

©Copyright 2020

Mohit Jain

# Making Chatbots More Transparent and Applicable to New Demographics

Mohit Jain

A dissertation  
submitted in partial fulfillment of the  
requirements for the degree of

Doctor of Philosophy

University of Washington

2020

Reading Committee:

Shwetak N. Patel, Chair

Khai N. Truong

James Fogarty

Program Authorized to Offer Degree:  
Computer Science & Engineering

University of Washington

## **Abstract**

Making Chatbots More Transparent and Applicable to New Demographics

Mohit Jain

Chair of the Supervisory Committee:

Dr. Shwetak N. Patel

Computer Science & Engineering and Electrical Engineering

Conversational agents, popularly called chatbots, received significant attention in the last few years. The major reason behind the success of these systems is that chatbots use the already familiar conversational interface. However, chatbots are still in their nascent stage: They have a low penetration rate as 84% of the Internet users have not used a chatbot yet. First, we conducted a study with 16 first-time chatbot users interacting with eight chatbots over multiple sessions on the Facebook Messenger platform. Analysis of chat logs and user interviews revealed several major problems with the current chatbots, including (a) mismatch between the chatbot’s state of understanding (also called *context*) and the user’s perception of the chatbot’s understanding, (b) limitations in natural language understanding technologies leading to dialog failures, and (c) targeting chatbots specifically towards the Internet-savvy technically-advanced users.

Second, we focused on these three problems and developed solutions, respectively: (a) Convey: stands for CONtext View, is a window added to the chatbot interface, displaying the conversational context and providing interactions with the context values, which we evaluated with 16 participants; (b) Resilient Chatbot: explores user preferences for eight repair strategies taken from commercially-deployed chatbots (e.g., confirmation, providing options) as well as novel strategies explaining characteristics of the underlying machine learning algorithms, was evaluated with 216 MTurkers; and (c) FarmChat: is a multi-modal multi-lingual

conversational agent, to meet the information needs of rural low literate farmers, and evaluated with 34 farmers in Ranchi, India. To summarize, we propose ways to make chatbots more transparent, and extend its applicability to new demographics.

## TABLE OF CONTENTS

	Page
List of Figures . . . . .	iii
List of Tables . . . . .	v
Glossary . . . . .	vi
Chapter 1: Introduction . . . . .	1
1.1 Research Aim 1: To bridge gap between the chatbot’s understanding and the user’s perception of its understanding . . . . .	3
1.2 Research Aim 2: To design repair strategies to help users recover from dialog failures . . . . .	4
1.3 Research Aim 3: To broaden chatbot’s applicability to new demographics, not limiting chatbots to technically-advanced users . . . . .	5
1.4 Thesis Statement . . . . .	7
Chapter 2: Related Work . . . . .	8
2.1 Basics of a Conversation System . . . . .	8
2.2 Evaluating Conversational Agents . . . . .	9
2.3 Stale User Interface of Chatbots . . . . .	10
2.4 Breakdowns, Repairs in Communication and Human-Agent Interaction . . . . .	11
2.5 Explainable Artificial Intelligence . . . . .	13
2.6 UIs for Low Literate Users, especially Farmers . . . . .	14
Chapter 3: Evaluating and Informing the Design of Chatbots . . . . .	17
3.1 Introduction . . . . .	17
3.2 Study Design . . . . .	17
3.3 Quantitative Data Analysis . . . . .	22
3.4 Findings: Qualitative Data Analysis . . . . .	26

3.5	Discussion . . . . .	35
Chapter 4:	Convey: Exploring the Use of a Context View for Chatbots . . . . .	43
4.1	Introduction . . . . .	43
4.2	Design of Convey . . . . .	43
4.3	Study Design . . . . .	46
4.4	Results . . . . .	48
4.5	Discussion and Design Implications . . . . .	52
Chapter 5:	Resilient Chatbots: Repair Strategy Preferences for Conversational Break- downs . . . . .	55
5.1	Repair Strategies and Research Questions . . . . .	55
5.2	Methodology . . . . .	58
5.3	Results . . . . .	62
5.4	Discussion . . . . .	70
Chapter 6:	FarmChat: A Conversational Agent to Answer Farmer Queries . . . . .	75
6.1	Why Chatbots for Farmers? . . . . .	76
6.2	Formative Findings . . . . .	77
6.3	FarmChat System Design . . . . .	82
6.4	FarmChat Comparative Evaluation . . . . .	89
6.5	Results . . . . .	92
6.6	Discussion . . . . .	104
Chapter 7:	Conclusion and Future Directions . . . . .	109
7.1	Future Directions . . . . .	111
7.2	Conclusion . . . . .	112
	Bibliography . . . . .	115

## LIST OF FIGURES

Figure Number	Page
3.1 (a) <i>Top</i> : Total chat messages exchanged; (b) <i>Bottom</i> : Division of those messages among chatbots and participants. . . . .	23
3.2 (a) <i>Top</i> : Content of the chatbot messages; (b) <i>Bottom</i> : Content of the participant (human) messages. . . . .	23
3.3 Interaction Time spent by the participant per chatbot. . . . .	23
3.4 <i>Alterra</i> chatbot showing different UI elements . . . . .	27
3.5 <i>CNN</i> chatbot showing different UI elements . . . . .	28
3.6 Likert-scale rating by the participants (with standard deviation shown by error bars) . . . . .	36
4.1 Shoe shopping default chatbot. . . . .	44
4.2 Shoe shopping chatbot with <i>Convey</i> at the top. Screenshot of the interface used in the user study. . . . .	45
4.3 Likert-scale rating by the participants (with standard deviation shown as error bars) . . . . .	50
4.4 Other <i>Convey</i> use cases - Left: News; Right: Support . . . . .	54
5.1 Eight repairs for the successful banking condition. Top left image initial prompt a user is given. From top to bottom: Top, Repeat, Confirmation, Options, Defer, Keyword Highlight Explanation, Keyword Confirmation Explanation, Out-of-Vocabulary Explanation. . . . .	60
5.2 Bradley-Terry rankings of repair strategies. From left to right, rankings for: all data, successful conditions, unsuccessful conditions. From top to bottom: lowest ranked to highest ranked. . . . .	65
6.1 System Design of FarmChat . . . . .	83
6.2 User Interface of Audio-only FarmChat . . . . .	86
6.3 User Interface of Audio+Text FarmChat . . . . .	86
6.4 Likert-scale rating considering ratings for the interface preferred by the participant . . . . .	94

6.5	FarmChat speech-based interaction . . . . .	96
7.1	Combining learning from the messenger study, Convey, Resilient Chatbot and FarmChat to propose a final chatbot design. . . . .	113

## LIST OF TABLES

Table Number	Page
3.1 Description of the selected eight chatbots for the user study. . . . .	19
4.1 Results from Log Data, mean±std (bold with p<0.05) . . . . .	49
5.1 Significant values, after Bonferroni adjustment (p <0.05/8), are noted with **. Marginally significant values (p <0.1/8) are noted with *. . . . .	63
5.2 Strength and weaknesses of repairs as reported by participants. . . . .	64
6.1 Current information sources and their challenges summarized in the square brackets. . . . .	81
6.2 Demographic data for the user study (refer Section 6.4.1). Note: The same 34 participants are categorized with respect to their literacy level, smartphone usage and profession. (A-o: Audio-only; A+T: Audio+Text; I: Illiterate; SL: Semi-literate; L: Literate; m=mean; std: standard deviation) . . . . .	93
6.3 Results from log data, mean±std. No significant difference was obtained be- tween the two FarmChat interfaces. (*Cannot be calculated accurately, as a majority of literate participants read the response multiple times instead of tapping the chat bubble. The reported number is for the sake of completeness.)	97

## **GLOSSARY**

AI: Artificial Intelligence

CA: Conversational Agent

HCI: Human-Computer Interaction

ICT4D: Information and Communication Technologies for Development

IVR: Interactive Voice Response

KCC: Kisan Call Center

MTURK: Mechanical Turk

NLP: Natural Language Processing

QA: Question and Answer

UI: User Interface

VUI: Voice User Interface

XAI: Explainable Artificial Intelligence

## ACKNOWLEDGMENTS

I would like to thank all the people who have helped me in achieving this significant milestone. I can not have gone this far without your support.

First, I would like to thank my family. Your love and support motivate me and make nothing seem impossible. Second, I would like to thank people who are instrumental to my PhD - Shwetak Patel, Khai Truong, Pratyush Kumar, and Zainul Charbiwala. Shwetak, thank you for being an amazing advisor. In spite of the timezone difference and your busy schedule, you made this off-campus working-from-India PhD work. I highly admire your enthusiasm, creativity, and encouraging and supportive attitude, and have learned a lot from you. Khai, thank you for your instrumental mentorship throughout my research career, starting from my early days at the University of Toronto. You will always be the first person whom I can reach out to to discuss any research idea. Pratyush, thank you for your constant support and guidance. I can still clearly remember our days at IBM Research when we started working on DDUIAI (Data-Driven UI for AI), which later got renamed to BeeHive, and then got matured into Convey, putting the seeds for my PhD work. Zainul, thank you for having constant faith in me that I can do a PhD remotely, and a big thank you for steering through the IBM processes to obtain all the approvals required to make it work.

I am extremely lucky to have spent one year of my life with an incredible group of people at the Ubicomp Lab. Mayank, I have always been amazed by your research breadth and your out-of-the-box thinking. Thank you for providing your insightful feedback during numerous discussions about blood pressure estimation and glove based DigiTouch. Tien, thank you for guiding me during the blood pressure estimation project. It was my first PhD research project. Still remember the last day of the deadline working at your house, crunching the

data. Fun times! Alex, a big thank you for reviewing all my papers, and fixing numerous grammatical mistakes. You have all the traits of becoming an awesome Professor! Elliot, thank you for teaching me the basics of signal and audio processing. Even now, you are my go-to person for any EE-related doubts. Hanchuan, I will always remember your simple yet tricky questions about India, and thank you for introducing me to Din Tai Fung. Edward, I highly admire your relentlessness, especially in pushing the blood pressure estimation work for multiple years. You will be an amazing Professor! Josh, thank you for teaching me the basics of soldering and everything related to working on the bench. Lilian, I remember our long conversations about life, mental health and random stuff. Thank you for sharing. Also thank you for teaching me 3D printing. Eric, you are an amazing engineer with a full grasp on hardware and software, and I highly admire your system-building skills. I still remember our times working on the glove, with you even teaching me how to stitch. Ruth, you were fun and inspiring to talk with. Morelle, CJ, Farshid, Manuja and other juniors, it was fun meeting you during my random visit to Seattle and at CHI/Ubicomp conferences. I am sure you all are very talented, and will achieve a lot in the following years.

Special thank you to Elise and Lindsay for being wonderful graduate program advisors, who were always there for me and helped me through out the journey!

I would like to thank my industry collaborators at IBM Research who had made my PhD journey truly special including: Ramachandra Kota, Jagabondhu Hazra, Vera Liao, Justin D. Weisz and Zahra Ashktorab.

Also, I would like to thank James, Kurtis and Leah, to agree to be on my committee and for providing excellent constructive feedback.

Finally, I would like to thank my dear wife Ishita. I have been so fortunate to have you by my side. You have been very supportive, not only in taking good care of the household, but also helping me in several of the research works which are part of this thesis.

## DEDICATION

to my loving, supportive and crazy typical Indian family,  
especially my dear wife, *Ishita*

## Chapter 1

### INTRODUCTION

Licklider’s ‘Man-machine symbiosis’ [125] was one of the earliest discourses from an HCI perspective that visualized humans interacting with machines in a natural manner. The first conversational agent (CA) emerged in 1966 from MIT, called ELIZA [211], which emulated a Rogerian psychotherapist. Such turn-based conversational agents are popularly called *chatbots*. ELIZA worked on simple declarative rules: if a certain keyword was identified in the user text, it responded with one or more pre-defined outputs. Subsequently, in the latter chatbots, the rules used for both natural language understanding and natural language generation were enriched. Ontologies were used to represent word meanings, reasoning was used to identify user intent, and memory was used to continue a contextual dialog [38, 170, 213, 224]. The notable follow-up chatbots (in 2000s) included MegaHAL [96], ALICE [1], and Elizabeth [184]. The main purpose of all these chatbots was to maintain a conversation with a human user. Indeed, naturalness was the most important metric for evaluating chatbots. In 1990, the Loebner Prize was instituted as an annual competition to award the most human-like chatbot. Recent examples from the Loebner winners are Mitsuku [2] and Rose [3]. Popular chatbots that have recently emerged from the industry are Xiaoice, Tay and Zo from Microsoft.

More recently, the term ‘Conversational Agent’ has come to mean a wide variety of systems with varying capabilities and purposes, with the underlying assumption that human interactions with the systems resemble normal conversations. In the last decade, the tech industry has mainly been devoting its efforts towards *utility-driven chatbots* – those designed with the goal of accomplishing specific task(s), *e.g.*, Dominos chatbot for ordering pizzas. The main idea is to enable users to effortlessly ‘converse’ with intelligent agents of businesses, just

as they converse with their friends and family. Chatbots have appeared on a variety of mobile and ubiquitous platforms, including phones, smart speakers, VR/AR devices, smartwatches and even operating systems (Siri by Apple and Cortana by Microsoft). Technology companies have raced to deploy platforms for developing such chatbots with built-in natural-language capabilities (such as Facebook Messenger, Amazon Lex, Google Dialog Flow and IBM Watson Conversation). Consequently, a large number of chatbots have been developed recently – *e.g.*, over 100,000 have been created just on the Facebook’s Messenger platform alone within a year of its release in 2015 [81] – for varied use-cases ranging from pizza ordering (Domino’s) to shopping (Burberry), from connecting like-minded humans (Chatible) to cab booking (Uber), and from chit-chatting (Pandorabots) to reading news (CNN). Developers are moving from app-first design – where each app comes with its own interface, thus incurring a small learning curve – to a chatbot-first model, which uses the already familiar conversational interface [69]. The primary focus of these chatbots is not to mimic human conversation but to enable tasks through the ease of conversation. Anthropomorphism in these chatbots, when it exists, seeks to augment the efficiency of the task-solving process. These conversational agents are typically available in two modalities – speech (such as Siri, Alexa, Cortana) and text-messaging (such as Domino’s, CNN, Pandorabots, Burberry, *etc.* found on Messenger, Slack, and/or Skype platform). Proponents of chatbots embrace their many strengths: *e.g.*, user’s familiarity with the conversational interaction, seamless natural-language interface across use-cases, offering direct information access and simplified navigational paths, and the promise of personalized and evolving intelligence [40, 127].

In spite of this rapid growth, the adoption of chatbots is still in its nascent stage, as a majority of users are first-time chatbot users; 84% of the Internet users have not used a chatbot yet [90]. Regular chatbot users were found to be disappointed and even frustrated with the current bots [127], and most chose to limit the usage of chatbots to simple tasks (*e.g.*, setting alarms), as the users were not certain of the system’s intelligence level. Even high-profile chatbots, like Microsoft Tay, were found to be inadequate at handling all of the complexities of natural language interactions, resulting it to be listed as one of the biggest

technology failures of 2016 by MIT Technology Review [162]. Other works evaluating users' experience with chatbots have discovered a gulf between experience and expectation with respect to the capabilities of chatbots [124,127,199]. Thus, a lot of research work is needed to make chatbots usable and acceptable.

This thesis mainly focuses on increasing the acceptability of chatbots by making them more transparent and broadening the chatbots' use-case to new demographics.

To start with, it is crucial to understand the interaction pattern of chatbot users to inform and guide the design of future bots. In particular, our work (details in chapter 3) focuses on first-time chatbot users, as studying first-time users can be more insightful compared to experienced users who might have grown accustomed to the limitations of chatbots and learned to adapt around them. While the HCI community has studied how conversational agents are used in different settings [105,123,124,127,199], none of them focuses on first-time chatbot users. Towards this goal, in chapter 3, we study the experience of sixteen first-time chatbot users interacting with a curated list of eight chatbots on the Facebook Messenger platform. We chose Messenger as it hosts the maximum number of chatbots [49] and is the second-most popular text-messaging app [194]. For our analysis, we combined qualitative findings from the semi-structured interviews with the quantitative findings from  $\sim 10,000$  messages that the participants exchanged with the chatbots. Apart from findings to evolve the design of future chatbots, we identified several issues with the state-of-the-art chatbots, which we discuss in detail below, along with proposing a solution for each of them:

### ***1.1 Research Aim 1: To bridge gap between the chatbot's understanding and the user's perception of its understanding***

Recent works by several researchers [124,127,199], including our initial study with first-time chatbot users, discovered a gulf between experience and expectation with respect to both intelligence and the user interface of chatbots. Although users enjoy chatbots that can continue a conversation specifically by retaining *conversational context*, there is a mismatch between the chatbot's real context versus the user's perception of the chatbot context, *i.e.*,

there is a difference between their mental models [123]. This is even more problematic in the case of lengthy, complex conversations. Similarly, users were found to be apprehensive in using conversational systems (*e.g.*, Siri, Cortana) for complex tasks, as the users were not certain of the system’s intelligence level and had a poor mental model of its contextual state [123, 127, 169]. Moreover, certain chatbot assumptions are not evident to the user, further exacerbating this issue.

