressiveness (*i.e.,* less-expressive and expressive word) on building user trust. We need to compare the result from **Condition 1 to Condition 2**, **Condition 3 to Condition 4**, and **Condition 5 to Condition 6**. A two-way ANOVA result (cf. Table 5.3) shows no significant effect of word on building user trust ($p>0.05$), thus it is not necessary to conduct a follow-up comparison. **H1b** is rejected.

H1c: *effect of prosodic expressiveness on perceived usability*

To analyze the main effect of prosodic expressiveness (*i.e.,* low pitch and high pitch) on usability, we need to compare the result from **Condition 1 to Condition 3**, and **Condition 2 to Condition 4**. The two-way ANOVA results 5.4 show no significant effect of prosody on perceived usability ($p>0.05$). **H1c** is rejected.

H1d: *effect of lexical expressiveness on perceived usability*

Similar to **H1b**, To analyze the main effect of lexical expressiveness (*i.e.,* less-expressive and expressive word) on perceived usability, we need to compare the result from **Condition 1 to Condition 2**, **Condition 3 to Condition 4**, and **Condition 5 to Condition 6**. A two-way ANOVA test was conducted. The result (cf. Table 5.4) showed significant effects of word ($p < 0.05$) on usability. A post-hoc Tukey test(cf. Table 2.6) was performed. we found a significant difference ($p<0.05$) between condition **low prosody, high-expressive words** and **high prosody, low-expressive words**, indicating the usability score of the former was significantly higher than the latter. The interaction effect is partially supported. However, the results do not provide any direct support for the impact of lexical expressiveness on perceived usability, **H1d** is rejected.

H2a: *effect of modality on building user trust*

To analyze the main effect of modality (*i.e.,* voice vs text) on building user trust, we need to compare the result from **Condition 1,3 to Condition 5**, and **Condition 2,4 to Condition 6**, while the lexical expressiveness is controlled as ”low”. We can approximately consider text-based CUI to have prosody as ”none” to conduct a two-way ANOVA. The result (cf. Table 5.3) showed significant effects of prosody ($p < 0.05$) on the TiA-Trust. However, the follow-up Tukey HSD test shows no significant difference between any pair of groups. This indicates the experimental results do not provide any to support the impact of modality on building user trust. Thus, **H2a** is rejected.

H2b: *effect of modality on perceived usability*

Similar to **H2a**, to analyze the main effect of modality (*i.e.,* voice vs text) on perceived usability, we need to compare the result from **Condition 1,3 to Condition 5**, and **Condition 2,4 to Condition 6**, while the lexical expressiveness is controlled as ”high”. The result (cf. Table 5.4) showed significant effects of word ($p < 0.05$) on usability. However, the follow-up Tukey HSD test shows no significant difference between the text-based group and the voice-based group. Thus, **H2d** is rejected.

Factor	Sum of Squares	*df*	*F*	*p*
Prosody	4.047	2.0	3.360	0.037
Word	0.467	1.0	0.775	0.380
Prosody:Word	2.190	2.0	1.818	0.166
Residual	87.334	145.0	-	-

Table 5.3: Results of a two-way ANOVA on TiA-Trust against Prosody and Word.

Factor	Sum of Squares	*df*	*F*	*p*
Prosody	1.668	2.0	2.547	0.082
Word	1.487	1.0	4.540	0.035
Prosody:word	1.301	2.0	1.987	0.141
Residual	47.487	145.0	-	-

Table 5.4: Results of a two-way ANOVA on Usability against Prosody and Word.

In summary, all the hypotheses were rejected. Our experimental results do not provide any support to the impact of prosodic expressiveness or lexical expressiveness on building user trust and perceived usability. However, we found a significant difference (p<0.05) between condition **low prosody, high-expressive words** and **high prosody, low-expressive words**. The interaction effect is partially supported.

5.3. Exploratory Findings

5.3.1. User Behavior across CUIs

To further analyze the impact of prosodic expressiveness, lexical expressiveness, and CUI modalities, we compared participants' behaviors across experimental conditions. To this end, we considered participants' accuracy, efficiency (*i.e.,* time spent), round of interactions, and the number of viewed recommendations. Although no significant differences were found among different experimental conditions for these behaviors, we would like to share some findings from the result (cf. Table 5.5)

Submission accuracy: For voice-based CUI, We observe the biggest increase in submission accuracy is from 72.2% to 88% when the prosodic expressiveness rises from low to high as the lexical expressiveness is low. It showed the highest submission accuracy (88%) when both prosodic expressiveness and lexical expressiveness scores were high. In the text-based CUI, the submission accuracy in conditions with low and high lexical expressiveness scores was fairly close (84.0% and 81.5%), suggesting that lexical expressiveness does not play a significant role in both voice-based and text-based CUIs. Overall, text-based CUI (82.75%) has a higher submission accuracy than voice-based CUI (77.37%).

Time Spent: An observation is the overall higher interaction time on voice-based CUI (4.01 min) as compared to text-based CUI (3.47 min). This could potentially be attributed to the nature of interaction in voice-based CUIs, where participants are required to both talk to and listen from the CUI. This process could be more time-consuming than the typing and reading activities involved in text-based CUIs.

Submission viewing All Recommendations: In the voice-based CUI, the percentage of users who viewed all recommendations seemed to increase as prosodic expressiveness and lexical expressiveness increases from low to high. For instance, we can see an increase from the condition **low prosodic expressiveness and low lexical expressiveness** to condition **high prosodic expressiveness and low lexical expressiveness** and from **high prosodic expressiveness and low lexical expressiveness** to **high prosodic expressiveness and high lexical expressiveness**. In contrast, in text-based CUIs, more users appeared to view all recommendations in the condition with low lexical expressiveness compared to high lexical expressiveness. Overall, the percentage of submissions viewing all recommendations in text-based CUI (46.4%) is higher than in voice-based CUI (37.5%).

Number of Viewed Recommendations : This data indicates the total number of houses viewed by each participant in two tasks. Each participant will view at least 4 houses in total, and they can submit the right houses after viewing at least 7 houses in total. In the voice-based CUI, the percentage of users who viewed all recommendations seemed to increase as the emotional lexical expressiveness increases from low to high. For instance, we can see an increase from the condition **low prosodic expressiveness and low lexical expressiveness** (*i.e.,* 6.33) to condition **low prosodic expressiveness and high lexical expressiveness** (*i.e.,* 7.26) and from **high prosodic expressiveness and**

low lexical expressiveness (*i.e.,* 6.46) to **high prosodic expressiveness and high lexical expressiveness** (*i.e.,* 7). In contrast, in text-based CUIs, users appeared to view more recommendations in the condition with low lexical expressiveness (*i.e.,* 9.64) compared to high lexical expressiveness (*i.e.,* 7.81), and this is aligned with the Submission viewing of all recommendations. Overall, the number of Viewed recommendations in text-based CUI (*i.e.,* 8.73) is higher than in voice-based CUI (*i.e.,* 6.76).

User Interface	Prosody	Word	Submission Accuracy(%)	Time Spent(mins) (*M* ± *SD*)	Submission Viewing All Recommendations(%)	Number of Viewed Recommendations (*M* ± *SD*)
Voice-based Interface	low	low	72.2	4.09 ± 1	66.67	6.33 ± 1.47
	low	high	72.2	3.83 ± 0.95	74.07	7.26 ± 2.89
	high	low	77.08	3.9 ± 0.88	79.17	6.46 ± 1.08
	high	high	88.0	4.21 ± 1.08	84	7 ± 2.68
Text-based Interface	-	low	84.0	3.3 ± 1.34	92	9.64 ± 8.56
	-	high	81.5	3.64 ± 1.47	81.48	7.81 ± 4.01

Table 5.5: Conversational interface user behaviour analysis.