Against this background, in chapter 4, we propose *Convey* (CONtext View), a window added to the chatbot interface that displays the (inferred and assumed) context of the conversation to the user. It also provides intuitive interactions on the context values, enabling users to modify them in a simple and efficient manner. The *Convey* content gets updated as the conversation proceeds, thus always showing the latest understanding of the chatbot. To evaluate the effectiveness and usability of the proposed design, we conducted a 16 participant user study centered around a chatbot for buying shoes. The results show that participants preferred using chatbot with *Convey* and found it to be easier to use, less mentally demanding, intuitive, and faster compared to the default chatbot without *Convey*.

## **1.2 Research Aim 2: To design repair strategies to help users recover from dialog failures**

While interacting with the eight chatbots on the Facebook Messenger platform, several participants experienced dialog failures, mainly due to limited natural language understanding technologies. Breakdowns in understanding user input happen often, and they can have profound impact on how users perceive and interact with a chatbot. In the worst case, users may abandon the chatbot or the current task. Or, they may need to endure a haphazard trial-and-error process to recover from the breakdown. Both breakdowns and current recovery processes decrease users’ satisfaction, trust, and willingness to continue using a chatbot [127]. A universal challenge faced by chatbot developers is how to design appropriate strategies that mitigate the negative impact of breakdowns. Previous work [120, 199, 215] studied strategies that aim to alleviate users’ negative emotional response from agent or robot breakdowns,

such as showing politeness and apologetic behaviors. However, in utility-driven chatbots, these strategies may not be very effective when the user still fails to accomplish the task. In chapter 5, we focus on strategies that support *repair* – recovering from the breakdown and accomplishing task goal.

Repair is a ubiquitous phenomenon in human communication. When a breakdown happens in a conversation, people take a variety of actions such as repeating, rephrasing, or clarifying, to repair it. Although chatbot users should be skillful in using similar actions as the *speaker*, the repair task becomes challenging as the *listener* is no longer a fellow human. Chatbots rely on machine learning algorithms to process a user’s input, which are “black boxes” for the user. Though these interfaces are deemed “conversational,” they may not be repaired in the same way as talking to another person [158]. In chapter 5, we study repair strategies that a chatbot could adopt to tackle these problems – providing evidence for the breakdown and supporting repair towards a desirable direction for the system model. We note that many commercial chatbot products are already adopting repair designs to serve these goals. One example is to ask for *confirmation* when the system has low confidence, which gives a clear signal of a potential breakdown and allows the user to initiate repair without the system mistakenly executing a task. Another example is to provide *options* of tasks that the chatbot can handle based on their proximity to the user input, which not only indicates a breakdown but also drives the interaction to the scope of the system model’s capabilities. To systematically understand user preferences for different repair strategies, including the underlying reasons and individual differences, we conducted a scenario-based study with Mechanical Turk workers (N=216) comparing eight different repair strategies.

### **1.3 Research Aim 3: To broaden chatbot’s applicability to new demographics, not limiting chatbots to technically-advanced users**

Most current chatbots targets literate, technologically-advanced users. In our initial study with first-time chatbot users (discussed in chapter 3), the participants were technologically-advanced literate Indians knowing about chatbots, but had not used them before. Even

these participants faced several hurdles in interacting with chatbots. Specifically, for two out of the eight chatbots, most participants were unsure of their core functionalities, even after interacting with them over multiple sessions. In chapter 6, we discuss our work focusing on developing chatbots for low literate novice smartphone users who have little knowledge or preconception of the chatbot technology. Specifically, we take the conversational agent on the ubiquitous smartphone to rural farmers in the developing world, broadening its scope to a new demographic. This new user population and usage require answers to questions concerning (a) how to encode farming related queries efficiently in a conversational system; (b) the robustness of speech and language technologies in the local language; (c) the acceptability and usability of chatbot technologies for rural farmers; and (d) interaction modality preferences of the farmer population.

To answer these questions, we designed *FarmChat*, a conversational agent to meet the information needs of rural Indian farmers. Speech-based conversational interface has several potential benefits for our targeted farmers population. Besides little requirement for literacy, it offers a natural and familiar modality that does not require a user to learn new technical concepts or interaction methods. This could be important as rural farmers often have low technology literacy and self-efficacy. Recent work suggests that interface design should support semi-literate users differently from illiterate users [73]. For semi-literate users, text can offer faster and unambiguous mode of interaction [73], while for illiterate users, the appearance of text negatively impacts their performance [132]. Further, text output allows more flexibility to process information and persistent access to messages, which we identified as a design requirement in a formative study. Hence, we built and compared two interfaces of FarmChat: (1) *Audio-only* (input: speech; output: audio) and (2) *Audio+Text* (input: speech, button; output: audio, text, image).

Currently, FarmChat supports Hindi—the most widely spoken Indian language—and answers queries about potato farming as a use case for the study. The knowledge base embedded within FarmChat on potato farming was derived by analyzing the query logs from Kisan Call Center (KCC) and a formative study with 14 farmers and 2 agri-experts. Thereon, we ran

a task-based user study with 34 farmers in villages around the city of Ranchi in the state of Jharkhand in eastern India. From the 626 inputs provided by the participants, we found that farmers appreciated FarmChat’s precise and localized responses, showed great interest and trust on the information, and generally found a conversational agent easy to use, thus hinting that a chatbot has the potential to meet their farming-related information needs at scale and a potential use-case of chatbots to a new demographic.

#### **1.4 Thesis Statement**

Throughout this dissertation, we provide support for the thesis statement:

**The acceptability of chatbots can be increased by making them more transparent and broadening their applicability to new demographics.**

Starting with understanding challenges faced by first-time chatbots users (chapter 3), we identified several issues with current chatbots. Our research aims at proposing solutions for three of the identified problems. First, there is a mismatch between the chatbot’s state of understanding (also called *context*) and the user’s perception of the chatbot’s understanding. We propose *Convey* (chapter 4), short of CONtext View, a window added to the chatbot interface that displays the context of the conversation to the user, so that both the user and chatbot has the same mental model. Second, chatbots have low retention rate due to dialog failures. We compared eight different repair strategies (chapter 5) – five strategies adopted from the literature, and three novel strategies proposed by us based on the concept of explainable Artificial Intelligence. Third, most current chatbots are targeted towards technically-advanced users only. In chapter 6, we focus on developing chatbots for low literate novice smartphone users who have little knowledge or preconception of the chatbot technology. Specifically, we take the conversational agent on the ubiquitous smartphone to rural farmers in the developing world, broadening the scope of chatbots to a new demographic.

## Chapter 2

### **RELATED WORK**

Our initial user study (details in chapter 3) evaluates state-of-the-art conversational agents on the Facebook Messenger platform, in order to identify limitations of current chatbots. Hence we start with discussing prior work related to the evaluation of conversational agents. Before that we provide an overview of the basics of a conversation system. Based on the findings from the initial user study, we developed three different chatbots, each aiming to tackle a major issue identified with the current chatbots. First, to resolve the mismatch between the chatbot’s understanding and the user’s perception of its understanding, we discuss the basics of a conversation system (including the importance of context), along with providing our view on the stale user interface of current text-based chatbots. Second, to understand breakdowns and recovery strategies in human-agent interaction, we delve deep into repairs in human communications theory and human-robot interactions. Third, to develop chatbots for a new demography, specifically low literate farmers, we provide an overview of previous works in developing user interfaces for illiterate users.

#### **2.1 Basics of a Conversation System**

A conversation system identifies *intents* and *entities* from user’s input, to understand the meaning of user text. The user’s *intent* is the current goal or purpose of their interaction with the chatbot. The *entities* add value to that purpose and narrow it further to make it specific. For any chatbot, the intent and entity types are defined by the chatbot designer based on the purpose of the chatbot. As the conversation involves multiple back-and-forth rounds between the user and the chatbot, the conversation system maintains *context* to keep track what the user and the chatbot have been discussing. The context values comprise of

a combination of intents and entities. Therefore, without context, a user’s new input would be analysed completely oblivious of their previous inputs.

As an example, here is a typical conversation with a delivery-ordering chatbot:

Human: *i want to order a hawaiian pizza*

Bot: *ok, anything else?*

Human: *yeah make that medium size, and add a coke*

From the first message by the user, the chatbot recognizes that their intent is to *order food*, with the entity being ‘*hawaiian pizza*’. As the conversation proceeds, the chatbot maintains the context of pizza ordering so that ‘*medium size*’ can be related to ‘*hawaiian pizza*’. Without the propagation of context, ‘*medium size*’ is just another entity which is not attached to any intent. Thus, maintaining appropriate context of the conversation is crucial to a chatbot’s success [123, 169].

## **2.2 Evaluating Conversational Agents**

Several recent works [105, 123, 124, 127, 199] evaluated chatbots from an HCI perspective. Luger and Sellen [127] evaluated speech-based CAs that act as virtual personal assistants, specifically, user interactions with Siri, Google Now and Cortana. They found that users restrict their usage to simple tasks such as setting alarms or reminders. In their study, the principle use-case of speech-based chatbots turned out to be enabling hands-free interactions that save time [127]. One of their central findings is participants complaining about the inaccuracies in speech-to-text conversion. Similarly, Jiang *et al.* [105] evaluated different tasks in Cortana, a speech-based chatbot, and found major issues with the quality of speech recognition and agent’s intent classification.

In contrast to the speech-modality of the studied chatbots, the text-modality have very different user expectations and interaction patterns. Thies *et al.* [199] conducted a Wizard-of-Oz study with 14 participants to understand chatbot personalities that are most compelling to young, urban users in India. They simulated interactions with three hypothetical chatbots

with varying personalities. Participants wanted a chatbot which can add value to their life by making useful recommendations, endowed with a sense of humor, while being reassuring, empathetic and non-judgmental [199]. Thus the paper shows that users have very high expectations from chatbots. However, it lacks insight on how well these expectations are met by the current chatbots. Liao *et al.* [123,124] studied deployment of a Human Resource (HR) chatbot in a workplace environment. Apart from functional usage, they found participants getting involved in playful interactions with the chatbot, which are rich signals to infer user satisfaction with the chatbot.

Contrary to the previous works, we study chatbots on the Messenger platform where the input is limited to text or button entry. Furthermore, our study captures user interactions with a wider variety of chatbots built specifically for different domains and hence equipped with differing capabilities. Finally, participants in [105, 127] studies were ‘regular’ users of speech-based CA, participants in Liao *et al.* [123, 124] studies were new office joining employees, and participants in Thies *et al.* [199] were in a single-session controlled lab study, while we focus on studying the real-world experience of first-time chatbot users over multiple sessions interacting with 8 different chatbots across domains on the Messenger platform.

### **2.3 Stale User Interface of Chatbots**

In spite of the growing industry adoption and the advancements in AI to make chatbots ‘smarter’ and more ‘easy-to-use’, the user interfaces of chatbots have not evolved much. They still closely resemble a messaging interface, wherein a user or a bot response results in a message bubble. While some chatbot platforms may have a few multimedia and interactive elements (such as buttons, hyperlinks, carousels, gifs, videos, and so on) to enhance interactivity, the essence of chatbot interfaces has remained unchanged. The main benefit of persisting with such an interface is that it is highly flexible and familiar to anyone who has used a messaging app before. In contrast, each website/app has its own interface, thus incurring a small learning curve [69]. We augment the familiar chatbot user interface with *Convey* (discussed in chapter 4). The aim of *Convey* is to enhance the effectiveness of using

chatbots without losing the flexibility afforded by the messaging interface.

## **2.4 Breakdowns, Repairs in Communication and Human-Agent Interaction**

Recently there has been a growing volume of research on human-agent interaction. A common theme shared by work studying everyday use of conversational agents is users’ struggle with natural language interactions [123, 124, 127, 137, 158]. Though this is new to the HCI community, linguistics and sociologists have long been interested in studying repairs in human communications, defined as “the replacement of an error or mistake by what is correct” [176]. Schegloff *et al.* made the distinction between self- and other- repair [176], referring to the correction made by the speaker or the listener, respectively. A distinction is also made between the initiation and the outcome of a repair. The person who initiates a repair is not necessarily the one who concludes it. Empirically, Schegloff *et al.* concluded a preference for self- over other-repair regardless of who initiates it.

Repair is also frequently studied under the framework of *grounding in communication*, proposed by Clark and Brennan [56]. Grounding describes conversations as a form of collective action to achieve common ground or mutual knowledge. As the speaker presents an utterance, *evidence* of understanding, whether explicit or implicit (*e.g.*, providing a correct response), is expected. If there is a lack of evidence or presence of negative evidence, the speaker may choose to initiate a repair. The theory uses the concept of *cost* to explain why a repair strategy is used, or a breakdown is ignored without repair. For example, formulation cost predicts that a speaker prefers simple ways of rephrasing (*e.g.*, correcting a partial sentence) instead of a perfect description to repair. It also explains the preference for self- over other-repair by minimizing turn-taking cost (potential turns needed) and fault cost (*i.e.*, perceived at fault).

In a serial work to adapt the grounding framework for human-computer interaction [41, 44], Brennan highlighted that the understanding models are private to each party, and dialog partners can only estimate how to collectively converge them. When the dialog partner is a machine, its private model or capability is significantly mismatched from the human speaker,

posing critical challenges for grounding or repair. Brennan derived a theory-driven model for a spoken dialog system to adapt its feedback to its understanding status by explicitly indicating in which state the breakdown happens, such as attending, recognizing, interpreting, or acting.

Myers et al. studied chat logs of a voice based interface (VUI) to identify types of errors and users' coping tactics [137]. They found NLP errors – misreading a user's utterance – to be the most common type of error, and users engaged in a variety of tactics including hyper-articulation, simplification, and providing more information in attempts to repair. But, some users also shied away from repairing by quitting, settling, or starting a new task. Porcheron *et al.* conducted a field study of user interactions with Amazon Alexa [158] at homes and found that a significant amount of interactions were dedicated to repair. They attributed the challenge of user repair to a lack of indicators of the form of trouble in Alexa's error messages: “[Alexa] provides no mechanism for further interaction, and does not make available the state of the system, allying the VUI with notions of a ‘black box’ ”. This conclusion echoes a long-standing concern on the limitation of conversational agent interfaces – a lack of transparency on system status and affordance [127, 189].

Besides these studies providing a descriptive account of breakdown and repair, work that suggests design solutions to support the repair process has been limited. A distinction should be made between *agent-initiative* and *user-initiative* systems [92]. In the former case, systems with the dialogue initiative can restrict users' responses by asking close-ended questions. It is in the latter case where breakdowns are common, as users can ask free-form questions and repairing breakdowns is challenging because users are uncertain about the system's status and capabilities. Information search agents, such as Apple's Siri and Amazon's Alexa are mostly user-initiative. They are also considered *goal-oriented*, or *utility-driven*, because users have an information goal to achieve from the interaction. For *non-goal-oriented* chatbots (i.e., for chit-chat), Yu et al. [225] enumerated a list of design strategies such as repeating parts of the user utterance, switching topics, and telling jokes, but they aim to engage users for further interaction instead of supporting repair.

The related human-robot interaction (HRI) community has studied designs to mitigate the negative effects of robot breakdowns. With humanoid robots, the focus has been on social behaviors that make users more tolerant or willing to help. For example, multiple studies explored using politeness and apology strategies to request help when the robot malfunctions [71, 121, 193]. Most relevant to ours is the work by Lee *et al.* [120]. Using a scenario-based survey, they studied three strategies for a robot to recover from a breakdown: apologies, compensation, and providing options. They found individual differences in repair preferences based on service orientation: those with a relational orientation preferred apologies, while those with a utilitarian orientation (interactions with the bot are purely transactional) preferred compensation [120]. In our Resilient Chatbot work (chapter 5), we borrow the methodology of a scenario-based survey as it provides a means to gather a large quantity of data for our collection of repair strategies, and it allows us to strictly control the interaction process and outcomes to evaluate the *perception* of different repair strategies.

## **2.5 Explainable Artificial Intelligence**

In recent years, there has been a surge of interest in explainable AI [20, 83]. In explainable AI, people are provided with an explanation on how an intelligent system arrived at a particular decision or conclusion. Prior work suggests that people face difficulties when they are not able to understand how a technology works (*e.g.*, smart thermostats [220] or conversational systems [123]), which limits their usage of the technology to non-complex tasks [127]. In a desire to broaden the adoption of autonomous AI systems, Shneiderman *et al.* [190] argue that a level of human control ought to be maintained via interfaces that allow users “*to better understand underlying computational processes.*” In line with this view, AI researchers have begun organizing their own work on machine learning systems around three dimensions – fairness, transparency, and accountability (FAT) – in order to ensure due process and understandability in decision-making [25].

For text classifiers, explanations are generated from their features, such as the words used in the documents they classify. A common approach is to highlight keywords in a

document that have the highest weights for the classifier’s decision – “this document is classified as sports news because it contains the keyword *football*’. Stumpf *et al.* explored user response for machine learning algorithms explained by keyword highlighting, together with explanations using rules of words and similar sentences [195], and showed that exposing the mechanism of an algorithms could enable people to provide rich feedback for improving the algorithm. For the Resilient Chatbot work (chapter 5), we draw inspiration from such recent works on explanation of machine learning algorithms by highlighting [76, 83, 195].

## **2.6 UIs for Low Literate Users, especially Farmers**

Most chatbots target technically-advanced literate users. No prior work exists developing conversational agents for low literate users (in particular, farmers). However, designing user interfaces for low literate populations is a growing area of research [73, 110, 130, 132, 153]. It spans multiple application domains, including agriculture [61], health care [82], citizen journalism [135], video search [61] and social networking [133]. Research has shown that users with low levels of literacy perform better with user interfaces which use minimal or no text and represent textual information using graphics/photographs and audio [130, 132]. Even interfaces that use graphics liberally, such as a job search web portal for illiterate users [132], rely on audio to provide descriptions and instructions. Voice as an interaction modality is well-suited for low literate users, as it is a natural means of expression. Voice-only citizen journalism portals [135], voice-based Q&A forums for rural farmers [154], and IVR systems [82, 108] are a few successful examples, relying heavily on speech for input and audio for output. With audio as the output modality, researchers have compared speech versus DTMF/keypad for input and obtained contrasting results with low literate users [82, 153, 186]. Interestingly, most low literate users are numerically literate; hence, using numbers (both for input in keypad buttons and output as text) has been found to be acceptable [132]. This has paved the way for multimodal interfaces [61, 133] that embed graphics, voice, and numbers for low literate users.

Recent research highlighted the interaction differences between illiterate and semi-literate

users and suggested to treat the two groups differently in interface design [73]. Comparing them in tasks on Audio+Text versus Text-only interfaces, Findlater *et al.* [73] found that text was important for semi-literate users since it offers a faster and less ambiguous mode of interaction. Importantly, text allows for opportunistic language learning. Interestingly, social factor also results in the preference of a text-based interface since it avoids the stigmatized perception of illiteracy [110]. For illiterate users though, the appearance of text negatively impacts their task performance [73, 132]. Hence, we designed and compared two FarmChat interfaces: Audio-only and Audio+Text (details in chapter 6).

ICT4D (Information and Communication Technologies for Development) research has contributed immensely in developing technological solutions to help farmers in developing regions with their information needs [64, 75, 154, 163, 187]. Farmers are usually located in rural areas with low literacy levels; hence, a majority of the proposed solutions rely heavily on voice as an interaction modality. Two of the widely adopted approaches are automated calls providing agriculture-related knowledge and IVR systems. These are highly scalable solutions, but are limited to providing generic crop-related advice, which may not work for a majority of farmers due to variations in crop, soil type, climate, *etc.* Automated calls have been adopted by several governments around the world, including India's [143]. Khedut Saathi [148] took automated calls a step further, allowing farmers to forward the received audio message to five other phone numbers. IVR systems use a computer-based back-end with a keypad/voice-based input, and audio output to provide farmers with relevant information related to weather, fertilizers, and market prices [163, 187]. Accessing the hierarchical navigation is a major usability issue with IVR systems [130]. Instead of restricting their system to audio alone, Digital Green [75] focuses on creating videos by farmers and uses human-mediated instruction for disseminating these videos to other farmers. VideoKheti [61] proposed a multimodal method for illiterate farmers to search specific videos.

Other solutions offer custom advisory services to farmers. The Indian government program KCC (Kisan Call Center) [78] allows farmer to dial a toll-free number and get responses to their specific queries. Since the demand for such information is too large to be met by a

manually operated system, most of the KCC calls go unanswered because the phone lines are usually busy [31]. Aavaaj Otalo [154] proposed a voice forum for asking agriculture-related questions to experts and peers. It is an asynchronous system, meaning that responses are not given in real-time. Furthermore, questions can be answered by anyone, which may result in incorrect answers and distrust towards the system.

All these work suggests that to fully support the agriculture-related needs of farmers, constant, real-time access to specific information is needed. None of the systems in prior work achieve all of those criteria; FarmChat chatbot (discussed in chapter 6) is an attempt to fulfill these requirements.

## Chapter 3

# EVALUATING AND INFORMING THE DESIGN OF CHATBOTS

### **3.1 Introduction**

Chatbots are still in their nascent stage: They have a low penetration rate as 84% of the Internet users have not used a chatbot yet. Hence, understanding the usage patterns of first-time users can potentially inform and guide the design of future chatbots. In this chapter, we report the findings of a study with 16 first-time chatbot users interacting with a curated list of eight chatbots over multiple sessions on the Facebook Messenger platform. We chose Messenger as it hosts the maximum number of chatbots [49] and is the second-most popular text-messaging app [194]. For our analysis, we combined qualitative findings from the semi-structured interviews with the quantitative findings from  $\sim 10,000$  messages that the participants exchanged with the chatbots. We found that users preferred chatbots that provided either a ‘human-like’ natural language conversation ability, or an engaging experience that exploited the benefits of the familiar turn-based messaging interface. Furthermore, we identify key implications on the design of chatbots and the design of messaging interface provided by the chatbot-hosting platform. Chatbot designers should ensure that chatbots understand and sustain conversation context, provide a clear and ongoing indication of the chatbot capabilities, engage in small talk, indicate when the chatbot fails to perform a task, and end a conversation gracefully.

### **3.2 Study Design**

To study user interaction with chatbots, we chose a set of chatbots that are representative of the diverse use-cases of chatbots. This section describes our choice of chatbots and

participants, and further continues with the study procedure.

### 3.2.1 Chatbots

Several messaging platforms (such as Facebook Messenger, WeChat, Kik, Slack, Telegram and Skype) support chatbots. In addition, there are individual chatbots, such as Google Assistant, Microsoft Zo, *etc.* When confronted with the choice amongst these for the study, we applied three guiding principles: (a) the study would focus on a single platform so as not to confound the comparisons across chatbots with platform variations, (b) the platform must be familiar to the users, and (c) the platform must have received significant developer interest as evidenced by its chatbot catalog. Facebook Messenger was the clear choice. Users are familiar with the platform as it is the second most popular messaging platform [194] after WhatsApp (which does not as yet support chatbots). Also, it has been the favorite platform amongst developers with over 34,000 chatbots as of Nov 2016 [54]. Thus, our study exclusively focuses on chatbots on the Messenger platform.

The aim of the chatbot selection process was to select a set of chatbots on the Messenger platform with which a new user is most likely to interact. We started the selection process by considering the top 100 Messenger chatbots listed on Chatbottle [4]. In 2016, Chatbottle was the only search engine and ranking provider of Messenger bots. Based on the chatbot descriptions, we identified eight major domains: News, Travel, Shopping, Social, Game, Utility, Chit-chat, Entertainment. For each of these domains, we selected the highest rated chatbot (using the rating from the Chatbottle [4] website), while ensuring that the chatbot has received more than 1000 likes on Facebook and the chatbot is functional in India, as the participants and authors of this study were based in India. Thus, we selected chatbots that are popular and diverse. (Note: As chatbots and their rankings are continuously evolving, we only considered the state of the chatbots and their ratings in Nov 2016). The selected chatbots are described in Table 3.1.

Table 3.1: Description of the selected eight chatbots for the user study.

<b>Chatbot</b>	<b>Domain</b>	<b>Description</b> (in their own words, or *from <a href="https://chatbottle.co">https://chatbottle.co</a> )
Alterra	Travel	Hi! I'm an AI travel agent. I can book flights and book hotels. If you haven't decided where to go I can give you vacation ideas, and tell you what to see there.
Call of Duty	Entertainment	*Experience the excitement of Call of Duty like never before.
chatShopper	Shopping	Hi, I'm Emma, your personal shopping chatbot. I can search for fashion items, shoes and accessories.
CNN	News	Chat with me for the news as it unfolds. I'll send you top stories every day, or you can ask me about a topic that you want to learn more about.
Hi Poncho	Utility	Hi, I'll give you a personal weather forecast that will make you smile, whatever the weather.
Pandorabots	Chit-chat	Hi, I'm Mitsuku! *You need never feel lonely again! Mitsuku is your new virtual friend and is here 24 hours a day just to talk to you. She learns by experience, so the more people talk to her, the smarter she becomes.
Swelly	Social	*Vote for cool stuff and help other people with their daily decisions. A swell contains a question and 2 options. A or B? High Heels or Sneakers? Hot or Not? Start voting!
Trivia Blast	Game	*Trivia Blast is the new quiz game to play with the bot or between friends.

### 3.2.2 Participants

In order to understand users first-time experience interacting with chatbots, we recruited individuals with no prior experience with chatbots. Regular chatbot users might have been accustomed to chatbots' limitations and would have learned to adapt. As participants were not compensated to participate in the study, we recruited individuals with strong *intrinsic motivation* to explore and experience chatbots as a new technology. Hence, we ensured these requirements were fulfilled: (a) the participant must be an avid Messenger user, using it at least once every 4 hours, and (b) the participant must be a technology-enthusiast, using phone, tablet, and/or laptop for 10+ hours a day. Also, the participant must be based locally for face-to-face interviews. To recruit participants, we used word-of-mouth and snowball sampling. Prospective participants were asked to fill out a questionnaire. In a week, we received 31 responses to our questionnaire, out of which 16 fulfilled our criteria.

Sixteen participants (10 male and 6 female, mean age of 32.1 years,  $sd=6.9$ , age range 23-45 years) were recruited. Ten of them had an engineering background, and the remaining six of them were from non-technical backgrounds, including operations, finance and social sciences. Most participants were young technology enthusiasts. While this is a very specific sample from the general population, the participants adequately represent technology early adopters who will likely constitute the majority of chatbot users in the near future. All of them understood chatbots at a conceptual level but had no prior experience with them. All of them had a Bachelor's or higher degree. None of them were native English speakers, but rated themselves to be fluent in English. Participants self-reported an average 11.8 hours ( $sd=1.3$ ) of daily computer and phone usage. Ten of them reported using Messenger every hour of the day, while the rest reported using it every four hours daily. All of them reported themselves to be frequent readers of tech-related news articles, spending on an average 0.5 hours ( $sd=0.2$ ) every day.

### 3.2.3 Procedure

During the first face-to-face meeting with the study facilitator, participants were informed about the definition of chatbot with a few generic examples and the goal of the study. Participants were provided with a list of 8 chatbots (*Alterra* [7], *Call of Duty* [8], *chatShopper* [9], *CNN* [10], *Hi Poncho* [11], *Pandorabots* [12], *Swelly* [13], *Trivia Blast* [14]) in a randomized order to counteract order effects. They were asked to interact with each chatbot for at least 3-5 minutes daily for the next three days. The facilitator sent each participant a personalized daily reminder on Messenger. The reminder consisted of web links, which opened a direct conversation with the chatbot on Messenger. The participants were not instructed on how to interact with the chatbots, what the chatbots were about, or what kind of tasks to perform using the chatbots. This was done to encourage exploration and open-ended usage of the chatbots, to capture a range of perspectives. Instead of asking participants to perform specific tasks with each bot which has been found to be insufficient for chatbot evaluation [105,168], we chose an exploration-based study because of these three reasons: (a) first-time users tend to explore which in turn can help to understand their learning curve, (b) we did not want to bias or influence participants' first interaction with the chatbots in any manner, and (c) with exploration, participants' opinions would not form based on the particular tasks that they were asked to do, thus exploration has potential to provide varying observations across participants.

After the three days of interaction with the chatbots, the participant had a face-to-face semi-structured interview with the facilitator. Interview questions sought to elicit participants' understanding of the chatbots, their perceived benefits/limitations, any interesting conversations and/or experiences, and areas for improvement. The interview durations ranged from 40-60 minutes. At the start of the interview, participants were asked to rank the chatbots and rate them with respect to different metrics, including learning curve, frustration level, and fun to use [86]. All the interviews were conducted in either the participants' office or home. The interviews were voice-recorded after receiving permission from the par-

ticipants, and later transcribed in English. At the end of the interview, a copy of the chat log (in HTML files) was downloaded by the facilitator after taking participants' permission to use it only for research purposes. It should be noted that no keyloggers were used for this study. This is for two reasons: privacy concerns, and freedom to switch between devices. Pre-study demography questionnaire, chat log files, post-study rating questionnaire, and interview transcripts, were used in our data analysis.

### **3.3 Quantitative Data Analysis**

In this section, we present the quantitative analysis of the chat logs. The analysis shows that the participants were engaged throughout the study, as hinted by their high level of participation. In total, participants and chatbots exchanged 9968 messages interacting for 25 hours (Figure 3.1a, 3.3) across 379 sessions. Six participants predominantly used the chatbots on their phone, one on her tablet, and nine on their laptop.

Conversational agents are usually evaluated using three measures – Task Completion Rate (TCR), Number of Turns, and Total Time [208]. TCR is not relevant for our study, as participants were not asked to complete any specific task with the chatbots. Number of Turns is defined as the number of messages exchanged between the user and the bot, and we refer it as ‘Message Count’. Message Count and Total Interaction Time are indications of how effectively the chatbot can engage a user. Also, from an HCI perspective, we explore the types of interactive elements constituting the message. Hence we focus on *Total Interaction Time*, *Message Count*, and *Interactive Elements* in this section. We conducted a mixed-model analysis of variance – on the total interaction time, total number of messages exchanged between the chatbots and participants, total chatbot message count, total participant message count, and average character length per message – treating Chatbot as a fixed effect and participant as a random effect.

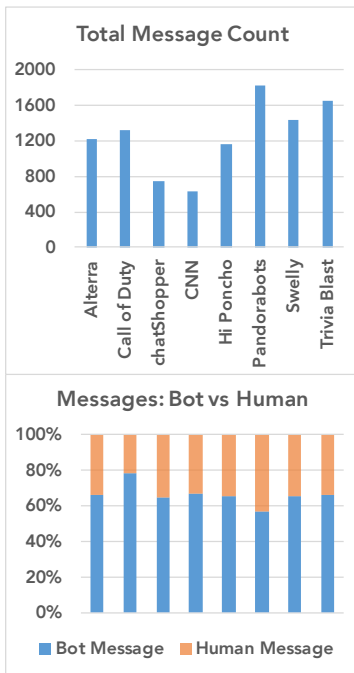


Figure 3.1: (a) *Top*: Total chat messages exchanged; (b) *Bottom*: Division of those messages among chatbots and participants.

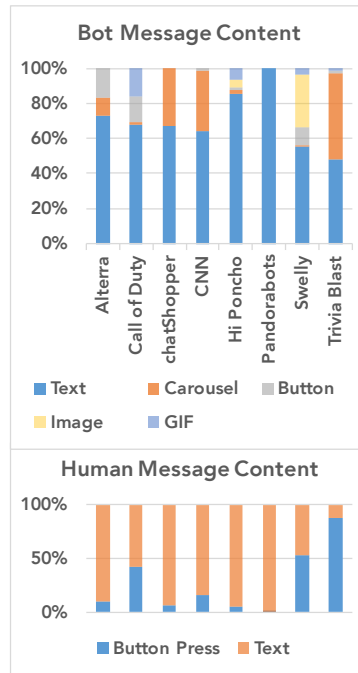


Figure 3.2: (a) *Top*: Content of the chatbot messages; (b) *Bottom*: Content of the participant (human) messages.

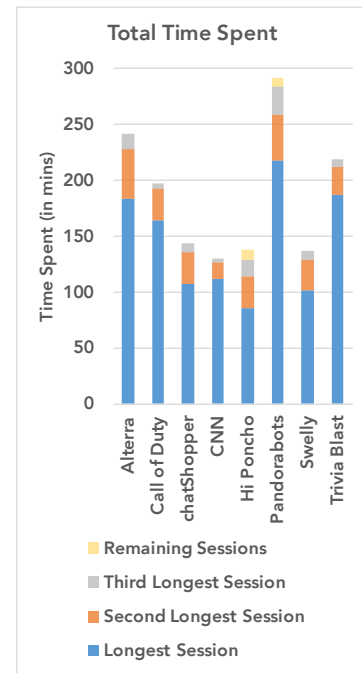


Figure 3.3: Interaction Time spent by the participant per chatbot.

### *Total Interaction Time*

Over the course of the study, the time spent by a participant interacting with the chatbots was  $93.5 \pm 53.9$  mins, and the number of sessions was  $23.7 \pm 5.6$ . For the total time of interaction, no significant main difference was found among the chatbots ( $F_{7,103}=1.77$ ,  $p=0.1$ ) (Figure 3.3). This hints that the participants spent similar amounts of time with each of the chatbot as the study facilitator instructed. Figure 3.3 shows time spent by the participant with each chatbot, and a session-wise split of duration. Interestingly, only with *Pandorabots* and *Hi Poncho*, the participants interacted for four or more sessions. In *Hi Poncho*, it was mostly chatbot-initiated in the form of weather notifications, while with *Pandorabots*, it was always participant-initiated interaction.