5.3.2. Further Analysis on TiA-Trust

In addition to the role of prosodic expressiveness and lexical expressiveness on trust. In this section, we further explore the impact of interaction effect and covariates (*i.e.,* ATI & TiA-PtT). To illustrate the impact of prosodic expressiveness, we showed the interaction plot in Figure 5.1. It is worth noting that, the two lines based on different lexical expressiveness are not parallel. The interaction effect indicates that high lexical expressiveness combined with low prosodic expressiveness yielded the highest trust, and conversely, high lexical and prosodic expressiveness resulted in the lowest trust. We found that changes in prosodic expressiveness (from low to high) substantially affect trust levels, regardless of the level of lexical expressiveness. Thus, it suggests that prosodic expressiveness has a substantial impact on the relationship between TiA-Trust and lexical expressiveness.

In our subsequent analysis (cf. Table 5.6), we conducted a two-way Analysis of Covariance (ANCOVA) to control for the effects of **TiA-PtT** and **ATI** after finding non-significant effects from the initial two-way ANOVA. The results of the ANCOVA indicated that neither prosody expressiveness levels nor lexical emotional expressiveness significantly influenced the user's TiA-Trust when controlling for **TiA-PtT** and **ATI**. Additionally, the interaction between prosody expressiveness and lexical emotional expressiveness also failed to reach statistical significance. In contrast, both **TiA-PtT** (F= 115.2, p< 0.001) and **ATI** (F= 11.03, p= 0.001) were found to be significant covariates, exerting a substantial impact on the **TiA-Trust**. We used Spearman's Rank Correlation to test the relationships between TiA-trust, TiA-PtT, and ATI. We found strong positive correlations between TiA-trust and TiA-PtT (r(149) = 0.667, p< 0.001) and between TiA-trust and ATI (r(149) = 0.334, p< 0.001).

Thus, after controlling for these variables, the primary factors of interest - prosody and word expressiveness - did not demonstrate a significant effect on **TiA-Trust** in the AI-assisted decision-making context.

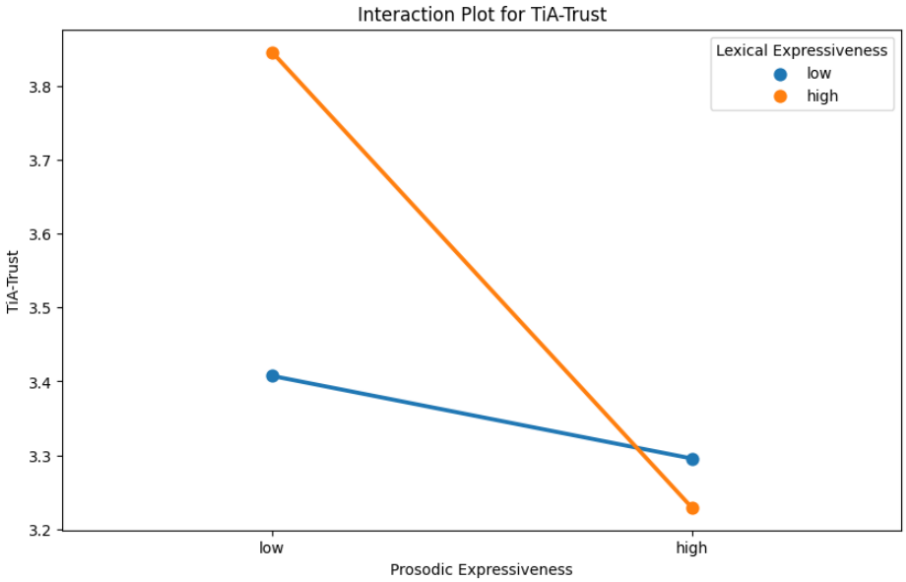

Figure 5.1: Interaction Plot for TiA-Trust.

Factor	Sum of Squares	*df*	*F*	*p*
Prosody	0.519	2.0	0.862	0.425
word	0.001	1.0	0.000	0.979
Prosody:word	1.794	2.0	2.979	0.054
TiA-PtT	34.69	1.0	115.2	<0.001
ATI	3.323	1.0	11.03	<0.001
Residual	43.07	143.0	-	-

Table 5.6: Two-way ANCOVA on TiA-Trust.

5.3.3. Further Analysis on Usability

In this interaction plot 5.2, the lines are parallel. This indicates no interaction effect occurs.

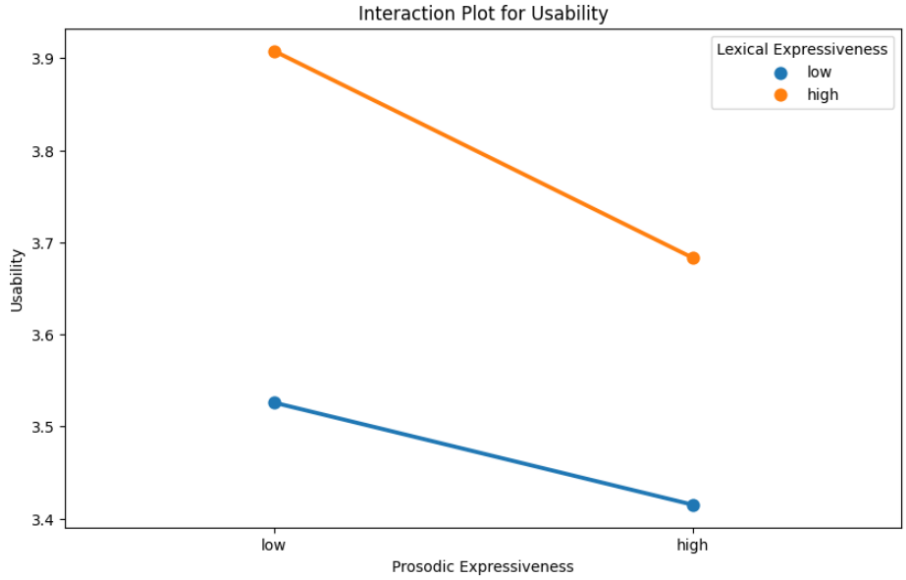

Figure 5.2: Interaction Plot for usability.

To further analyze the impact of covariates on usability, a two-way Analysis of Covariance (ANCOVA) (cf. Table 5.7) was adopted to control for the potential impact of **ATI** and **TiA-PtT** on usability, following our initial finding of the significance of lexical expressiveness from the two-way ANOVA. Upon examining the ANCOVA results, neither prosodic expressiveness nor lexical expressiveness significantly influenced the user's perception of usability, once **ATI** and **TiA-PtT** were taken into account. Similarly, the interaction effect between prosody expressiveness and lexical emotional expressiveness was not significant. On the other hand, the covariates **ATI** (F= 18.732, p< 0.001) and **TiA-PtT** (F= 31.229, p< 0.001) were found to exert substantial influence on usability, both reaching a highly significant level. We used Spearman's Rank Correlation to test the relationships between usability, **TiA-PtT**, and **ATI**. We found strong positive correlations between usability and **TiA-PtT** ($r(149) = 0.466$, $p< 0.001$) and between usability and **ATI** ($r(149) = 0.402$, $p< 0.001$).

Therefore, after controlling for these covariates, our primary variables of interest - prosody and word expressiveness - did not show a significant impact on usability. These results underline the importance of these covariates in the perception of the usability of a conversational agent.

Factor	Sum of Squares	*df*	*F*	*p*
Prosody	0.205	2.0	0.444	0.642
word	0.493	1.0	2.134	0.146
Prosody:word	0.639	2.0	1.382	0.254
ATI	4.329	1.0	18.732	<0.001
TiA-PtT	7.217	1.0	31.229	<0.001
Residual	33.045	143.0	-	-

Table 5.7: Two-way ANCOVA on usability.