### *Message Count*

Out of the total 9968 messages, 65.8% of messages were by the chatbots and 34.2% by the participants. The ANOVA test showed a significant main effect of Chatbot on the total messages exchanged between the chatbot and the participant ( $F_{7,103}=3.93$ ,  $p<0.0001$ ) (Figure 3.1a). This means that although the participants spent similar amount of time with each chatbot, the total count of messages exchanged significantly varied across the chatbots. This prompted us to investigate pairwise differences. We employed Tukey’s HSD procedure to address the increased risk of Type I error due to unplanned comparisons. We found that the number of messages exchanged with *Pandorabots* and *Trivia Blast* were significantly higher than *chatShopper* and *CNN* ( $p<0.01$ ). This may be due to the fact that *Pandorabots* falls in the Chit-chat domain. Its known that users tend to chat more with a conversation partner, rather than a human assistant [196]. In the case of *CNN*, the news article opens up in a new browser window (leaving the Messenger interface), hence limiting the interaction with the chatbot.

With respect to chatbot-only message count, we found that *Call of Duty* was significantly higher than *chatShopper* ( $p<0.01$ ) and *CNN* ( $p<0.01$ ), with 78% of the *Call of Duty* total

messages comprising of chatbot messages (Figure 3.1b). This may be because *Call of Duty* is verbose, and continues with the game story regardless of human input. The count of participant messages to *Pandorabots* was significantly higher than *Alterra*, *Call of Duty*, *chatShopper*, *CNN* and *Hi Poncho* ( $p < 0.01$ ) (Figure 3.1b). This may be because *Pandorabots* is a chit-chat bot, and requires minimal mental effort (as in the case of texting a friend).

### *Interactive Elements*

We analyzed the composition of chatbots' messages (Figure 3.2a) and participants' messages (Figure 3.2b). *Pandorabots* lacked interactive elements (Figure 3.2) as it is a completely text-based chatbot, while *Trivia Blast* (Figure 3.2a) was predominantly click-based. In Messenger, human input was limited to text messages or button clicks (Figure 3.2b). As no key loggers were used, only the button presses that resulted in input text (e.g., Figure 3.4, 'book hotels' button) were logged and hence considered in the analysis. Other button presses such as buttons which disappear after selection (e.g., Figure 3.5, 'auto-suggestion buttons') or buttons that lead to external websites (e.g., Figure 3.4, buttons as part of 'carousel') could not be counted. Overall, participants typed 43,844 characters (maximum 35.3% on *Pandorabots* followed by 20% on *Alterra*). The average number of characters typed by the participants per message showed a significant main effect for Chatbot ( $F_{7,103} = 11.4$ ,  $p < 0.0001$ ). On *Alterra* and *CNN* participants typed significantly more number of characters per message than *Call of Duty*, *Swelly* and *Trivia Blast* (all  $p < 0.01$ ). This indicates that *Alterra* ( $23.5 \pm 19.5$  characters/message) and *CNN* ( $21.7 \pm 26.4$ ) needed longer text input as in a search query, while *Call of Duty* ( $13 \pm 10.1$ ), *Swelly* ( $9.2 \pm 7.6$ ) and *Trivia Blast* ( $12.7 \pm 9.9$ ) were predominantly click-based.

Chatbot messages comprise of several interactive and rich media elements (Figure 3.2a). 70.1% of messages comprises of only text. The character count per chatbot message response was highest for *CNN* ( $84.3 \pm 60.2$ ), closely followed by *Hi Poncho* ( $79.4 \pm 55.1$ ) and *Call of Duty* ( $73.8 \pm 38.1$ ), and lowest for *Pandorabots* ( $41.9 \pm 38.7$ ). This hints that a few chatbots were verbose in their response, which participants complained about (specifically *Call of*

*Duty*). After ‘text’, the second most common chatbot message content was ‘carousel’ (Figure 3.4), constituting of 14.5% messages (Figure 3.2a). We define *seat* as an UI element that comprises of an image with a header and 1-3 buttons below the image. Messenger provides a way to combine multiple such *seats* to form a horizontal scrollable *carousel* (Figure 3.4). Carousels were extensively used to show news in *CNN*, quiz questions in *Trivia Blast*, and shopping items in *chatShopper*.

### 3.4 Findings: Qualitative Data Analysis

In this section, we discuss the comments made by the participants on their experiences with the chatbots as recorded during the face-to-face semi-structured interviews, followed by the chatbot ratings collected in the post-study questionnaire. The interview coding and analysis was done in an iterative fashion. Three of the co-authors met as a group to explore the data. Each interview transcript was projected on a large screen and discussed to identify interesting comments. In total, 957 comments were identified, which were coded iteratively. The three co-authors met multiple times as a group to refine and coalesce the initial 49 codes into 4 high-level themes representative of the data – *functionality*, *conversational intelligence*, *personality*, and *interface*.

The purpose of this study was not to compare the different chatbots; however, most participants’ comments were chatbot-specific, as each chatbot was very different in its capability and domain. Thus, at the end of each theme, we provide a summary generalizing the bot specific comments. Note:  $P_{i,j}$  refers to a comment by participant  $i$  for the chatbot  $j$ .

#### 3.4.1 Functionality

The first theme that emerged concerns the functionality of a chatbot. In other words, did the chatbot do what it is supposed to do, and if so, how good was it? Fourteen participants praised at least one chatbot for successfully accomplishing its *primary task*. Participants defined primary task as the task stated as part of the chatbot description.  $P_8$  liked *Trivia Blast* as it helped him pass time during commute. However, for *CNN*, participants complained

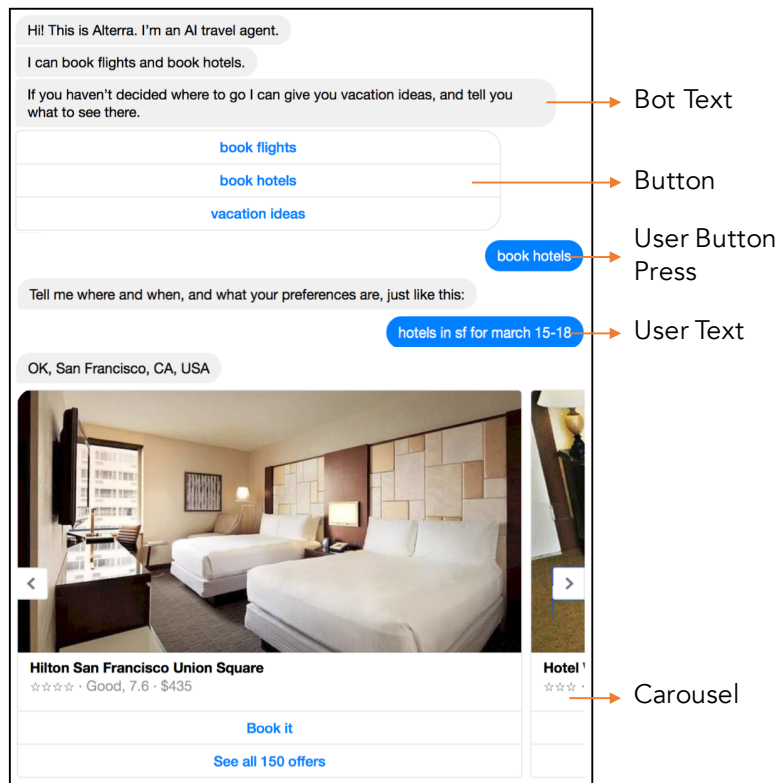


Figure 3.4: *Alterra* chatbot showing different UI elements

that it “shows mostly old stale news” -  $P_{9,CNN}$ , and “It doesn’t even understand weather, Pakistan, migrations... doesn’t work at all” -  $P_{13,CNN}$ . Participants commented extensively about the subjective utility of the primary task. Participants found a few chatbots to be not useful, either because those chatbot domains were not useful for their specific use cases, or the chatbots were found lacking compared to their website/app counterparts. For instance,  $P_{11}$  was not interested in video games, hence she didn’t like *Call of Duty*. Participants mentioned that for certain domains, such as news and utility, chatbots are a good fit. “News by bot makes sense, though CNN ain’t good.” -  $P_{12,CNN}$ .

Comparing these chatbots with existing alternatives, including websites, phone apps, and search engines, to accomplish the same task was a constant theme across participants.  $P_{10}$  praised *Swelly* as an “awesome idea” as there is no alternative, “I can’t google for opinions”.



Figure 3.5: CNN chatbot showing different UI elements

However, participants mentioned that flight-booking websites are better than *Alterra*, as “*I can quickly browse through hundreds of flights*” - P<sub>2,Alterra</sub>. This again relates back to the domain of the chatbot, as certain domains which requires choosing from a large number of available options (such as shopping) are less suited for a chatbot interface, compared to domains requiring specific answers (such as news, weather).

As the participants were interacting with eight chatbots at the same time, they intrinsically compared them against each other, and hence were setting the threshold of acceptable failure for each chatbot, not only based on alternatives, but also based on their simultaneous experience with the other bots. Participants appreciated chatbots which were able to perform “*tough*” tasks, where they initially expected the chatbot to fail, thereby exceeding their expectation. For *Hi Poncho*, P<sub>15</sub> expected it to just provide weather information based on the location input, however found “*It worked even for ‘rain in Bangalore’, ‘hiking in London’, ‘umbrella in Seattle’. It just works!*”. Similarly, P<sub>1</sub> liked *Alterra*: “*It was able to understand ‘second Sunday of march’, we can’t do (that) on a website*”.

When the chatbots did not fulfill their expected functionalities or did not behave as expected, participants started doubting and blaming themselves, such as “*maybe I don’t know how to use it? or how to properly communicate with it?*” - P<sub>3,Call of Duty</sub>. This is consistent with Norman’s theory of “human error” [140]. Three participants showed these traits, and instead of resolving their problems, all of them abandoned the specific chatbot completely. When the researcher explained the specific chatbot purpose to the participants during the

interview, the participants were surprised and wanted the chatbot to “*clearly specify it in their description*”. Finally, a few participants complained that some basic functionalities were missing from the chatbots. “*CNN should understand simple search query, and provide latest relevant news*”-P<sub>13,CNN</sub>. At other times, participants were not aware of the existence of certain functionalities. For instance, eight participants did not realize that even they can post questions to *Swelly* which other *Swelly* users will answer.

*Summary:* A chatbot must accomplish its primary task, and must outperform its existing website, app, or search engine alternatives by offering diverse and/or enhanced functionalities. Moreover, chatbots must communicate their functionalities to the users, and check for domain suitability.

### 3.4.2 Conversational Intelligence

The second theme revolved around the “brain” of the chatbot, *i.e.*, its ability to converse intelligently. This represents participants’ interest in the quality of the conversation over and beyond mere functionality. The most common comments were related to a chatbot’s understanding of the input text. Participants considered this as an important criterion to determine whether its a “chatbot” or not. For *Call of Duty*, *Trivia Blast* and *Swelly*, multiple participants commented that it is “*not a chatbot, as (it) can’t chat*” - P<sub>9,Trivia Blast</sub>. The major complaint with the *Call of Duty* chatbot was that it was “*completely scripted*”, and ignores the user input text. Most participants got annoyed by it, as evident from the ratings (Figure 3.6). *Pandorabots* delighted several participants with its natural language understanding as it was able to understand and respond appropriately to most conversations: “*It is as good as talking to a human*” - P<sub>9,Pandorabots</sub>. “*It answers like my spouse*” - P<sub>1,Pandorabots</sub>. Five participants tried to “*break the (intelligence of) Pandorabots*”, similar to [199], which might be one of the reasons for maximal interaction with *Pandorabots* (Figure 3.1a, 3.3). Participants expected other chatbots which have basic “*keyword-understanding*” (such as *Alterra*, *chatShopper*) to have “*human-like*” conversational abilities. Six participants also mentioned that *Alterra*, *chatShopper* and *Hi Poncho* do not understand statements with

negation, “*I asked for not red shirts, and she started showing me red shirts*” - P<sub>5,chatShopper</sub>.

Participants discussed how chatbots handle such failures with regards to understanding users’ text or finding a suitable response. A few participants wanted the chatbot to cover-up with a smart response, while others wanted it to admit that it failed. P<sub>9</sub> asked *chatShopper* for “*eye-liner*” and it responded with socks instead. P<sub>9</sub> wanted the chatbot to admit its failure and respond with a “*big NO, the very first time... clearly stating which products she can help me with*”. Four participants were pleasantly surprised by *Pandorabots* ability to cover-up its lack of knowledge by providing smart responses. P<sub>8</sub> asked, “*among the US 2016 presidential candidate, who is more popular?*”, to which *Pandorabots* responded, “*The one who has the greatest number of fans and friends.*”

Participants were impressed with chatbots that continued a conversation by retaining *conversational context*. For instance, P<sub>1</sub> mentioned that she was “*super happy to use it (chatShopper)*” because *chatShopper* was able to follow up on her query of “*shoes*”, followed by “*in red*”. Similarly, P<sub>10</sub> highlighted that *Pandorabots* was able to understand and retain context even in a complex conversation - “*I told Pandorabots that X is my friend and Y is his wife. Later I asked her, who is Y, and she correctly said X’s wife!*” A few participants mentioned that they found mismatch between the chatbot’s real context versus their assumption of the chatbot context. “*I wasn’t sure if the bot understood ‘brown shoes’, as a few shoes were black and red*” - P<sub>14,chatShopper</sub>. Two participants expected the chatbot to retain context across chat sessions, thus providing users with personalized recommendations learned over multiple interactions between the chatbot and user. P<sub>3</sub> asked *chatShopper* to “*recommend shoes to go with the dress that I selected yesterday*”, and was disappointed by the results. This is in accordance with previous findings of maintaining a sense of continuity over time [34], similar to human conversations.

Furthermore, participants suggested several features to improve the conversation efficiency. Participants expected chatbots to proactively ask questions in order to reduce the search space. P<sub>12</sub> appreciated questions asked by *Alterra* to refine the flight search, while P<sub>5</sub> was disappointed with *chatShopper* for not asking questions. Participants also recommended

a few advanced features, such as ability to edit a previous message, either using the UI or “*using newer text message starting with an asterisk, as we do in current messaging apps*” - P<sub>5</sub>.

*Summary:* A chatbot needs to have ‘human-like’ conversational capabilities, including context preservation (intra- and inter-session), understanding of negative statements, cover-up smartly or admit failure, and ability to ask intelligent questions to engage the user in a meaningful conversation, along with helping the user with the task.

### 3.4.3 Chatbot Personality

Participants enjoyed chatbots with a distinct personality. They expected the chatbot personality to match its domain, *e.g.*, a news chatbot should be professional, while a shopping chatbot can be casual and humorous. Moreover, personalities have a strong impression, as most participants referred to *chatShopper* and *Pandorabots* with gendered pronouns (‘he’, ‘she’), while *CNN* and *Trivia Blast* were considered as tools (‘it’). Previous work with a teaching bot found that using pronouns (‘we’) rather than ‘it’ significantly correlates with student learning [146]. Most participants started their conversation with a ‘*hi*’, expecting the chatbot to respond back. Since participants expected a conversation, they assumed that the chatbot would engage in small talk. *E.g.*, “*didn’t even respond to how are you?... not even to hi*” - P<sub>15,Call of Duty</sub>. A few participants expected the chatbots to be more personal. “*She was not addressing me by my name... very impersonal.*” - P<sub>15,Pandorabots</sub>. All these – using pronouns to refer to the bot, engaging in small talk, expecting the bots to be personal in their response – hints that the participants were assuming and expecting the bots to be more human-like.

Apart from the small talk, humor was prominently mentioned by the participants. Ten participants mentioned that they had a “*fun*” conversation with *Pandorabots* and/or *Hi Poncho*, as these kept them “*engaged*” with their “*humorous*” and “*highly diverse responses*”. For instance, P<sub>6</sub> mentioned that when she asked *Hi Poncho* for weather forecast of a city, it responded with “*Cool, I DJ’ed there once. Good crowd. Right now it is 28 C and clear*”

there.”, and P<sub>15</sub> stated that when he asked *Pandorabots* “*why are you learning about humans?*”, it responded with “*Because if I know a lot about human behaviour, it will be easier to erase your species.*” This is also corroborated with earlier work [123, 124], and the chat logs showing that participants spent the highest amount of time with *Pandorabots*. Participants even asked for jokes to these two chatbots, and were delighted to find that they support such requests.

All the participants mentioned that the chatbot must explicitly convey its capabilities as part of the introduction. Twelve participants stated that they didn’t understand the functionality of *Swelly*, and four participants complained about *Call of Duty* and *Pandorabots*. “*What is it? A pseudo girl-friend?*” - P<sub>4,Pandorabots</sub>. However, none of the participants mentioned searching/googling to learn about the chatbot functionality. All of them explored chatbot capabilities using a “*trial-and-error*” method. This can be one of the reasons for participants being intrinsically motivated to interact with the chatbots. Six participants liked the fact that *Hi Poncho* “advertised” its capabilities later in the conversation, by stating, “*Try a few of these commands: Is it snowing in New York? ... Do I need an umbrella today? And if you ever need help, just type HELP.*” Without the upfront knowledge of chatbots’ limitations and capabilities, it seems that the participants assumed high potential in chatbots, but were later disappointed when the bots fail to accomplish those tasks. As part of the study design, we intentionally did not provide any information about the chatbots to the participants, thus unearthing these issues.

Finally, a majority of the participants (11) reported being annoyed with chatbots that do not end a conversation. “*It was impossible to end the conversation. I tried ‘exit’, ‘quit’, ‘stop it’, ‘end this’, still it kept talking.*” - P<sub>9,Call of Duty</sub>. According to P1, “*closure... exiting gracefully is super crucial.*”

*Summary:* Chatbot should have an apparent personality suiting its domain, which can help in retaining users. The chatbot should be able to introduce and advertise its functionalities, engage users in small talk, provide a personal touch, respond humorously, and exit gracefully.

#### 3.4.4 Chat Interface

The last theme discusses the interface that the participants used to interact with the chatbots. Although some of these comments refer to the interface choices in the Facebook Messenger platform, they are representative of users' expectations of chatbots interface beyond natural language text exchange. Messenger provides several interactive UI elements (Figure 3.2a). Eight participants appreciated interacting with the option buttons and auto-suggestion buttons. *Option buttons* appear as part of the bot message and are static in nature (Figure 3.4), while *auto-suggestion buttons* appear dynamically to reduce typing effort and disappear after one of the buttons is clicked or text is entered (Figure 3.5). P<sub>5</sub> liked *Trivia Blast* as “*it doesn't require typing, just interacted with the buttons.*” Participants cited “*time saving*” as the main reason to be in favor of buttons. This is similar to findings from a previous study [127], where in participants used speech-based CAs to save typing time.

With respect to the other UI elements, five participants liked the horizontal carousel (Figure 3.4) to view a list of catalog items. However, some felt that limiting the carousel to only five items at a time was restrictive (a limitation of the Messenger platform). P<sub>1</sub> suggested “*it (chatShopper) should keep populating more items on the right side, whenever I press this (carousel) right button.*” Moreover, participants asked for direct interaction with the object, rather than interacting with a button placed next to the object. For instance, in *Swelly*, participants ended up clicking the image several times, instead of clicking the button placed below the image.

Clicking on certain interactive elements opens the content in a new window detached from the chat interface, which six participants complained about. For instance, in *CNN*, clicking on ‘*Read this Story*’ button opens a new CNN webpage with the full news article. P<sub>5</sub> complained that the chatbot “*... has to leave the current (browser) tab. With 10+ tabs open, coming back to that tab is tricky*”. In contrast, previous work [199] recommends putting such external links as part of the chatbot response, from data collected using a Wizard-of-Oz study.

Apart from UI elements, participants wanted a persistent display of certain handy information such as description of the chatbot capabilities with a few examples, and a menu option to access the chatbot main functionalities. P<sub>10</sub> commented that in the Messenger platform, the “*chatbot description, summarizing its capabilities, disappears as soon as the ‘Get Started’ button is pressed to start the conversation*”. According to P<sub>15</sub>, “*in IVRS, it says, press star (\*) to go back to the main menu... In chatbots we should have something similar. I usually end up asking ‘show me the options again’ - and a few bots fail (to respond correctly).*”

*Summary:* Along with text input, interactive elements in the interface works in favor of chatbots. For user retention, a chatbot should have minimal external links. The interface should show certain information, including chatbot’s description and main menu, persistently to the user.

### 3.4.5 Chatbot Ratings

Participants were asked to rate each chatbot on six different metrics on a 5-point Likert scale (Figure 3.6). For all metrics, except *Frustrating*, higher score is better. *Pandorabots* was rated the best in all criteria, except *task success* (whether the bot was successful in performing the task), perhaps because *Pandorabots* is for chit-chat with no specific task to accomplish. *Trivia Blast* and *Hi Poncho* were consistently in the top three ratings, while *CNN*, *Call of Duty* and *Swelly* were in the bottom three, except that *CNN* received high ratings for *Future use*, due to its relevant domain. The ANOVA test showed a significant main effect of Chatbot on: *Fun to use* ( $F_{7,120}=5.9$ ,  $p<0.0001$ ), *Frustrating* ( $F_{7,120}=3.7$ ,  $p<0.001$ ), *Will use in future* ( $F_{7,120}=4.1$ ,  $p<0.001$ ), and *Overall* ( $F_{7,120}=3.4$ ,  $p<0.01$ ), while *Ease of use* and *Task success* did not showed any significant difference. With respect to *Fun*, *Hi Poncho* ( $3.9\pm 0.8$ ) and *Pandorabots* ( $4.2\pm 1.3$ ) were rated significantly higher than *Call of Duty* ( $2.7\pm 1.3$ ), *CNN* ( $2.4\pm 0.9$ ) and *Swelly* ( $2.7\pm 1.1$ ), with  $p<0.01$ . For *ease of use*, the sentiment towards all the bots were generally positive with *Swelly* achieving the minimal score of  $3\pm 1.2$ . Also *Call of Duty* ( $3.7\pm 1.2$ ) and *Swelly* ( $3.7\pm 0.9$ ) were found to be the most frustrating, and were

significantly worse than *Pandorabots* ( $2.4 \pm 1.3$ ) and *Trivia Blast* ( $2.5 \pm 0.9$ ), with  $p < 0.01$ . Regarding using the bot in future, the general opinion was unfavorable, still *Pandorabots* ( $3.5 \pm 1.4$ ) was voted higher than *Call of Duty* ( $1.8 \pm 1.3$ ) and *Swelly* ( $1.8 \pm 1.1$ ) ( $p < 0.01$ ). In *Overall* rating, *Pandorabots* ( $3.9 \pm 1.2$ ) was rated significantly higher than *Call of Duty* ( $2.4 \pm 1.3$ ), *CNN* ( $2.5 \pm 1.1$ ) and *Swelly* ( $2.5 \pm 1.1$ ) ( $p < 0.01$ ).

Participants were asked to rank the 8 chatbots (1 being the best). *Hi Poncho* was ranked the highest with 12 participants ranking it in the top half (ranking= $3.1 \pm 1.5$ ). *Pandorabots* (ranking= $3.5 \pm 2.7$ ) was a close second with 11, and *Trivia Blast* (ranking= $3.7 \pm 2.3$ ) was third with 8. The worst ranked was *CNN* (ranking= $6.3 \pm 1.8$ ) which was ranked in the bottom half by 13 participants. Several interesting associations between the bots emerged. All, except two, participants ranked *Hi Poncho* and *Pandorabots* in the same half (either top or bottom), while 13 participants ranked *Pandorabots* and *Trivia Blast* in opposing halves. This hints that the participants were divided into two classes: (i) preferring *Hi Poncho* and *Pandorabots*, (ii) preferring *Trivia Blast*. *Trivia Blast* indulges in no conversation with the user (it is click-based), while *Hi Poncho* and *Pandorabots* are capable of having a ‘human-like’ conversation with the user preserving the conversation context. Another association that emerged is between *Alterra* and *chatShopper*. Both performed average and their rankings were in the mid-range (3–6); 13 participants ranked them adjacent to each other. This may be because both *Alterra* and *chatShopper* provides similar functionality of e-commerce. This hints that the functionality provided by the bot played an important role in the participants’ perception of the bots.

### 3.5 Discussion

Chatbots benefit from several significant strengths - users’ familiarity with the messaging interface, seamless natural-language interface across use-cases, and the promise of personalized and evolving intelligence driving them. Still, the overall verdict is that participants’ expectations from the technology of chatbots was not met by the sampled set of Messenger chatbots. Participants were disappointed and even frustrated with mediocre natural lan-

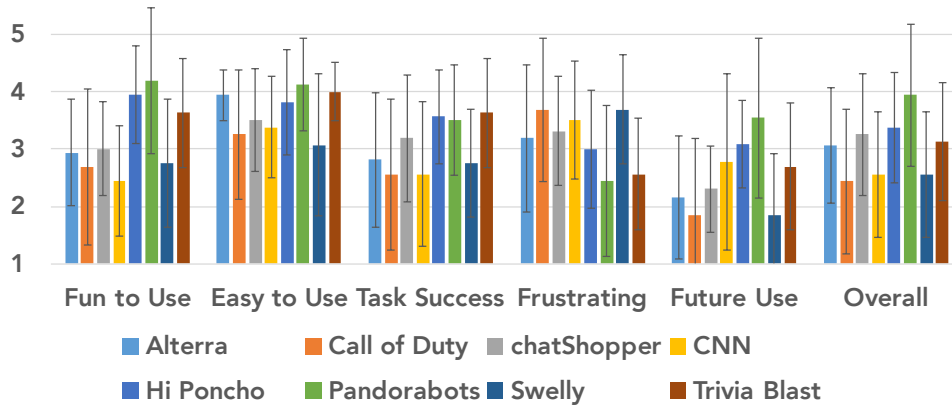


Figure 3.6: Likert-scale rating by the participants (with standard deviation shown by error bars)

guage capabilities. Particularly, they felt that the chatbots often did not understand their input text or comprehend their intention, resulting in chatbots being unable to engage or answer them efficiently. Similar to the findings of [127], users were not able to assess the intelligence of the bots. These drawbacks compounded by the limited set of features offered by the chatbots meant that the participants did not see themselves re-using most chatbots in future. Given this critical feedback, it is clear that chatbots need to evolve quickly on core competencies to engage and retain users effectively, and future attempts to address this expectation mismatch will drive innovation on generic AI abilities of language processing.

Directions for such an evolution are provided as part of the positive comments received by a few chatbots. Specifically, participants liked the witty human-like conversational skills of *Pandorabots*. It seemed to understand user’s input and could generate appropriate and smart responses. A subset of these conversational skills were well-received in *Hi Poncho*, which was perceived to have a funny enjoyable personality. On the other hand, participants also appreciated *Trivia Blast*; although it was non-chatty and click-based, it provided an engaging quiz experience within a messaging interface. These three top-rated bots thus encapsulate the key insights for future chatbots: chatbots must provide either a natural

language driven functionality served with adequate conversational delight, or an engaging app-like experience specifically designed for the familiar turn-based messaging interface.

The insights from the study complement and expand on results from earlier studies on speech-based CAs [105, 127] and chatbots [123, 124, 199]. Existing work emphasizes the speech modality with features such as ease of hands-free interaction [127] and inaccuracies in speech-to-text conversion [105, 127]. This emphasis is evident in participants choosing to perform simple tasks (*e.g.*, setting alarms), which require neither a turn-based conversation nor maintenance of context. In contrast, participants in our study performed more complex tasks (*e.g.*, planning a vacation or buying clothes). Earlier works in messaging-based chatbots evaluation [123, 124, 199] share some of our findings, including the value of playful interactions with the chatbots, and the mismatch between the user’s expectation and chatbot’s capabilities. However, previous studies were conducted with experienced users and failed to identify specifics that emerged from our study of the first-time chatbot users, such as their initial misunderstanding of the bot’s expertise, the value of ending a conversation gracefully, and the mismatch between application domain and interface. Next, we will discuss the design implications.

### *3.5.1 Design Implications for Chatbot Designers*

Below is a list of essential cross-domain design implications for developers building chatbots.

#### *Clarify capabilities at the start and on-demand*

The messaging interface is powerful, allowing unrestricted interactive patterns with natural language, in contrast to specific UI elements of websites and apps. However, a natural language interface increases the users’ expectations on the capabilities of the bot. Similar to our findings, even Luger *et al.* [127] found that insufficient visibility of the limits and capabilities of speech-based conversational agents was a major problem. To reduce the expectation gap of users, based on our study findings, we recommend that the chatbot must clearly specify what it can do. The chatbot should explicitly describe its capabilities with

examples not only as part of introduction at the start of an interaction, but also later in the conversation (as in *Hi Poncho* that was appreciated by some study participants); both during times of low engagement and after failures in the dialog. This can also help the user to transition from a novice to an expert chatbot user.

### *Evaluate application-interface match*

Chatbot designers must first identify if the application is suitable for the messaging interface. Conversational or turn-based features should be essential for the application. The application should also be restricted to the chat interface, as adding links to external webpages is not recommended (*e.g.*, *CNN*). This is in contrast with previous findings [199], as they recommend providing useful links in the conversation. Designers must ensure that they provide value over existing alternatives such as search engines, webpages and native mobile apps. This is in line with prior research that found conversational agents to be frustrating for the users when agents default to Google search [127]. Furthermore, tasks requiring exploratory search across a large number of available options, such as clothes shopping, might not be best suited for the chatbot interface. In comparison, tasks requiring minimal input, such as grocery shopping and news, fit better with the chatbot interface. This indicates the value in understanding the usage patterns – users wishing to browse versus having choices made for them – to decide if a chatbot is the right interface for a specific application.

### *Enable dialog efficiency through context resolution*

Humans need context dependence in the conversation and expect connectedness across the whole sequence of conversation [40]. Designers must aim to improve dialogue efficiency by resolving and maintaining context from earlier user messages. To resolve context, the chatbot must proactively ask intelligent questions in order to reduce the search space, and engage the user in a meaningful conversation. Maintaining context increases the input efficiency of users, as it minimizes the user input required at any instance. This ability can range from preserving context within a conversation to preserving context across conversation

sessions. Users interpret such context resolution as properties of a personalized, empathetic and intelligent chatbot (such as in *Pandorabots*).

#### *Consistent personality with small-talk and humor*

Users relate better with a chatbot that exhibits a consistent personality, *e.g.*, cat weather-expert *Hi Poncho* and shopping assistant Emma of *chatShopper*. Users expect human-like conversational etiquette from an automated chatbot, specifically introductory phrases (*‘hi’*, *‘how are you’*) (also reported in [169]) and concluding phrases (*‘bye’*). Although most designers do build dialogue flows for introductory phrases, they miss out on the concluding phrases entailing a sense of dissatisfaction among the users (as in *Call of Duty*). Moreover, designers should enrich the conversation with humor, and a large diversity in chatbot responses. In previous work [123, 124, 127, 199, 215], humor, sarcasm, and playfulness have been identified as positive traits, while excessive politeness is considered a negative trait of CAs.

#### *Design for dialog failures*

Inevitably, interaction through a free-form messaging interface can cause conversational flows that are not modeled and thus leading to a dialog failure. Designers must explicitly design for such situations, by either admitting failure and showing a list of capabilities with examples (as in *Hi Poncho*), or providing a witty conversational cover-up (as in *Pandorabots*).

### *3.5.2 Design Implications for Chatbot Platform UX Designers*

While the study focused on Facebook’s Messenger platform, the following platform-related implications are generic and applicable for most other platforms as well.

#### *Combine text-based interface with buttons and media*

The Messenger platform combines the use of text with buttons and media content such as images, GIFs, and videos. Participants found this natural and engaging. Participants

expressed dissatisfaction when the chatbot passed on the control to an external interface, such as opening a news article in a new browser window. Platforms should have a feature to allow such links to open in-line. Also, the Messenger platform provides a ‘Menu’ button persistently (Figure 3.5), though none of the participants ever used it. It seems to be under-advertised by Facebook. A messaging platform must highlight such features to the users. Finally, carousels with suggested items (such as shopping or news suggestions) should allow for a much larger number of items rather than the current limit of five items.

#### *Enable efficient input from users*

Participants commented that auto-suggestion buttons improved their interaction efficiency. Even in Luger’s work with speech-based CAs [127], time-saving was a universal theme. The messaging platform should help in reducing the interaction cost. It should allow for easy editing of user’s last few messages. This is specifically important when the edit changes a single parameter in a search query (such as changing price in a shopping/travel chatbot). Also, specific to the Messenger platform, click interaction should be enabled on images directly instead of only supporting clicks on buttons.

#### *Provide persistent view on chatbot capabilities and context*

To avoid the expectation mismatch, the platform must provide a persistent view of the chatbots’ capabilities. In Messenger, a description of the chatbot is shown at the start of an interaction, but it disappears after the first message. As an advanced feature, conversation context can also be shown to the user persistently. This will allow user to identify with the bot’s contextual state and its assumptions, and help the chatbot and the user to have the same state-of-mind.

### *Provide effective chatbot discovery*

While not experienced by the participants of this study, the authors faced the problem of discovering chatbots with specific functionalities. Each chatbot platform must enable an effective way to discover chatbots based on the bots' capabilities and popularity (a *Google Play* equivalent for discovering chatbots). As the list of chatbots keeps growing, such discovery and consequent search engine optimizations would be crucial for attracting users. Recently, in April 2017, Facebook announced launching chatbot discovery tab in Messenger [72].

### *3.5.3 Potential Future Usage of Chatbots in HCI Research*

Furthermore, as we conducted our survey of top-rated Messenger chatbots, we came across several chatbots that help with personal logging. *E.g.*, *Forksy* and *Fitmeal* log meals, *UR-report Global* is a civil reporting chatbot, and *Swelly* crowdsources votes on A-vs-B questions. Logging data as part of HCI-relevant diary studies [62] could benefit from the use of chatbots. Advantages include a normalized interface across different studies and the ability to proactively solicit feedback. For instance, a food-tracker bot knows which restaurant you are in, based on your Facebook check-in, can ask in a conversational manner if you are eating the same meal as the last time. A strong advantage of chatbots is that getting started is just a text away with virtually no barrier to entry; in contrast, a study using a custom mobile app loses participants in motivating them to install the app. Thus, in future, we expect increased use of chatbots for HCI research studies.