5.3.4. Ordering Effect

The study design incorporated two task orderings to mitigate the impact of system accuracy on participants' responses. In the first task ordering, participants were initially presented with a task featuring immediate correct recommendations. Subsequently, they encountered a task with 3 incorrect recommendations, followed by the option to click the "view more recommendations" button for access to 3 additional houses, one of which contained the correct house. The second task ordering followed the opposite order. Importantly, the occurrence of each task ordering was evenly distributed. This investigation aimed to explore the influence of task ordering on the **TiA-Trust** and **usability** of participants.

An independent samples t-test (cf. Table 5.8) was conducted to compare the responses of participants who received task ordering 1 (n = 72, M = 3.49, SD = 0.84) and participants who received task ordering 2 (n = 79, M = 3.52, SD = 0.74). The results revealed no significant difference between the two groups ($t(149) = -0.20$, $p = 0.84$). Similarly, an independent samples t-test was conducted to compare the responses of participants who received task ordering 1 (n = 72, M = 3.65, SD = 0.58) and participants who received task ordering 2 (n = 79, M = 3.74, SD = 0.59). The results revealed no significant difference between the two groups, $t(149) = -0.20$, $p = 0.84$.

Therefore, it can be concluded that the **Usability** and **TiA-Trust** of the two groups are not biased by task ordering effect.

	System accuracy in task 1	System accuracy in task 2	TiA-Trust ($M \pm SD$)	Usability ($M \pm SD$)
task ordering 1	High	Low	3.49 ± 0.84	3.65 ± 0.58
task ordering 2	Low	High	3.52 ± 0.74	3.74 ± 0.59

Table 5.8: Task ordering effect analysis.

5.3.5. Compliance with Scenario

Our experiment involved 10 scenarios with 4 constraints each. Participants had to provide at least 2 inputs to receive recommendations from the system. If the inputs were less than 4, the system prompted them to complete all preferences for better results. The mean number of inputs across all

groups was above 7, indicating high compliance with the system's suggestions. However, in the "high - low" condition, one outlier entered 10 preferences in two tasks, skewing the mean (8.04).

Table 5.9 shows the mean and standard deviation for the number of inputs for each condition. Overall, the number of participants' inputs in the text-based CUI (*i.e.,* 7.77) was close to that of the speech-based CUI (*i.e.,* 7.87). Both numbers are close to 8. Thus, people are willing to input all constraints. To further analyze the difference, we conducted Kruskal-Wallis H Test to examine the differences among the 6 conditions. The result $H(5) = 5.15$, $p = 0.40$ indicates no significant differences among the conditions. Thus, we may conclude that the willingness to enter constraints is not influenced by emotional expressiveness or the modality of interaction.

Interface	Prosody	Word	Number of Inputs ($M \pm SD$)
Voicebased Interface	low	low	7.74 ± 0.80
	low	high	7.89 ± 0.83
	high	low	8.04 ± 0.45
	high	high	7.8 ± 0.49
Textbased Interface	-	low	7.8 ± 0.80
	-	high	7.74 ± 0.75

Table 5.9: Mean and Standard deviation of number of inputs per condition.

5.3.6. Qualitative Feedback Analysis

In total, we have 29 valid responses from participants, which can be used to analyze user opinions and feedback to our CUIs. We manually annotated the user sentiment (positive / negative) and reasons for the sentiment (cf. Table 5.10). The positive sentiment encapsulated the users' intrinsic interest and the enjoyment they derived from the system's utilization. On the contrary, negative responses were predominantly driven by concerns over voice recognition, user interaction speed, system flexibility, and an inherent distrust of chatbots. We present exemplar feedback (cf. Table 5.11) to provide a more comprehensive understanding of the user experience.

Although the voice-based CUI (*i.e.,* 98) has nearly twice the number of participants compared to its text-based counterpart (*i.e.,* 53). Out of the responses from the voice-based CUI, 17 conveyed negative sentiments against a mere 2 positive sentiments. In contrast, the text-based CUI yielded 5 positive sentiments against 7 negative sentiments. These results might suggest that, despite having fewer participants, the text-based interface could be more favorably received than the voice-based CUI.

Sentiment	Reason	Voice-based CUI	Text-based CUI	Total
Negative sentiments	Voice recognition	8	-	22
	Interaction Speed	5	-	
	control and flexibility	2	3	
	Distrust of chatbot	2	2	
Positive sentiments	interesting, enjoyment	2	5	7
Total		19	10	29

Table 5.10: Feedbacks per reason under voice-based and text-based CUIs.

Sentiment	Reason	Participant Feedback
Positive	Interesting	(1) Really interesting system! (2) It was an interesting task. Thank you.
	Enjoyment	(1) it all seems very straight forward! (2) Hello - just completely your task and just wanted to say that I really enjoyed it! :)
Negative	Voice Recognition	(1)Doesn't pick up certain accents (2)the system was hard to use, it kept getting my words wrong. it was frustrating
	Interaction Speed	(1)It was fine but the system was slow. I'd expect it to be faster than using filters provided in a drop down menu. (2)I didn't feel in control of the process and it felt slow and awkward.
	Control And flexibility	(1)didn't like the mix of buttons to select and voice options, would have preferred one or the other. (2)I would like to input more than one criteria at a time
	Distrust of Chatbot	(1)I usually like to visit somewhere in person or see a video. I think it's a good tool to narrow down choices of places to live but not to actually go ahead and sign a contract.People may not trust the system. (2)The system is beneficial but I would still be hesitant about trusting the info, input data can be captured incorrectly and the system will not know if there were errors in the user input

Table 5.11: Example of participants' responses Regarding the usability and trust of the system.

A detailed analysis of reasons for the negative sentiments that the voice-based Computer User Interface (CUI) may have some significant areas that need improvement. Firstly, many users reported issues with the **voice recognition** feature, particularly in recognizing accents. This indicates that the system may not have a comprehensive enough dataset to accurately understand and respond to diverse accents even though our participants all have English as their first language. The voice-based CUI may need to improve its ability to understand different accents, making it more accessible and user-friendly for a broader range of users globally. Secondly, users expressed dissatisfaction with the **speed of interaction**, indicating that the system is not as responsive as they expect. This could involve delays in processing voice commands or providing output. Nowadays, users expect near-instantaneous responses, and any obvious delay can lead to frustration and decreased overall usability. Thirdly, **control and flexibility**, shows that users desire a more streamlined and efficient user experience. For example, feedback suggests that users disliked the combination of selecting options via buttons and using voice commands. They would have preferred a consistent mode of interaction - either all voice or all button-based. This signals a need for the design to offer a more uniform interaction pattern. Lastly, 4 participants mention **Distrust** issue. Users are hesitant to rely on it for significant decisions and are concerned about potential data inaccuracies. The implications of qualitative analysis for designing interfaces for DSS are further discussed in the discussion section 6.4.

6

Discussion

This chapter discusses the limitations of our experiment and highlights the key conclusions of our experiment. To further explain our findings, we identified a potential cause — uncanny valley effect.

6.1. Potential Cause: Uncanny Valley Effect

From the result, we observe an increase in usability and trust from conditions **low prosodic and lexical expressiveness** to **low prosodic expressiveness and high lexical expressiveness** and a drop when shifting to **high prosodic and high lexical expressiveness**. Zhu *et al.* [76] identify that increased prosodic expressiveness can result in both higher human likeness and emotional expressiveness. This discovery points to a sequential hierarchy of perceived emotional expressiveness and human likeness given that we employed a similar setup for the manipulation of prosody and word:

> **"Less-expressive word and prosody" ≈ "Expressive word" < "Expressive prosody" ≈ "Expressive word and prosody"**

The envisioned scenario is that the trust score is in accordance with the same sequential hierarchy. However, Table 5.6 shows that the trust score increase and then decreases as the emotional expressiveness and human-likeness increase. Additionally, our results demonstrate a significant positive correlation between TiA-Trust and Affinity ($r(149) = 0.334$, $p < 0.001$). In an effort to encapsulate these findings, one potential explanation is the "uncanny valley" theory [46], which suggests that as an entity adopts an increasingly human-like appearance, its emotional response or affinity amplifies positively until abruptly declining into a negative sentiment.

Figure 6.1 hypothesize how TiA-Trust is moderated by emotional expressiveness in line with the uncanny valley theory (Affinity is moderated by human-likeness). Our analyzed graph depicts a proximal level of emotional expressiveness between the conditions "Less-expressive word and prosody" and "Expressive word," both trailing behind "Expressive prosody" and "Expressive word + prosody." These results cohere with our organized statistical findings 5.2. We infer that elevated emotional expressiveness and human-likeness, associated with expressive prosody (as validated by Zhu *et al.* [76]), exert an inverse effect on the affinity, which bears a substantial correlation with TiA-Trust.

To enhance the persuasiveness of these observations, further research could deploy an even broader range of emotional expressiveness conditions. Specifically, we could implement a study using a linear regression model to delve into whether varying levels of emotional expressiveness have a significant influence on anthropophilic or uncanny reactions. The model would take into account fixed aspects like the conditions of lexical expressiveness and prosodic expressiveness, as well as the ratings for human likeness and likability of CUI. Interaction between these elements should also be considered. The key objectives of the study would be to ascertain if the level of human-likeness rated by participants has a direct impact on the trust of the CUI and to evaluate the interaction between human-likeness ratings and prosody conditions.

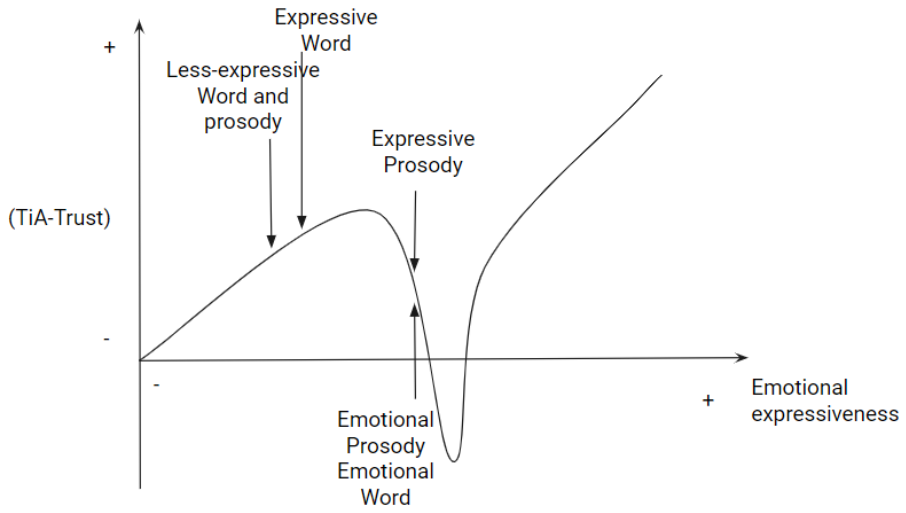

Figure 6.1: The graph (adapted from [46]) proposed a relation between the emotional expressiveness of the voice-based CUI, and the TiA-Trust of it.

6.2. Trust and Usability in CUIs

Our exploration focused on the effects of prosodic expressiveness and lexical emotional expressiveness on user trust and usability in conversational agents. In addition, we investigated the influence of voice-based and text-based Conversational User Interfaces (CUIs) on user trust and usability in AI-assisted decision-making.

For trust, as measured by the TiA-Trust, it was found to be significantly influenced by prosody, with a p-value of 0.037 from our two-way ANOVA test. However, no significant pairwise differences between prosodic and lexical expressiveness were found via the post-hoc analysis. Subsequent Analysis of Covariance (ANCOVA) suggested that neither prosodic expressiveness nor lexical expressiveness had a significant influence on user trust when factors such as TiA-PtT and ATI were controlled. In contrast, both TiA-PtT (F= 115.2, p< 0.001) and ATI (F= 11.03, p= 0.001) were determined to be significant covariates that exert influence on user trust. The relationship between trust and lexical expressiveness was also found to be contingent on prosody, as evidenced by our interaction plot.

Turning to usability, the two-way ANOVA test unveiled a significant effect of prosody (F= 2.547, p= 0.035). A post-hoc test further revealed that significant differences existed between the condition (high prosodic expressiveness, low lexical expressiveness) and condition (low prosodic expressiveness, high lexical expressiveness). Usability scores were significantly higher in the condition: low prosodic expressiveness, and high lexical expressiveness. However, subsequent ANCOVA findings indicated no significant influence on usability when factors such as ATI and TiA-PtT were controlled. Instead, ATI (F= 18.732, (p< 0.001) and TiA-PtT (F= 31.229, (p< 0.001) were found to be significant covariates exerting substantial influence on usability.

Regarding the comparison between voice-based CUI and text-based CUI, our study found no significant difference in user trust and usability between the two modalities, thereby challenging the assumptions that voice-based CUIs inherently foster greater user trust and perceived usability. In conclusion, after controlling for the influence of significant covariates, neither prosody nor word expressiveness significantly impacted user trust or usability. This underscores the importance of factors such as TiA-PtT and ATI in shaping perceptions of trust and usability in a conversational agent.

6.3. Efficiency vs. Efficacy and the Role of Trust

Our research findings offer a nuanced insight into the efficiency and efficacy of text-based versus voice-based CUIs, as depicted in Tables 5.5. When comparing the time required to complete tasks through different CUIs, we found participants' efficiency is slightly higher when using text-based CUIs. We speculate that this discrepancy may be rooted in the interactive nature of voice-based interfaces, where users are required to both speak and listen to responses, which potentially consumes more time than text-based CUIs.

Upon closer inspection of the voice-based CUI, we observed an intriguing pattern. The condition characterized by high prosodic expressiveness and high lexical expressiveness resulted in the highest submission accuracy, at 88%. It also led to the highest percentage of participants reviewing all recommendations, at 84%, despite having the longest task completion time, averaging 4.21 minutes. A plausible interpretation of these findings might be that users, perhaps distrusting the conversational interface, felt compelled to cross-verify the provided constraints and meticulously examine all recommendations before selecting the appropriate house. Such behavior would invariably increase both the time taken and the accuracy of the task.

Despite the higher peak accuracy achieved with voice-based CUIs, text-based interfaces had an overall higher submission accuracy, at 82.75%, compared to the 77.37% achieved by their voice-based counterparts. Furthermore, the efficiency of text-based interfaces stood out, with tasks being completed approximately half a minute faster than with voice-based CUIs.

Our research diverges from previous studies [12, 25, 55] on trust and work performance, which primarily focused on output quality and time spent on tasks. In contrast, we identify a trade-off among the trust inspired by the interface, the active task completion time, and user performance. This expanded perspective holds the potential to offer more detailed insights into the intricacies of user interactions within CUIs.

6.4. Implications for Designing Conversational DSS

The results provide some insights into the impact of emotional expressiveness on the user trust and usability of decision support systems. For future development, for the purpose of high usability and trust, designers should consider implementing a voice-based CUI with low expressive expressiveness and high lexical expressiveness. However, it has a lower submission accuracy for the task. This suggests that we should be cautious against the over-trust of CUI. The goal is to facilitate proper trust between the system and the user, avoiding over-trust or under-trust.