### *3.5.4 Limitations of the study*

Our study is an initial step towards understanding the first-time usage experience of chatbots. There are several limitations of our work. First, chatbots are continuously evolving, hence we completed the study in a 10-day period (during the second week of Feb 2017), assuming that the chatbots' remained the same. The participants' experiences thus only reflect the status of the chatbots at the time of the study. Second, our study participants were highly

educated, with all of them working in IT or financial firms. This population is at one end of the spectrum, albeit those more likely to be early-adopters, and enthusiastic to play with chatbots for extended duration. Third, there might be idiosyncrasies to this first-time chatbot users group that might not extend to other groups. Fourth, the small sample size limited our analyses. A larger number of participants is required to identify broader trends. Finally, the study was limited to the Facebook Messenger platform with our curated list of chatbots.

## Chapter 4

# CONVEY: EXPLORING THE USE OF A CONTEXT VIEW FOR CHATBOTS

### 4.1 Introduction

Recent work on evaluating chatbots and our study with first-time chatbots users found that there exists a mismatch between the chatbot’s state of understanding (also called context) and the user’s perception of the chatbot’s understanding. Users found it difficult to use chatbots for complex tasks as the users were uncertain of the chatbots’ intelligence level and contextual state. In this chapter, we discuss *Convey* (CONtext View), a window added to the chatbot interface, displaying the conversational context and providing interactions with the context values. It also provides intuitive interactions on the context values, enabling users to modify them in a simple and efficient manner. The *Convey* content gets updated as the conversation proceeds, thus always showing the latest understanding of the chatbot. To evaluate the effectiveness and usability of the proposed design, we conducted a usability evaluation of *Convey* with 16 participants. The results show that participants preferred using chatbot with *Convey* and found it to be easier to use, less mentally demanding, faster, and more intuitive compared to a default chatbot without *Convey*.

### 4.2 Design of Convey

In this section, we discuss the design of *Convey*. Please refer section 2.1 to understand the basics of a conversation system.

Our proposed design (the *Convey* box) explicitly displays the context (including assumptions) of the conversation system to the user and provides a way to efficiently interact with the context values. Below are the primary features of the proposed *Convey* (a sample can

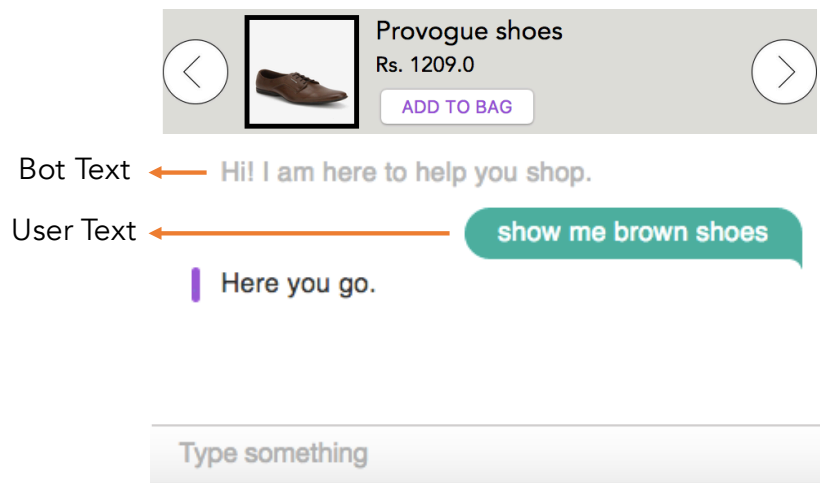


Figure 4.1: Shoe shopping default chatbot.

be seen in Figure 4.2).

#### 4.2.1 *Convey: Showing Context*

Context can be of two types: *inferred* and *assumed*. Inferred contexts are extracted from the conversation between the user and the chatbot. In the example shown in Figure 4.2, the user typed ‘*show me brown shoes*’, so ‘*brown*’ is an inferred context value. Additionally, a chatbot may typically assume a few context values based on the input. For example, on asking for ‘*brown shoes*’, the chatbot might automatically assume that the user is looking for ‘*male*’ shoes (perhaps based on user history). Depending on the chatbot design, even the price range of the shoes can be assumed based on the buying history of the specific user.

*Convey* shows both these contexts differently such that it is clear to the user whether the context was inferred or assumed. In *Convey*, inferred contexts are shown in black, while assumed contexts are shown in gray (Figure 4.2). The context values in *Convey* get updated in real-time as the conversation proceeds. Moreover, all displayed contexts are interactive, as indicated using a dotted underline. When a participant updates an assumed context, either by interacting with *Convey* (e.g., by clicking ‘*men*’ and selecting ‘*women*’ from the drop-down list), or by stating the updated value as part of the conversation (e.g., by texting ‘*looking for*

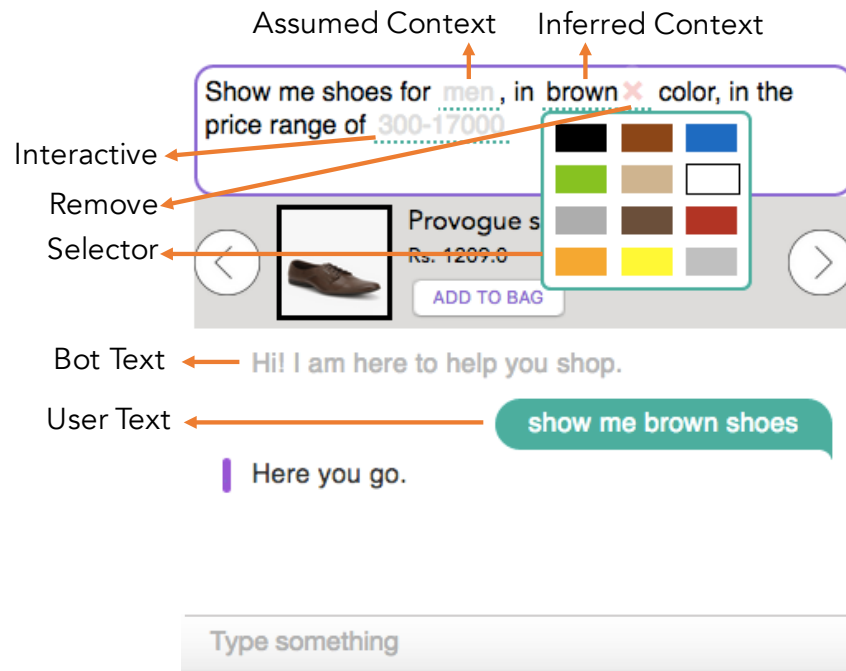


Figure 4.2: Shoe shopping chatbot with *Convey* at the top. Screenshot of the interface used in the user study.

*female shoes*'), *Convey* converts the assumed context into an inferred context. Alternately, based on chatbot design, a user confirmation can be attached to assumed context values, (e.g., the chatbot asking the user explicitly, 'are you looking for male or female shoes?').

#### 4.2.2 *Convey*: Interaction

As stated earlier, the displayed context values in *Convey* are interactive in nature. The user can perform three actions: *confirm context*, *modify context*, and *remove context*. Confirming assumed context is as discussed in the previous section. Another way to confirm assumed context is by long pressing (*i.e.*, holding one's finger/mouse over an item for more than 0.5s) it. Note that this long-press feature is not visible to the user, so it can add to the learning curve. For this reason, the long-press feature was not included in *Convey*'s evaluation.

The user can modify a context by clicking on it. Each context value has a specific UI element associated with it. The element is populated with domain-specific options extracted

from the chatbots’ catalog in the database. Clicking the context value in *Convey* shows the UI with options as a *selector* pop-up. For example, clicking on the ‘*brown*’ color context value in *Convey* pops up a color palette showing colors available in the catalog for male shoes (Figure 4.2). Similarly, clicking on the price range ‘*Rs 300-17000*’ shows a slider-based price selector, while clicking on the gender ‘*male*’ shows a drop-down menu with two gender options (*male* and *female*) to choose from.

Any of the context values can be removed by the user. Apart from users wishing to modify their preference, context may have been wrongly inferred or assumed by the chatbot, which also necessitates deletion by the user. Deleting context has been found to be an issue with current chatbots [127]; either the chatbot does not support deleting context, or it is hard for users to specify the deletion request in text so that the chatbot is able to correctly understand it. In our example, typing ‘*show all colored shoes*’ results in removal of the ‘*brown*’ color context value. *Convey* makes the deletion task much easier by allowing users to click on the cross (‘x’) button next to the context value (as shown in Figure 4.2 next to ‘*brown*’). The *cross* button for deletion, along with the *selector* pop-up for modifying context, only appear after clicking a particular context value. Deleting context is subject to the consistency of *Convey* after their removal.

Finally, the *Convey* design ensures symmetry between the two user modalities: typing and clicking to interact with context. Any interaction with a context value in *Convey* is logged as an equivalent message on the messaging window, which helps the user recognize exactly what happened and also learn additional phrases to message the bot. Both user modalities have equivalent capabilities, and interacting with either of them updates both the *Convey* window and the messaging window. Thus, the two modalities complement each other and can be used interchangeably.

### 4.3 Study Design

In this section, we present the study design by describing the participants, the systems used and the study procedure.

### 4.3.1 Participants

Sixteen participants (11 male and 5 female, mean age = 32.5 years, sd = 7.4 years) were recruited for the study by emailing employees of a local IT company and snowball sampling. Fourteen of them had an engineering background, and the remaining two were from non-technical backgrounds (finance and social sciences). All participants held a Bachelor’s or higher degree. Although none of the participants were native English speakers, all rated themselves fluent in English. Five of them reported using Facebook Messenger every hour of the day, while the rest reported using Messenger at least every four hours daily. All participants understood chatbots at a conceptual level, while five had prior experience interacting with chatbots on the Facebook Messenger platform. Two participants stated that “*proper context understanding*” was one of the major difficulties they faced while interacting with chatbots in the past.

### 4.3.2 System Description

For the user study, we developed a chatbot using IBM Conversation platform [15] with functionalities similar to an e-commerce chatbot for buying footwear. We used the shoes catalog data from jabong.com [5], an e-commerce website. The chatbot was designed to understand and filter shoes based on several features, including price, color, material, style, and brand, to help participants in their decision process. The user can click the shoe image to view a zoomed version of the image. Clicking on ‘*ADD TO BAG*’ (Figures 4.1 & 4.2) results in placing an order for the shoe .

### 4.3.3 Procedure

We conducted a within-subject user study with two interfaces: default chatbot and the same default chatbot with the added *Convey* feature (also referred as *Convey* chatbot). The ordering of the interfaces was randomized across participants to counter ordering effects. With each interface, participants were required to perform one of these two tasks: (a) Select

party footwear for yourself, and (b) Select a pair of sports shoes for the opposite gender. Half of the participants had to select a party footwear for themselves using the default chatbot and select sports shoes for the opposite gender using *Convey* chatbot, while the other half had to select a party footwear for themselves using *Convey* chatbot, followed by selecting sports shoes for the opposite gender using the default chatbot. For the tasks (a) and (b), the combined budget was 3000 INR (45 USD). To motivate the participants, the reward for participation was that a randomly-selected participant would receive his/her selected shoes as a free gift. No other rewards were given for participation.

At the start of interacting with each interface, a one-minute tutorial video (screen-cast with no audio) was played to showcase the capabilities of that interface. At the end, participants were asked to rate their experience on a 5-point Likert scale on several metrics, including ease of use, fun, and frustration [86] and also provide subjective feedback regarding the interface by typing their responses in an online form. After interacting with both the interfaces, participants were asked to compare the two interfaces, and specify which one they preferred and why. Every participant was asked to use their personal laptops/phones for the study with the URL provided by the study facilitator. Participants were not primed to use the *Convey* chatbot in any particular way. All input events were logged and saved on the server for analysis. The study took place in an IT office, and on an average, it took 45 minutes.

#### **4.4 Results**

In general, participants enjoyed their experience interacting with a chatbot for buying shoes, as a majority of them (11) were interacting with a chatbot for the first time: “*It was a fun exercise... got to know how to use chatbots.*” - P<sub>1</sub>, “*... was able to try lots of custom queries*” - P<sub>10</sub>. Also, 9 participants liked that the chatbot was “*very responsive*” and “*prompt*”. Seven participants stated that it was “*easy to use*”, and five appreciated the “*enormous catalog*”. Out of the 16 participants, 7 used their phone for the study, while remaining used their laptop/computer.

<b>Data type</b>	<b>Default Bot</b>	<b><i>Convey</i> Bot</b>
Time taken	7.6±2.1	8.3±2.0
Number of shoes viewed	71.5±21.5	77.6±22.7
Number of shoes zoomed into	9.8±5.3	10.6±9.2
<b>Number of typed messages</b>	9.8±4.4	6.3±3.7
<b>Number of words/message</b>	33.9±13.3	26.1±11.2
<b>Number of chars/message</b>	174.4±65.8	136.6±50.9
Number of <i>Convey</i> interactions		7.8±6.5

Table 4.1: Results from Log Data, mean±std (bold with  $p < 0.05$ )

#### 4.4.1 Log Data

On average, participants viewed  $77.6 \pm 22.7$  shoes with the *Convey* chatbot and zoomed into  $10.6 \pm 9.2$  of them before adding a shoe to the cart, while with the default chatbot, participants viewed  $71.5 \pm 21.5$  shoes and zoomed into  $9.8 \pm 5.3$  of them (Table 4.1). Participants took an average time of  $8.3 \pm 2.0$  mins to complete the task with the *Convey* chatbot, while with the default chatbot, they took  $7.6 \pm 2.1$  mins. This hints that the participants spent enough effort and time in shoe selection.

We conducted paired t-tests between the two interfaces on several parameters, including time taken to complete the task, total number of words input by the user, and total number of shoes browsed and zoomed into. Except for the total number of words input, we did not find any significant difference between the two interfaces. This might be attributed to the fact that the study task was not a performance-measurement task, rather it was a subjective decision-making task. It could also be due to the small sample size. As expected, participants typed significantly more text messages in the default chatbot interface ( $9.8 \pm 4.4$  messages) compared to *Convey* chatbot interface ( $6.3 \pm 3.7$  messages), with  $t_{15} = 1.9$ ,  $p < 0.05$ . Instead of typing, participants interacted with the context values in *Convey*. Moreover, participants also typed longer messages with default chatbot ( $174.4 \pm 65.8$  characters/message) compared to *Convey* chatbot ( $136.6 \pm 50.9$  characters/message) with  $p < 0.05$ , which was mostly attributed to the text messages for updating the price range. Overall, participants interacted with elements in *Convey* 124 times, using a combination of drop-down menus (67 times to

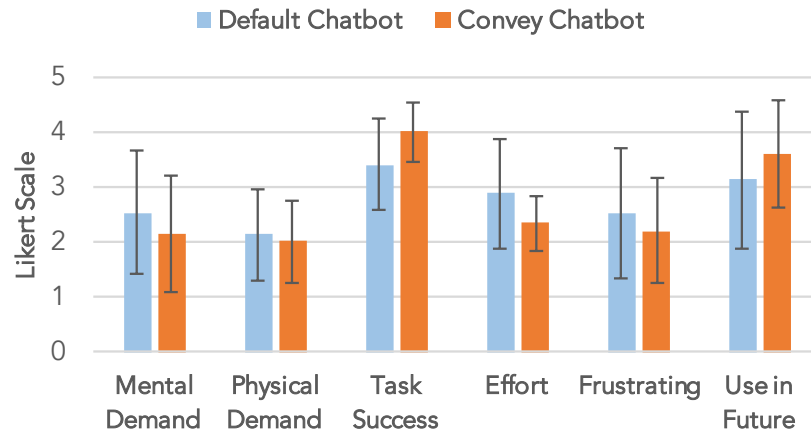


Figure 4.3: Likert-scale rating by the participants (with standard deviation shown as error bars)

select a brand, change gender, modify shoe type, *etc.*), range sliders (26 times to choose the price range), and button menus (21 times to select the shoe color). Also, the remove option on *Convey* was used 10 times.

#### 4.4.2 Ratings

Participants rated both the interfaces on a 5-point Likert scale rating [86] (Figure 4.3). Note that for all metrics other than *Task Success* and *Use in Future*, a lower score is better. We conducted a paired t-test analysis and found the *Convey* chatbot to be significantly better than the default chatbot, with respect to perceived success in performing the task ( $t_{15}=3.0$ ,  $p=0.01$ ), and potentially using it in future ( $t_{15}=3.1$ ,  $p=0.01$ ). The *Convey* chatbot also outperformed the default chatbot in the effort required to achieve the participants' level of performance ( $t_{15}=-2.4$ ,  $p=0.05$ ) and mental demand of the task ( $t_{15}=-2.3$ ,  $p=0.05$ ). The ratings clearly show that participants preferred the *Convey* chatbot over the default chatbot.

#### 4.4.3 Comparison

When asked to choose between the two chatbots for shopping in future, all 16 participants preferred chatbot with *Convey*. These positive comments about *Convey* summarize the participants' response: "*It (Convey chatbot) was more like a shopping experience, the other one was more like an exam!*" - P<sub>13</sub>, "*it felt good interacting with this (Convey) chatbot.*" - P<sub>4</sub>, and "*Well, just keep the GUI at the top, plz, it helps!*" - P<sub>12</sub>.