Upon deeper qualitative exploration, several observations were made. Firstly, voice recognition is the primary reason for negative sentiment in voice-based CUI, even though all participants were native English speakers, suggesting clear room for improvement. Secondly, a sense of frustration was noticed among users, attributed to the slower interaction speed and lesser flexibility in voice-based CUI, a scenario not present in text-based CUIs. This may be attributed to the voice-based CUI taking longer to process and respond to natural language. As a whole, designers should be aware of the shortcomings of voice-based CUI compared to text-based CUI.

In developing the conversational interface, we tried to maintain a sense of intelligence to build user trust. Despite encouraging exploration within the CUI, unforeseen user behavior could pose a challenge to designers in handling errors and crafting appropriate responses. In such cases, a lack of appropriate response could trigger user frustration. However, developing a CUI capable of addressing every possible scenario would pose significant implementation complexities.

Lastly, a theoretical proposition was made in explaining the reduction in trust as emotional expressiveness increased: the “uncanny valley” hypothesis. Mori *et al*.[46] explain the relation between the human likeness and affinity for an entity:“I predict that it is possible to create a safe level of affinity by deliberately pursuing a nonhuman design”. Thus the correlation between human likeness and affinity suggests a judicious design approach that maintains a degree of human resemblance without straying into the uncanny territory, thus ensuring an optimal level of affinity and, by extension, user trust.

6.5. Limitations

This current research possesses certain limitations that, if addressed, could pave the way for future explorations in the field. In our study, housing recommendation is adopted as a realistic use case for human-AI collaboration. This task may constrict the scope of discussion topics available to participants. Although participants tend to be familiar with the house-hunting situation. Some may find the topic less engaging, which could consequently result in lower ratings, independent of the various manipulations employed in the study. In accordance with the research conducted by Folstad in 2021 [18], it could be beneficial to transition the research domain from a predominantly formal context to one that is more socially driven. The idea here is to shift from an objective, concise approach to a more personal, informal one that allows for detailed, subjective discussions. For instance, Zhu *et al*. [76] apply the topic of music in voice assistant, which might be more engaging.

In our study, we utilized the platform, Prolific, for the purpose of data collection. The participant pool consisted of a total of 345 individuals, of whom 184 were able to successfully complete the given task. Conversely, we noted a total of 161 individuals who discontinued their participation prematurely. The rationale behind this significant dropout rate could be attributed to several factors such as the task's inherent complexity and the conversational user interface's (CUI) usability. It is worth noting that despite our efforts to facilitate task completion through means such as providing an instructional video and elaborating instructions, the CUI's usability still proved to be a challenge. Several influencing factors were identified, which include but are not limited to: 1. Unfamiliarity: A considerable number of participants were lacking in their understanding of the conversational agent and the requisite procedures necessary for task completion. 2. Compatibility issues: Some users reported technical difficulties due to specific web browsers not supporting voice input features. 3. Speed: Participants reported that the voice-based CUI was less efficient in responding when compared to its text-based counterpart. 4. Voice recognition: Certain factors, such as variances in accent and hardware quality, were found to pose challenges in voice recognition. Furthermore, we found that even among those who successfully finished the task, the workflow might have been perceived as irritating. Such negative experiences could potentially have influenced their overall evaluation of the chat interface. In light of these findings, future research initiatives are being planned. Lab investigations that will allow us to compare co-located human-human interactions with those involving a chatbot will be conducted. This will not only provide more nuanced insights into user behavior but also allow for more control over the experimental setup.

In addition to the focus on a single emotion, Future research efforts could broaden the scope to encompass other emotions such as sadness and anger, investigating their possible influence on interactions with voice-based CUIs. Moreover, these emotions could have a significant role in the formation of trust, a factor that remains yet to be fully explored.

Our study did not include any survey to evaluate participants' perceptions of emotional expressiveness in CUIs. This decision was made because existing research [76] found a significant influence of prosodic expressiveness on emotional expressiveness. There is a recognition of the need for further statistical validation due to the multiplicity of experimental designs in the field. Moving forward, we propose the inclusion of a measurement scale to directly assess emotional expressiveness. We're considering options such as a 100-point scale, or alternatively, Likert scales with either 5 or 7 points. The addition of such a metric would provide a more tangible and quantifiable measure of emotional expressiveness, contributing to the robustness of the research methodology and results.

6.6. Future Work

Based on this study, the following recommendations for further research are given:

- In this work, the undertaking of house selection yields a concrete outcome, thereby exhibiting a highly formalized nature. However, it may be advantageous to reorient the task towards a less structured and more socially interactive paradigm. An example of this could involve transforming the task to solicit advice or seek emotional support. This potential shift in focus is predicated on the presumption that individuals exhibit a greater propensity to engage in dialogue with a chatbot, rather than resorting to textual communication [18]. Indeed, people's inherent sociability may lead them to prefer interactions that mimic human conversation, thereby making such chatbot-assisted tasks more appealing and engaging.
- The investigation into the interaction effect between prosodic expressiveness and lexical expressiveness, specifically concerning the uncanniness effect, has been relatively limited. Our research, although not yielding significant evidence, aims to catalyze future studies.

7

Conclusion

In this paper, we use an empirical study to understand how a conversational agent's emotional expressiveness and modality influence user trust and usability in the domain of AI-assisted decision-making. Our research methodology included a 2x3 between-subjects experimental design involving 151 valid participants. We manipulated emotional expressiveness by modifying prosodic expressiveness (low vs high) and lexical expressiveness (low vs high), which were delivered through two modalities (voice vs text) in CUIs' acknowledgments.

Overall, the experimental results are insufficient to conclude the impact of prosodic expressiveness and lexical expressiveness on user trust and usability in CUIs. However, we find that the condition with low prosodic expressiveness and high lexical expressiveness has the highest perceived usability and trust among all conditions. Moreover, the interaction plot (cf. Figure 5.1) suggests the potential interaction effects of prosodic expressiveness and lexical expressiveness on trust. In addition, we also examined the ATI and TiA-PtT as covariates, both of which showed a strong positive correlation with Trust and usability.

The results highlight the importance of considering other influential factors such as the user's affinity to technology and propensity to trust. And we hypothesize that emotional expressiveness in a CUI at a point might detrimentally affect user trust and usability due to the potential uncanniness effect. These findings can potentially benefit the design and development of conversational agents used in AI-supported decision-making scenarios, by considering prosodic and lexical expressiveness.

Bibliography

[1] July 2020. URL: https://www.marketsandmarkets.com/Market-Reports/conversational-ai-market-49043506.html.

[2] Tanja Bänziger, Marcello Mortillaro, and Klaus R Scherer. “Introducing the Geneva Multimodal expression corpus for experimental research on emotion perception.” In: *Emotion* 12.5 (2012), p. 1161.

[3] Christoph Bartneck. “Affective expressions of machines”. In: *CHI’01 extended abstracts on Human factors in computing systems*. 2001, pp. 189–190.

[4] Diane S Berry and James W Pennebaker. “Nonverbal and verbal emotional expression and health”. In: *Psychotherapy and psychosomatics* 59.1 (1993), pp. 11–19.

[5] Luka Bradeško et al. “Curious Cat–Mobile, Context-Aware Conversational Crowdsourcing Knowledge Acquisition”. In: *ACM Transactions on Information Systems (TOIS)* 35.4 (2017), pp. 1–46.

[6] Ross W Buck et al. “Communication of affect through facial expressions in humans.” In: *Journal of personality and social psychology* 23.3 (1972), p. 362.

[7] Rafael A Calvo et al. *The Oxford handbook of affective computing*. Oxford Library of Psychology, 2015.

[8] Jiahao Chen et al. “Effects of anthropomorphic design cues of chatbots on users’ perception and visual behaviors”. In: *International Journal of Human–Computer Interaction* (2023), pp. 1–19.

[9] Michelle Cohn, Chun-Yen Chen, and Zhou Yu. “A large-scale user study of an Alexa prize chatbot: Effect of TTS dynamism on perceived quality of social dialog”. In: *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*. 2019, pp. 293–306.

[10] Louis Columbus. *What’s new in Gartner’s hype cycle for AI, 2020*. Oct. 2020. URL: https://www.forbes.com/sites/louiscolumbus/2020/10/04/whats-new-in-gartners-hype-cycle-for-ai-2020/?sh=746a15a8335c.

[11] Joe Crumpton and Cindy L Bethel. “A survey of using vocal prosody to convey emotion in robot speech”. In: *International Journal of Social Robotics* 8 (2016), pp. 271–285.

[12] Florian Daniel et al. “Quality control in crowdsourcing: A survey of quality attributes, assessment techniques, and assurance actions”. In: *ACM Computing Surveys (CSUR)* 51.1 (2018), pp. 1–40.

[13] Alexander Diel, Sarah Weigelt, and Karl F MacDorman. “A meta-analysis of the uncanny valley’s independent and dependent variables”. In: *ACM Transactions on Human-Robot Interaction (THRI)* 11.1 (2021), pp. 1–33.

[14] Mateusz Dubiel, Sylvain Daronnat, and Luis A Leiva. “Conversational Agents Trust Calibration: A User-Centred Perspective to Design”. In: *Proceedings of the 4th Conference on Conversational User Interfaces*. 2022, pp. 1–6.

[15] Justin Edwards and Elaheh Sanoubari. “A need for trust in conversational interface research”. In: *Proceedings of the 1st International Conference on Conversational User Interfaces*. 2019, pp. 1–3.

[16] Alexander Erlei et al. “Impact of algorithmic decision making on human behavior: Evidence from ultimatum bargaining”. In: *Proceedings of the AAAI conference on human computation and crowdsourcing*. Vol. 8. 2020, pp. 43–52.

[17] Franz Faul et al. “G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences”. In: *Behavior research methods* 39.2 (2007), pp. 175–191.

[18] Asbjørn Følstad and Cameron Taylor. “Investigating the user experience of customer service chatbot interaction: a framework for qualitative analysis of chatbot dialogues”. In: *Quality and User Experience* 6.1 (2021), p. 6.

[19] Thomas Franke, Christiane Attig, and Daniel Wessel. “A personal resource for technology interaction: development and validation of the affinity for technology interaction (ATI) scale”. In: *International Journal of Human–Computer Interaction* 35.6 (2019), pp. 456–467.

[20] Amos Freedy et al. “Measurement of trust in human-robot collaboration”. In: *2007 International symposium on collaborative technologies and systems*. Ieee. 2007, pp. 106–114.

[21] Robert W Frick. “Communicating emotion: The role of prosodic features.” In: *Psychological bulletin* 97.3 (1985), p. 412.

[22] Howard S Friedman et al. “Understanding and assessing nonverbal expressiveness: the affective communication test.” In: *Journal of personality and social psychology* 39.2 (1980), p. 333.

[23] Catherine Gao. “Use New Alexa Emotions and Speaking Styles to Create a More Natural and Intuitive Voice Experience”. In: *Amazon. com, Epub* 26 (2019).

[24] Akshit Gupta et al. “To Trust or Not To Trust: How a Conversational Interface Affects Trust in a Decision Support System”. In: *Proceedings of the ACM Web Conference 2022*. 2022, pp. 3531–3540.

[25] Lei Han et al. “The impact of task abandonment in crowdsourcing”. In: *IEEE Transactions on Knowledge and Data Engineering* 33.5 (2019), pp. 2266–2279.

[26] Jeffrey T Hancock, Christopher Landrigan, and Courtney Silver. “Expressing emotion in text-based communication”. In: *Proceedings of the SIGCHI conference on Human factors in computing systems*. 2007, pp. 929–932.

[27] Shlomo Hareli and Ursula Hess. “The social signal value of emotions”. In: *Cognition & Emotion* 26.3 (2012), pp. 385–389.

[28] Marc Hassenzahl, Sarah Diefenbach, and Anja Göritz. “Needs, affect, and interactive products–Facets of user experience”. In: *Interacting with computers* 22.5 (2010), pp. 353–362.

[29] Danula Hettiachchi et al. “” Hi! I am the Crowd Tasker” Crowdsourcing through Digital Voice Assistants”. In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 2020, pp. 1–14.

[30] Kevin Anthony Hoff and Masooda Bashir. “Trust in automation: Integrating empirical evidence on factors that influence trust”. In: *Human factors* 57.3 (2015), pp. 407–434.

[31] Daniel Holliday, Stephanie Wilson, and Simone Stumpf. “User trust in intelligent systems: A journey over time”. In: *Proceedings of the 21st international conference on intelligent user interfaces*. 2016, pp. 164–168.

[32] Simo Hosio et al. “Leveraging wisdom of the crowd for decision support”. In: *Proceedings of the 30th International BCS Human Computer Interaction Conference 30*. 2016, pp. 1–12.

[33] Krystal Hu. *ChatGPT sets record for fastest-growing user base - analyst note — reuters.com*. `https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/`. [Accessed 24-Jun-2023].

[34] Ji-Youn Jung et al. “Great chain of agents: The role of metaphorical representation of agents in conversational crowdsourcing”. In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 2022, pp. 1–22.

[35] Peter GW Keen. “Decision support systems: a research perspective”. In: *Decision support systems: Issues and challenges: Proceedings of an international task force meeting*. 1980, pp. 23–44.

[36] Mark L Knapp, Judith A Hall, and Terrence G Horgan. *Nonverbal communication in human interaction*. Cengage Learning, 2013.

[37] Moritz Körber. “Theoretical considerations and development of a questionnaire to measure trust in automation”. In: *Proceedings of the 20th Congress of the International Ergonomics Association (IEA 2018) Volume VI: Transport Ergonomics and Human Factors (TEHF), Aerospace Human Factors and Ergonomics 20*. Springer. 2019, pp. 13–30.

[38] Pavel Kucherbaev et al. “Crowdcafe-mobile crowdsourcing platform”. In: *arXiv preprint arXiv:1607.01752* (2016).

[39] John D Lee and Katrina A See. “Trust in automation: Designing for appropriate reliance”. In: *Human factors* 46.1 (2004), pp. 50–80.

[40] Poornima Madhavan and Douglas A Wiegmann. “Similarities and differences between human–human and human–automation trust: an integrative review”. In: *Theoretical Issues in Ergonomics Science* 8.4 (2007), pp. 277–301.

[41] Meeri Mäkäräinen, Jari Kätsyri, and Tapio Takala. “Exaggerating facial expressions: A way to intensify emotion or a way to the uncanny valley?” In: *Cognitive Computation* 6 (2014), pp. 708–721.

[42] Victoria Masterson. *What has caused the global housing crisis - and how can we fix it?* June 2022. URL: https://www.weforum.org/agenda/2022/06/how-to-fix-global-housing-crisis/.

[43] Panagiotis Mavridis et al. “Chatterbox: Conversational interfaces for microtask crowdsourcing”. In: *Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization*. 2019, pp. 243–251.

[44] Saif Mohammad. “Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words”. In: *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: Long papers)*. 2018, pp. 174–184.

[45] Roger K Moore. “A Bayesian explanation of the ‘Uncanny Valley’effect and related psychological phenomena”. In: *Scientific reports* 2.1 (2012), pp. 1–5.

[46] Masahiro Mori, Karl F MacDorman, and Norri Kageki. “The uncanny valley [from the field]”. In: *IEEE Robotics & automation magazine* 19.2 (2012), pp. 98–100.

[47] Clifford Nass, Jonathan Steuer, and Ellen R Tauber. “Computers are social actors”. In: *Proceedings of the SIGCHI conference on Human factors in computing systems*. 1994, pp. 72–78.