Seven participants mentioned that the *Convey* chatbot was easier to use ("*easier to find products with different combinations*" - P<sub>7</sub>, "*easier to narrow down products*" - P<sub>5</sub>, "*very intuitive*" - P<sub>15</sub>), and five participants stated that the *Convey* chatbot was faster than the default chatbot. The *Convey* chatbot was perceived to be faster as it "*saves typing effort*" - P<sub>1</sub>, P<sub>5</sub>, and helped in providing precise input, "*I gave less false inputs to the chatbots in case with top interactive part (Convey)*" - P<sub>11</sub>. False inputs have been reported as a major barrier to adoption of conversational systems in general [127].

Interestingly, six participants liked the *Convey* chatbot because it showed context; "*it (Convey chatbot) can keep track of what we are searching currently*" - P<sub>3</sub>, "*easy to see what are the choices made and edit them*" - P<sub>15</sub>, "*maintains effective cumulative history*" - P<sub>6</sub>, "*I was not clear on what filters were getting applied (with default chatbot). I had to go through the old chats to figure that out.*" - P<sub>12</sub>. By showing context, *Convey* reduced confusion; "*It (Convey) showed what the bot understood so there weren't any misunderstandings... It (Convey chatbot) is WYSIWYG of chatbots!*" - P<sub>4</sub>.

Five participants mentioned that *Convey* guided them by "*showing what options are available to choose from.*" - P<sub>8</sub>. For brands and colors in particular, participants were not sure of the available options in the default chatbot interface. Only one participant asked for the options by typing "*show all available brands*", and received a list of brands in response, but this was not obvious to other participants. Two participants pointed at the efficiency of the price range slider, as entering the price range using text was "*almost impossible*" for them; it requires typing "*more than Rs 1000 and less than Rs 2000*". Compared to

the default chatbot, one of the participants commented that he was “*easily able to remove preferences once selected (with Convey)*” - P<sub>7</sub>. Also, three participants mentioned liking the fact that even with *Convey*, they can always use the default text mode, if needed. None of the participants complained that *Convey* took up space at the top of the chatbot, even with the limited screen space of a mobile device.

#### **4.5 Discussion and Design Implications**

In our study, participants preferred the *Convey* chatbot, and found it to be easy to use, intuitive, less mentally demanding, and faster compared to the default chatbot. Interestingly, even though they interacted with the *Convey* interface for a short period of time, participants were cognizant of the benefits offered by *Convey* and appreciated them. Next, we briefly discuss design implications for chatbots, as derived from the positive comments by users while interacting with *Convey*

##### *Summary and Persistent View*

Participants perceived *Convey* as showing a summary of the conversation between the human and chatbot so far, which gets updated after every message turn. The default chatbot interface is non-persistent, *i.e.*, as the conversation proceeds, the text messages are eventually removed from the messaging window. Although user can always scroll up to view past messages, it quickly becomes cumbersome as conversation proceeds. A persistent summary of the conversation not only adds to the usability and but also helps ensure that the humans and the chatbot have the same mental model. This helps counter an important drawback of chatbots, as seen in earlier studies [123, 127, 169], that users lose track of the chatbot’s contextual state.

##### *Form-based UI*

Most existing chatbots do not provide value over alternatives such as search engines, web-pages and native mobile apps [123]. Current chatbots do not allow previous messages to be

edited. In certain scenarios, such as flight booking, changing one of the parameters (*e.g.*, departure date) is easier on a website due to the form-based UI. *Convey*, in a way, combines the benefits of form-based UI with the flexibility of a text-based chat interface. However, unlike a typical form-based UI, a chatbot should not be dense, it should not show all the options available to the user all the time. Instead, showing only those options that pertain to the context explicitly mentioned by the user or assumed by the system, makes it easier to interact.

### *Precise Input*

Participants enjoyed the fact that they could specify precise inputs with *Convey*, especially the price range selector. This is necessary at times, as text might be too cumbersome to type, resulting in lengthy chats to reach the desired outcome, and/or the chatbot might not be intelligent enough to understand complex input text. To elaborate, natural language input to chatbots is highly flexible as anything can be expressed. However, it has a low bit-rate since it requires time for users to type and intelligence for bots to understand. Future chatbots should combine natural language with standard UI elements to enhance the interaction medium between humans and computers in order to combine high flexibility with high bit-rate.

Finally, in this work, we centered our study on a shopping bot. However, *Convey* can be adapted for chatbots in other domains such as IT support, travel booking, news, movie booking, *etc.*, (see Figure 4.4), as the concept of ‘context’ remains consistent across utility-driven conversational systems. The results of the study should also be generalizable to other domains as *Convey* can provide a way for precise input, along with providing a persistent view summarizing the conversation. Adding these capabilities to the current-day chatbots will help in making them more user-friendly and bridging the gap between user experiences and expectations [124, 127, 199].

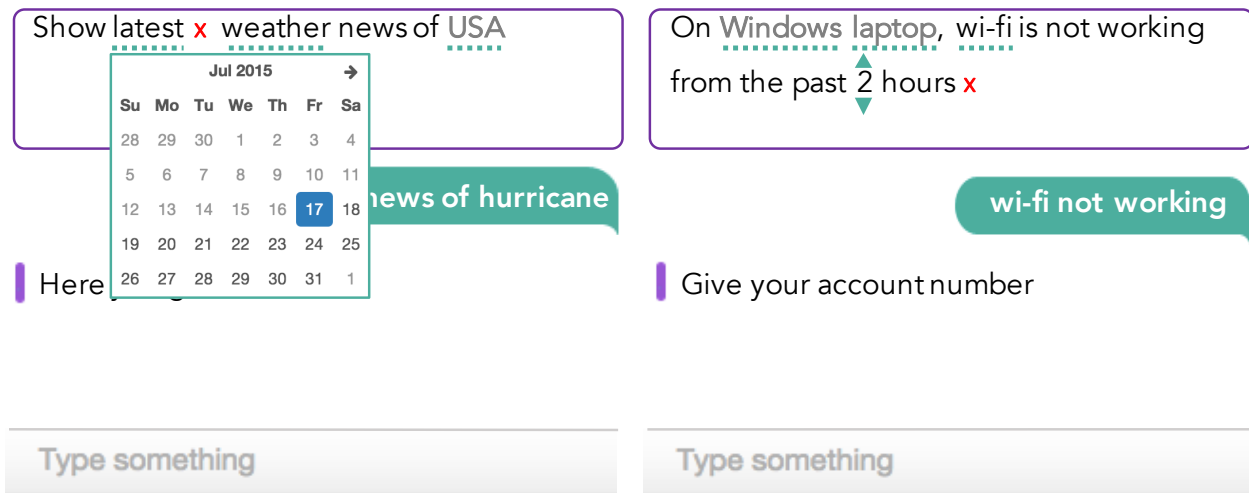


Figure 4.4: Other *Convey* use cases - Left: News; Right: Support

#### 4.5.1 Limitations

The shopping chatbot used for the study had limited capabilities, which participants pointed out. Participants suggested adding more items to the catalog (in particular, more brands), improving the understanding capability (NLP) of the chatbot, enabling viewing of multiple shoes in a carousel, adding images of the same shoe from different angles, enabling an option to maintain a list of shortlisted shoes, auto-correcting spelling mistakes, and providing user reviews and ratings. However, as the purpose of the study was to understand the usability of *Convey*, a chatbot without such advanced features sufficed. In fact, the capabilities of the chatbot used in this study is similar to most of the existing utility-driven chatbots.

## Chapter 5

# RESILIENT CHATBOTS: REPAIR STRATEGY PREFERENCES FOR CONVERSATIONAL BREAKDOWNS

Current natural language understanding technologies are not yet ready to tackle the complexities in conversational interactions. Breakdowns are common in human-chatbot interaction, leading to negative user experiences. Guided by communication theories, in this chapter, we explore user preferences for eight repair strategies that can be broadly applied when designing chatbots for information search tasks. We study strategies that are common in commercially-deployed chatbots (e.g., confirmation, providing options) as well as novel strategies that explain characteristics of the underlying machine learning algorithms. We conducted a scenario-based study to compare repair strategies with Mechanical Turk workers (N=216). We found that providing options and explanations were generally favored, as they manifest initiative from the chatbot and are more actionable to recover from breakdowns. Through detailed analysis of participants' responses, we provide a nuanced understanding on the strengths and weaknesses of each repair strategy.

### ***5.1 Repair Strategies and Research Questions***

We use several concepts from communication theory to drive the choices of repair strategies we study. First, we pay attention to the *initiation of repair* or *evidence of misunderstanding* from the agent. Given users' unfamiliarity with the agent's private model, it is necessary for the agent to indicate a potential misunderstanding. However, an HRI study found that users prefer the agent to ignore the uncertainty and carry on an action until the user initiates a correction [71]. Explicitly acknowledging a mistake lowers the likability and perceived intelligence of the agent, and may add friction to the interaction as the user is obliged to

respond to the initiation.

Second, we distinguish between self-repair and system-repair. Specific to an information search chatbot, users’ self-repair is usually limited to rephrasing the original input. System-repair may diverge from other-repair in human-human conversations given the underlying machine learning model and limited capabilities.

Lastly, we attempt to reduce users’ repair cost by exposing details of system’s understanding status, so users can engage in *assisted self-repair*. We draw inspiration from work on explainable machine learning and introduce three novel designs of agent explanation strategies.

Specifically, we focus on the following eight repair strategies (Figure 5.1). We opted out of a factorial design because the factors we are interested in are either dependent or orthogonal. To initiate system-repair or provide explanation, the agent must acknowledge the potential misunderstanding; however, engaging in system-repair precludes assisting in users’ self-repair. We focus on strategies that can be broadly applied to information-search chatbots, based on the assumption that an intent-based model is used [216]: A chatbot’s understanding capability relies on using a multi-classifier to classify a user utterance to one of its pre-defined intents, which would trigger a chatbot response for that intent. Specifically, classification of each intent has a *confidence*, and the one with the highest confidence is considered as the recognized intent. If all confidence levels are below a threshold, the chatbot recognizes that there is a potential breakdown. Our repair strategies are concerned with the immediate action that a chatbot would take after recognizing such a breakdown.

### 5.1.1 *No evidence of a breakdown*

**Top response:** Similar to the “ignore” strategy studied by Engelhardt et al. [71], the chatbot gives no evidence of a potential breakdown, but outputs the response to the intent with the highest confidence, even when it is below the threshold. In this scenario, the user would have to initiate the repair after seeing the wrong response.

### 5.1.2 *With evidence of a breakdown*

**Repeat:** The chatbot recognizes a potential breakdown and explicitly indicates it, and repeats the prompt requesting the user to ask a question.

**Confirmation:** The chatbot recognizes a potential breakdown when the top intent is under the confidence threshold. It then explicitly confirms the top intent (e.g., “sounds like you want to... is that correct?”). This strategy is considered more natural, and similar to how a human listener initiates a repair [176].

### 5.1.3 *With evidence of a breakdown, system-repair*

**Options:** The chatbot not only indicates a potential breakdown but also provides options of potential intents in which it has the highest confidence. The system attempts to repair by taking over the dialogue initiative to restrict interaction within its capabilities.

**Defer:** It is a common strategy for a chatbot to transfer a request it is unable to solve to a human agent. We consider deferring as a type of system-repair as it is a solution for the system to resolve breakdowns via human intervention.

### 5.1.4 *With evidence of a breakdown, assisted self-repair*

**Keyword highlight explanation:** Inspired by keyword-based explanations for text classifiers [195], we introduce a strategy that reveals why an intent is mistakenly recognized by highlighting keywords in the user’s utterance that contribute to the classifier’s decision. By exposing the chatbot’s understanding mechanism, it is expected to help the user rephrase by avoiding the keywords that the chatbot misunderstands or using words that are closer to the desired intent.

**Keyword confirmation explanation:** This strategy is similar to keyword highlighting, but instead of highlighting on the original user utterance, the chatbot explicitly explains its understanding to the user in a confirmation message. While it is more natural in a conversational form, it makes a trade-off in that it needs an additional conversational turn.

**Out-of-vocabulary explanation:** This strategy highlights words that the bot cannot understand to help the user rephrase. This explanation can be realized by extracting words that are distant or missing from the chatbot’s training data or knowledge base.

### 5.1.5 Research Questions

By conducting a scenario-based study to compare these strategies in pairs, we attempt to answer the following research questions:

- **RQ1:** Which repair strategies are preferred when a conversational breakdown with a chatbot occurs, and why?
  - **RQ1a:** Is it preferable to acknowledge breakdowns?
  - **RQ1b:** Is it preferable to provide system-repair?
  - **RQ1c:** Is it preferable to provide assisted self-repair by explaining system’s understanding?
- **RQ2:** How do different individual and task-related factors impact preferences for different repair strategies?

For **RQ2**, there are a number of individual factors worth considering, including social orientation with chatbots (i.e., tendency to engage in human-like social interactions [123, 124]), service orientation (i.e., viewing service interactions as either transactions or social interactions [120]), experience with chatbots, and experience with technology. For task-related factors, we focus on the repair outcome (successful or not) and the context (e.g. shopping vs. banking).

## 5.2 Methodology

To answer our research questions, we developed several scenarios in which a chatbot interacted with a human and a breakdown occurred. These scenarios were created for three

different task contexts in which chatbots are commonly used: shopping, banking, and travel. All scenarios contain the same breakdown, in which the user presents an utterance that the chatbot is unable to confidently understand (e.g., its confidence is below the intent threshold), shown in the “Initial Prompt” in Figure 5.1. The chatbot then attempts to repair the conversation using one of the eight repair strategies, shown in Figure 5.1. We considered cases in which the repair was either successful or not. Thus, we developed a total of 48 individual scenarios: 3 (context)  $\times$  8 (repair)  $\times$  2 (outcome success).

### 5.2.1 Paired Comparison Experiment

We adopted a pairwise comparison experiment to collect user preferences for repair strategies. Specifically, our experiment consists of tasks where we randomly showed participants two of the eight repairs, but with the same context (shopping/banking/travel) and outcome (successful/unsuccessful). We asked participants to select the scenario that appealed to them more and describe why they had made their selection.

Pairwise experiments are commonly used in various fields of research to determine participant judgments [53, 117]. Pairwise comparisons could yield more realistic results than Likert scales [21] because they take advantage of simple judgments and prioritize a small set of stimuli to learn user’s preferences [50, 63]. They also allow us to elicit qualitative responses on the desirable traits of one repair strategy over another. We perform rank analysis of our pairwise comparisons using the Bradley-Terry model [39].

### 5.2.2 Individual Factors Survey

We are interested in how the following individual factors impact user preference for repair strategies: social orientation toward chatbots, service orientation, prior experience with chatbots, and experience with technology in general. These factors are introduced in prior work on chatbots and were shown to impact user preference and user behaviors. All measures are self-reported using 5-point Likert scales.