[48] Clifford Nass et al. “The effects of emotion of voice in synthesized and recorded speech”. In: *Proceedings of the AAAI symposium emotional and intelligent II: The tangled knot of social cognition*. AAAI North Falmouth, MA. 2001.

[49] Lynne C Nygaard and Jennifer S Queen. “Communicating emotion: linking affective prosody and word meaning.” In: *Journal of Experimental Psychology: Human Perception and Performance* 34.4 (2008), p. 1017.

[50] Raja Parasuraman, Thomas B Sheridan, and Christopher D Wickens. “Situation awareness, mental workload, and trust in automation: Viable, empirically supported cognitive engineering constructs”. In: *Journal of cognitive engineering and decision making* 2.2 (2008), pp. 140–160.

[51] PricewaterhouseCoopers. *Consumer intelligence series: Prepare for the Voice Revolution*. Feb. 2018. URL: https://www.pwc.com/us/en/services/consulting/library/consumer-intelligence-series/voice-assistants.html.

[52] Pearl Pu, Li Chen, and Rong Hu. “A user-centric evaluation framework for recommender systems”. In: *Proceedings of the fifth ACM conference on Recommender systems*. 2011, pp. 157–164.

[53] Si Qiao and Roger Eglin. “Accurate behaviour and believability of computer generated images of human head”. In: *Proceedings of the 10th International Conference on Virtual Reality Continuum and Its Applications in Industry*. 2011, pp. 545–548.

[54] Si Qiao, Roger Eglin, and Ariel Beck. “Audience perception of computer generated human facial behaviour”. In: *GSTF International Journal on Computing* 1.3 (2011), pp. 61–65.

[55] Sihang Qiu, Ujwal Gadiraju, and Alessandro Bozzon. “Improving worker engagement through conversational microtask crowdsourcing”. In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 2020, pp. 1–12.

[56] Sihang Qiu, Ujwal Gadiraju, and Alessandro Bozzon. “Ticktalkturk: Conversational crowdsourcing made easy”. In: *Conference Companion Publication of the 2020 on Computer Supported Cooperative Work and Social Computing*. 2020, pp. 53–57.

[57] Sihang Qiu, Ujwal Gadiraju, and Alessandro Bozzon. “Towards memorable information retrieval”. In: *Proceedings of the 2020 ACM SIGIR on international conference on theory of information retrieval*. 2020, pp. 69–76.

[58] Minjin Rheu et al. “Systematic review: Trust-building factors and implications for conversational agent design”. In: *International Journal of Human–Computer Interaction* 37.1 (2021), pp. 81–96.

[59] Peter AM Ruijten, Jacques MB Terken, and Sanjeev N Chandramouli. “Enhancing trust in autonomous vehicles through intelligent user interfaces that mimic human behavior”. In: *Multimodal Technologies and Interaction* 2.4 (2018), p. 62.

[60] Christine Rzepka, Benedikt Berger, and Thomas Hess. “Voice assistant vs. Chatbot–examining the fit between conversational agents’ interaction modalities and information search tasks”. In: *Information Systems Frontiers* 24.3 (2022), pp. 839–856.

[61] Maha Salem et al. “Would you trust a (faulty) robot? Effects of error, task type and personality on human-robot cooperation and trust”. In: *Proceedings of the tenth annual ACM/IEEE international conference on human-robot interaction*. 2015, pp. 141–148.

[62] Harold Schlosberg. “Three dimensions of emotion.” In: *Psychological review* 61.2 (1954), p. 81.

[63] Marc Schroder. “Expressing degree of activation in synthetic speech”. In: *IEEE Transactions on Audio, Speech, and Language Processing* 14.4 (2006), pp. 1128–1136.

[64] Elaine Short et al. “No fair!! an interaction with a cheating robot”. In: *2010 5th acm/ieee international conference on human-robot interaction (hri)*. IEEE. 2010, pp. 219–226.

[65] Keng Siau and Weiyu Wang. “Building trust in artificial intelligence, machine learning, and robotics”. In: *Cutter business technology journal* 31.2 (2018), pp. 47–53.

[66] Yao Song, Yanpu Yang, and Peiyao Cheng. “The investigation of adoption of voice-user interface (VUI) in smart home systems among chinese older adults”. In: *Sensors* 22.4 (2022), p. 1614.

[67] Isabel Thielmann and Benjamin E Hilbig. “Trust: An integrative review from a person–situation perspective”. In: *Review of General Psychology* 19.3 (2015), pp. 249–277.

[68] Angela Tinwell, Mark Grimshaw, and Debbie Abdel-Nabi. “Effect of emotion and articulation of speech on the uncanny valley in virtual characters”. In: *Affective Computing and Intelligent Interaction: Fourth International Conference, ACII 2011, Memphis, TN, USA, October 9–12, 2011, Proceedings, Part II*. Springer. 2011, pp. 557–566.

[69] Angela Tinwell, Mark Grimshaw, and Debbie Abdel-Nabi. “The uncanny valley and nonverbal communication in virtual characters”. In: *Nonverbal Communication in Virtual Worlds* (2014), pp. 325–341.

[70] Suzanne Tolmeijer et al. “Second chance for a first impression? Trust development in intelligent system interaction”. In: *Proceedings of the 29th ACM Conference on user modeling, adaptation and personalization*. 2021, pp. 77–87.

[71] Edward C Tomlinson and Roger C Mryer. “The role of causal attribution dimensions in trust repair”. In: *Academy of management review* 34.1 (2009), pp. 85–104.

[72] Gerrit H Van Bruggen, Ale Smidts, and Berend Wierenga. “Improving decision making by means of a marketing decision support system”. In: *Management Science* 44.5 (1998), pp. 645–658.

[73] Frank MF Verberne, Jaap Ham, and Cees JH Midden. “Trust in smart systems: Sharing driving goals and giving information to increase trustworthiness and acceptability of smart systems in cars”. In: *Human factors* 54.5 (2012), pp. 799–810.

[74] Katharina Weitz et al. “” Do you trust me?” Increasing user-trust by integrating virtual agents in explainable AI interaction design”. In: *Proceedings of the 19th ACM international conference on intelligent virtual agents*. 2019, pp. 7–9.

[75] Michael Juntao Yuan et al. “Evaluation of user interface and workflow design of a bedside nursing clinical decision support system”. In: *Interactive journal of medical research* 2.1 (2013), e2402.

[76] Qingxiaoyang Zhu et al. “Effects of Emotional Expressiveness on Voice Chatbot Interactions”. In: *4th Conference on Conversational User Interfaces*. 2022, pp. 1–11.

A Acknowledgement Templates

We provide some examples (cf. Figure 7.2) to illustrate how expressive acknowledgment templates vary from less expressive acknowledgment templates in terms of lexical choice (cf. Figure 7.1). In the chatbot implementation, we use this acknowledgment template to respond to users' preference inputs(cf. Figure 7.3).

	NRC	Valence	Arousal	Dominance
emotionally less-expressive words	see	0.625	0.269	0.312
	yea	0.751	0.412	0.472
	sure	0.724	0.27	0.793
	ahhh	0.562	0.517	0.241
	understood	0.780	0.441	0.686
	noted	0.615	0.382	0.464
	confirmed	0.760	0.490	0.764
emotionally more-expressive words (valence score $>$ 0.9)	perfect	0.98	0.471	0.87
	wonderful	0.971	0.776	0.83
	fantastic	0.969	0.696	0.831
	great	0.958	0.665	0.81
	congrats	0.93	0.726	0.775
	nice	0.93	0.442	0.65
	interesting	0.927	0.726	0.731
	reasonable	0.902	0.337	0.726

Table 7.1: Example of emotionally 'less-expressive' and 'more-expressive' words and their scores for Valence, Arousal and Dominance

Condition **A** (less-expressive words)	Condition **B** (more-expressive words)
Bot: What is your preference? Human: I want to stay for more than 6 months Bot: **Understood.(low prosody)**	Bot: What is your preference? Human: I want to stay for more than 6 months Bot: **That is wonderful.(low prosody)**
Condition **C**(more-expressive Prosody)	Condition **D** (more-expressive Words and Prosody)
Bot: What is your preference? Human: I want to stay for more than 6 months Bot: **Understood.(high prosody)**	Bot: What is your preference? Human: I want to stay for more than 6 months Bot: **That is wonderful.(high prosody)**

Table 7.2: A dialogue clip with different levels of acknowledgment for prosodic and lexical expressiveness.

Preference Category	Emotionally less-expressive words	Emotionally more-expressive words
Budget	• Understood, we can work within your budget of xxx. • Noted. we can find some options for you at the budget of xxx. • That is possible. With a budget of xxx, there are options for a place to call your own.	• Perfect, we can work within your budget of xxx. • Nice, we can find some options for you at the budget of xxx. • Awesome! With a budget of xxx, there are options for a place to call your own.
House type	• Hmm. You can find some xxx here. • Oh, you're looking for a xxx. I'll keep that in mind. • Alright, you have a preference for xxx.	• Fantastic! You can find some xxx here. • Terrific, you're looking for a xxx. I'll keep that in mind. • Wonderful, you have a preference for xxx.
Commute	• Acknowledged. We'll work with the commute of xxx minutes to find your dream home. • I see. Your ideal scenario is a commute time under xxx minutes. • Yes. I've registered your desire for a commute time of less than xxx minutes.	• Fabulous. We'll work with the commute of xxx minutes to find your dream home. • Marvelous. Your ideal scenario is a commute time under 10 minutes. • Brilliant. I've registered your desire for a commute time of less than xxx minutes.
Duration	• Ok! We can work within this duration of xxx months. • Right. Your preference is for a duration of xxx months. I'll consider that as we proceed. • Okay. I've taken into account your preference for a xxx month's duration.	• Impressive! We can work within this duration of xxx months. • Remarkable. Your preference is for a duration of xxx months. I'll consider that as we proceed. • Extraordinary. I've taken into account your preference for a xxx month's duration.
Near supermarket	• Confirmed. Let's explore homes that near the supermarket • Got it. you're seeking a house that offers easy access to a nearby supermarket. • I see, I've taken note of your preference for a house near a supermarket.	• Excellent. Let's explore homes that near the supermarket. • Awesome. you're seeking a house that offers easy access to a nearby supermarket. • Outstanding, I've taken note of your preference for a house near a supermarket.
Registration	• Noted. You want to register an address. I'll consider that. • Understood, registering the address with the house is important to you. • Alright, you need to register the address with the house.	• Superb. You want to register an address. I'll consider that. • great. Registering the address with the house is important to you. • admirable, you need to register the address with the house.

Table 7.3: Templates of using emotionally 'less-expressive' and 'more-expressive' words, to demonstrate acknowledgment.

B Activity Diagram of House Selection

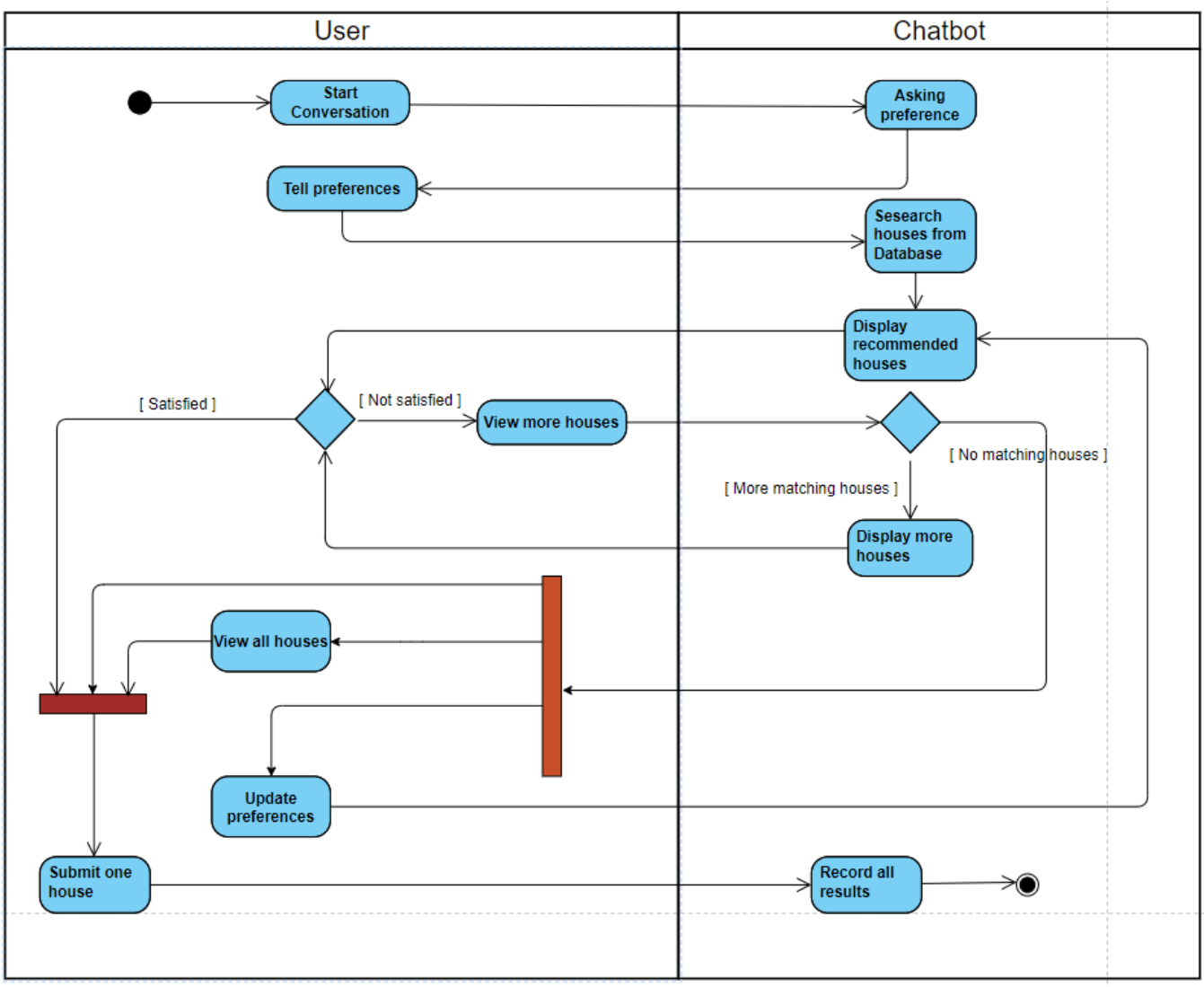

Figure 7.1: Activity Diagram illustrates the interaction of user and chatbot

C Consent Form

You are invited to participate in a research study on conversational agents. We are interested in understanding how people reason with conversational agents. Before deciding whether to participate, it is important for you to understand why the research is being conducted and what it will involve. Please read the following information carefully.

Purpose: The aim of this study is to analyze how individuals engage with conversational agents. More specifically, you will be prompted to express your preferences for a house in response to a given scenario. For each preference you articulate, we will supply a corresponding utterance to help you express it.

Confidentiality: All data collected during the experiment will be treated as confidential. Your responses will be recorded anonymously and we will not collect any personally identifying information.

Consent: By clicking the ”Confirm and Continue” button on this page, you confirm that you have read this consent form, understand the procedures involved in the experiment, and freely consent to participate in the study. You also understand that your answers will be recorded for research purposes and that your data will be kept confidential.