<p style="text-align: center;"><b>Initial Prompt</b></p> <p><b>Bot:</b> Hello, I'm your digital assistant. How can I help you with your banking needs today?</p> <p><b>User:</b> hi I'd like to add my daughter to my card</p> <p style="text-align: center;"> <input type="button" value="Show First Chat"/> <input type="button" value="Show Second Chat"/> </p>	<p><b>Bot:</b> I don't quite understand. How can I help you with your banking needs today?</p> <p><b>User:</b> I want to add my daughter sarah as an authorized user to my credit card</p> <p><b>Bot:</b> OK, I am now adding your daughter listed on file as Sarah Johnson to your credit card ending in 8741.</p> <p><b>User:</b> perfect thanks!</p>
<p><b>Bot:</b> I can help you with adding a credit card. First tell me the last 4 digits of the card you want to add?</p> <p><b>User:</b> no i want to add my daughter sarah as an authorized user to my credit card</p> <p><b>Bot:</b> OK, I am now adding your daughter listed on file as Sarah Johnson to your credit card ending in 8741.</p> <p><b>User:</b> perfect thanks!</p>	<p><b>Repeat</b></p> <p><b>User:</b> hi i'd like to <b>add</b> my daughter to my <b>card</b></p> <p><b>Bot:</b> I can help you with adding a new credit card. First tell me the last four digits of the card you want to add? <i>(I've highlighted keywords I understand in your request. If this is not what you meant, please rephrase to help me understand better.)</i></p> <p><b>User:</b> i want to add my daughter <b>sarah</b> as an <b>authorized user</b> to my credit card</p> <p><b>Bot:</b> OK, I am now adding your daughter listed on file as Sarah Johnson to your credit card ending in <b>8741</b>. <i>(I've highlighted keywords I understand in your request. If this is not what you meant, please rephrase to help me understand better.)</i></p>
<p><b>Top</b></p> <p><b>Bot:</b> Sounds like you want me to add a credit card to your account. Is that right?</p> <p><b>User:</b> no.. i want to add my daughter sarah as an authorized user to my credit card</p> <p><b>Bot:</b> Sounds like you want to add an authorized user to your credit card ending in 8741. Is this correct?</p> <p><b>User:</b> yes</p> <p><b>Bot:</b> OK, I am now adding your daughter listed on file as Sarah Johnson to your credit card ending in 8741.</p> <p><b>User:</b> perfect thanks!</p>	<p><b>Keyword Highlight Explanation</b></p> <p><b>Bot:</b> Hello. My understanding is that you want to <b>add a card</b> to your account. Is that correct?</p> <p><b>User:</b> no.. i want to add my daughter sarah as an authorized user to my credit card</p> <p><b>Bot:</b> OK. My understanding is that you want to <b>add Sarah</b> as an <b>authorized user</b>. Is that correct?</p> <p><b>User:</b> yes</p>
<p><b>Confirmation</b></p> <p><b>Bot:</b> I am not quite sure if I understood. Which of these are you trying to do today: 1) Add my daughter as an authorized user to my credit card. 2) Add a credit card to my account, or 3) Apply for a credit card for my daughter, 4) None of the above?</p> <p><b>User:</b> option 1 add my daughter</p> <p><b>Bot:</b> OK, I am now adding your daughter listed on file as Sarah Johnson to your credit card ending in 8741.</p> <p><b>User:</b> perfect thanks!</p>	<p><b>Keyword Confirmation Explanation</b></p> <p><b>User:</b> hi i'd like to add my <b>daughter</b> to my card</p> <p><b>Bot:</b> Sorry I don't understand. I've highlighted words I am unable to process. Can you please rephrase?</p> <p><b>User:</b> i want to add my child sarah johnson as an authorized user to my credit card</p> <p><b>Bot:</b> Sounds like you want to add Sarah Johnson listed on your family profile to your credit card ending in 8741. I can do that now.</p>
<p><b>Options</b></p> <p><b>Bot:</b> Sorry, I do not understand. For this request, I will need to refer you to a human agent who will address your request in the next few minutes.</p> <p style="text-align: center;"><i>A human agent comes online, reads the message and tells you he will add your daughter, Sarah, to your credit card ending in 8741.</i></p> <p><b>Defer</b></p>	<p><b>Out-of-Vocabulary Explanation</b></p> <p><b>User:</b> perfect thanks!</p>

Figure 5.1: Eight repairs for the successful banking condition. Top left image initial prompt a user is given. From top to bottom: Top, Repeat, Confirmation, Options, Defer, Keyword Highlight Explanation, Keyword Confirmation Explanation, Out-of-Vocabulary Explanation.

*Social Orientation toward Chatbots:* Introduced by Liao et al. [123, 124], this measure reflects individual tendency to engage in human-like social interactions, which is associated with a difference in mental models of agent systems as sociable entities versus machines. They found that users with high social orientation desire natural conversation and social designs from the agent while those low in the orientation use a chatbot like an information search engine. We used the scale introduced in [123]: “*I like chatting casually with a chatbot*” and “*I think ‘small talk’ with a chatbot is enjoyable.*” Cronbach’s  $\alpha$  was 0.84 indicating high reliability.

*Service Orientation:* In Lee et al.’s work studying recovery strategies for robot breakdown [120], they noted a preference difference in those with a utilitarian vs. a relational service orientation. We adapted two items from their work: “*Efficient customer service is important to me*” and “*I found it frustrating when a customer service representative could not immediately give me the information I need.*” However, Cronbach’s  $\alpha$  was 0.38 indicating poor reliability, so we include these items as two separate measures in our analysis: service frustration and service efficiency.

*Experience with Chatbots:* We assessed self-reported prior experience with chatbots: “*I am familiar with chatbot technologies*” and “*I use chatbots frequently.*” Cronbach’s  $\alpha$  was 0.71 indicating good reliability.

*Experience with Technology:* We assessed self-reported tech-savviness: “*I consider myself an advanced technology user*” and “*I am eager to try new technologies.*” Cronbach’s  $\alpha$  was 0.70 indicating good reliability.

### 5.2.3 Participants, Task, and Procedure

Participants were recruited on Mechanical Turk with the requirement of being 18 years or older. In each task, participants performed 10 pairwise comparisons between repair strategies for a given scenario and outcome. Each scenario was presented turn-by-turn, with three-second typing indication pauses in between chat bubbles to simulate the interactive experience of a chat. After reading the first scenario (shown on the left half of the screen),

participants clicked a button to show the second scenario (shown on the right half of the screen). After both scenarios were presented, participants were asked to select which chatbot they preferred and give some explanation as to why. Scenarios were selected randomly without replacement so the same participant did not see the same combination of factors twice, and two control scenarios were included as attention checks. The first repeated a previous scenario to see whether the participant gave it the same rating. The second provided a comparison between a chatbot that had successfully repaired the breakdown with one that did not, and participants were expected to express a preference for the one that was able to successfully repair. After finishing all 10 comparisons, participants filled out a survey that collected demographic information and measurements of individual factors as introduced in the section above. The overall task took about 10 minutes to complete, and participants were compensated \$1.50 USD for their participation (\$9 USD/hr).

We deployed a total of 340 tasks on Mechanical Turk. We filtered out 124 participants (36%) who did not pass the attention checks, yielding a final sample of 216 participants (141 male) and 1,728 pairwise comparisons. Of these, 124 (57%) held a bachelor’s degree, and 28 (13%) held a post-graduate degree. The average age of our participants was 34 years (SD=9 years). Most of our participants spoke English as their native language (N=189, 88%), and other native languages included Hindi (4%), Malay (3%), and Tamil (3%).

### **5.3 Results**

In this section, we describe user preferences for repair strategies and the underlying reasons (RQ1), where we pay attention to preferences with respect to the acknowledgement of breakdowns (RQ1a), system-repair (RQ1b) and assisted self-repair (RQ1c). We then explore how individual and task-related factors impact these preferences (RQ2).

#### *5.3.1 Preferences of Repair Strategies (RQ1)*

The Bradley-Terry model [39] is a mathematical model that estimates a vector of “ability scores” for a set of paired object comparisons, which yields an ultimate ranking of all ob-

Preferred repair vs. Rejected repair	p-value
Options vs. Keyword Highlight	0.000**
Options vs. Confirmation	0.000**
Options vs. Repeat	0.000**
Options vs. Top	0.000**
Options vs. Defer	0.000**
Options vs. Keyword Confirmation	0.000**
Options vs. Out-of-Vocabulary	0.002**
Out-of-Vocabulary vs. Confirmation	0.000**
Out-of-Vocabulary vs. Top	0.000**
Out-of-Vocabulary vs. Repeat	0.000**
Out-of-Vocabulary vs. Keyword Highlight	0.000**
Out-of-Vocabulary vs. Defer	0.000**
Out-of-Vocabulary vs. Keyword Confirmation	0.041
Keyword Highlight vs. Top	0.036
Keyword Highlight vs. Confirmation	0.058
Keyword Highlight vs. Keyword Confirmation	0.576
Keyword Highlight vs. Repeat	0.270
Keyword Confirmation vs. Confirmation	0.014
Keyword Confirmation vs. Top	0.008*
Keyword Confirmation vs. Defer	0.061
Repeat vs. Defer	0.855
Repeat vs. Keyword Confirmation	0.094
Defer vs. Keyword Highlight	0.199
Confirmation vs. Defer	0.546
Confirmation vs. Repeat	0.433
Top vs. Defer	0.413
Top vs. Repeat	0.315
Top vs. Confirmation	0.828

Table 5.1: Significant values, after Bonferroni adjustment ( $p < 0.05/8$ ), are noted with \*\*. Marginally significant values ( $p < 0.1/8$ ) are noted with \*.

jects. This model has been used in previous HCI studies conducting pairwise comparison experiments [23, 182]. We use the BradleyTerry2 R package [204] to generate an overall ranking of repair strategies followed by pairwise comparison tests for significance. For each repair, the model conducts a pairwise test that generates a p-value for each other repair it is compared to. We used a Bonferroni correction [210] to account for the number of individual comparisons made ( $p < 0.05/8$  for significance,  $p < 0.1/8$  for marginal significance [60]). In Figure 5.2 we show the overall rankings as well as separate rankings for when the scenario was successfully or unsuccessfully repaired. In Table 5.1 we present the p-values for pairwise comparisons.

<b>Repair</b>	<b>Strengths</b>	<b>Weaknesses</b>
<b>Top</b>	concise with no extraneous questions; simple interaction	began an unwanted process without confirming; lacks resource to resolve breakdown; unfriendly and rude
<b>Repeat</b>	concise; natural; explicit about lack of understanding	appears less intelligent; did not show interest of understanding; lack of resources for user to repair
<b>Confirmation</b>	verify before take an action; show understanding capability; polite; natural	longer conversation to respond to confirmation; appear less competent by repetitively confirming
<b>Options</b>	provide choices to resolve the issue faster; narrow down to what it can do; show understanding capability and intelligence; less typing required by user	complicates with clutter; unnatural; more reading
<b>Defer</b>	interaction with human is faster; human more likely to solve the problem; prefer interacting with a human	wait time and interaction with human slower; human intervention is unnecessary
<b>Keyword Highlight Explanation</b>	show understanding capabilities; help users to rephrase; teach user how to interact with the chatbot; proactively making an effort; intuitive explanation; resolve issue faster with less turns	verbose; repetitive description; highlighting is visually unappealing; less natural
<b>Keyword Confirmation Explanation</b>	show understanding capabilities; help users to rephrase; teach user how to interact with chatbots; proactively making an effort; polite; concise	highlighting is visually unappealing; longer conversation to respond to confirmation; less information provided
<b>Out-of-Vocabulary Explanation</b>	show understanding capabilities; help users to rephrase; teach user how to interact with chatbots ; proactively making an effort; polite; concise; specific about why it fails to understand	appear less competent unable to understand simple words

Table 5.2: Strength and weaknesses of repairs as reported by participants.

As seen in Figure 5.2, the Options repair was unarguably the most favored strategy, preferred in pairwise comparisons over all other strategies (Table 5.1). Assisted self-repairs – Keyword Highlight, Keyword Confirmation, and Out-of-Vocabulary Explanation – were generally favored, with Out-of-Vocabulary Explanation as the most preferred among the three. For the rest – Defer, Confirmation, Repeat, and Top – preferences were noisier. Part of the reason, as we observe in Figure 5.2, is that they were ranked differently in scenarios with successful and unsuccessful repair outcomes. Most evidently, Defer was outranked by all other repairs when the repair was successful, but ranked second when the repair was unsuccessful. This difference implies that if a breakdown can be easily repaired, users prefer to resolve it with the chatbot, whereas if the repair fails after an initial attempt, they desire a human agent to be involved, even if the human agent is unable to resolve it immediately

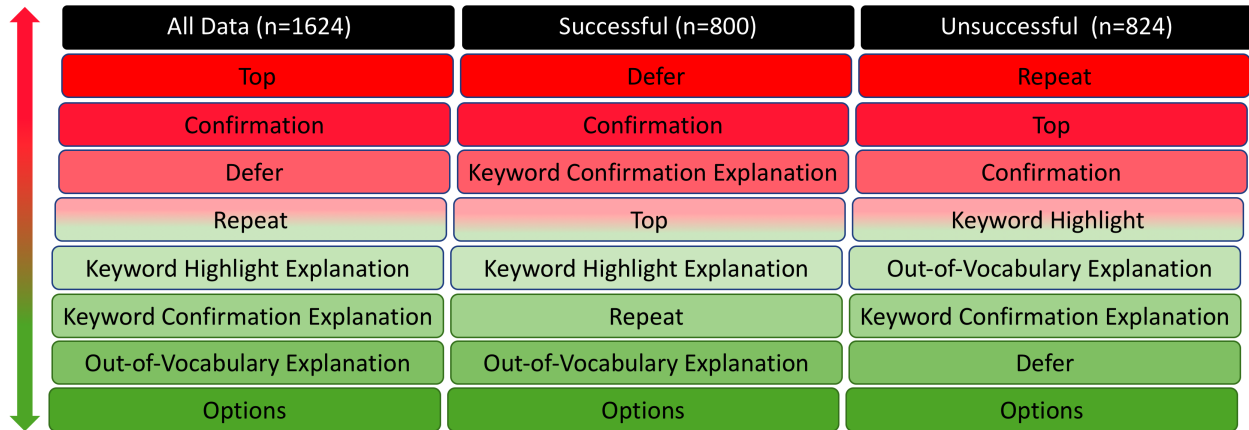


Figure 5.2: Bradley-Terry rankings of repair strategies. From left to right, rankings for: all data, successful conditions, unsuccessful conditions. From top to bottom: lowest ranked to highest ranked.

(as in the scenario). We also observe that simple strategies – Top and Repeat – were ranked higher in successful than unsuccessful scenarios. This finding suggests that if the breakdown is straightforward enough to repair with one attempt, chatbots that don’t offer evidence of breakdown or repair assistance are acceptable.

### 5.3.2 Reasons for Preferences (RQ1)

Along with collecting user preferences, we asked participants to give reasons why they selected a repair strategy over another. The authors individually reviewed this data and used open coding [70] to extract themes in the open-ended answers. Codes were harmonized after two iterations of review and discussion, resulting in the final set of themes shown in Table 5.2. A few common themes were observed across repair strategies, reflecting general user desires for repair design: 1) *efficiency and efficacy* were desired when recovering from the breakdown to accomplish the information search goal, as demonstrated by codes such as “faster,” “concise” (easy to read), “help to rephrase,” and “less typing required”; 2) some strategies increased *perceived intelligence and capability*, especially when the agent demonstrated its understanding through confirmation or explanations, or when it proactively assisted repair

via explanation or directly providing options; 3) *politeness* was demonstrated in strategies that presented an understanding before executing a response (e.g. confirmation, explanations); and 4) *naturalness*, in which users feel that interactions faithfully resemble human conversations, was not felt in strategies that highlighted keywords or provided options. Based on the results shown in Table 5.2, we focus on addressing our research questions regarding breakdown evidence, system-repair, and assisted self-repair.

### *Explicit Acknowledgement of Breakdown (RQ1a)*

Our ranking results suggest that participants preferred chatbots to explicitly acknowledge a potential breakdown, as Top was generally less favored. Our qualitative data reveals that to proceed with a wrong response is not only unhelpful for resolving the breakdown, but is also perceived as rude, unfriendly, and putting in no effort. Although for scenarios in which the breakdown was resolved in one attempt, participants were more tolerant of the Top strategy, and some also favored its simplicity.

Two similar strategies that acknowledge potential breakdowns – Repeat and Confirmation – had interesting trade-offs. Participants perceived Confirmation to be more polite (verifying before taking an action) and intelligent (showing its understanding capability) than Repeat, but some found it more burdensome to have to read and respond to the confirmation. Both strategies were considered natural as they resemble ways that a human listener would initiate a repair.

Our results also suggest that while participants like chatbots to acknowledge potential breakdowns, they may be turned away by messages that are redundant and repetitive, as what appeared in Keyword Highlighting.

### *System-Repair (RQ1b)*

We introduced two distinct strategies for system-repair: Options and Defer. Participants favored Options because of it was efficient and required less effort from the user in formulating and typing. They also perceived the chatbot to be more intelligent by taking the

dialogue initiative. We note that the scenario-based method does not completely capture the actual success rate in finding the desired intent. However, we found that this strategy was favored even in the unsuccessful scenarios, and one participant commented that it “*ends the conversation quicker when it doesn’t understand instead of stringing me along.*” (P76, *Options vs. Top*). Users also liked to have the “none of the above” option to explicitly exit a conversation: “*It at least did provide a way to say that it was on the wrong track: i.e. none of the above*” (P117, *Options vs. Out-of-Vocabulary Explanation*).

As discussed earlier, the status of a breakdown (successful/unsuccessful) affects users’ preferences. When the repair fails, Defer is a preferred strategy as a human agent is more likely to resolve a difficult issue. In contrast, if success can be achieved through a single repair, participants generally found the intervention of a human agent and the waiting time to be unnecessary. “*I liked the fact that the bot continued to try to work out what was being asked rather than immediately referring the user to a human agent, which defeats the purpose of the bot.*” (P77, *Keyword Confirmation vs. Defer*).

### *Assisted Self-Repair (RQ1c)*

Repair strategies that aid with self-repair, by exposing the chatbot’s understanding model, were generally ranked highly compared to strategies that provide no evidence of misunderstanding (Top) or simple acknowledgement (Confirmation). The qualitative results reveal several themes shared by these strategies. First they provide actionable resources for the user to resolve the breakdown, by either avoiding undesirable words or using words more specific to the targeted intent when rephrasing: “*I really like seeing the keywords highlighted since it gives me insight into the logic behind the bot’s responses, which will assist me if it does not provide the response I want.*” (P108, *Keyword Highlighting vs. Repeat*).

Second, they make the chatbot to appear more intelligent, not only by exhibiting its understanding capabilities, but also showing pro-activeness to help repair—“*bot is interactive and appears to have interest in understanding question by asking questions to clarify.*” (P42, *Keyword Confirmation vs. Repeat*);

Lastly, some participants noted an educational aspect as well, in that the explanations helped them better understand how the chatbot worked by *“teach[ing] you how to speak to the bot”* (P151, Out-of-Vocabulary vs. Confirmation). Conversely, the explanation-based strategies were considered less natural as they did not resemble human conversations due to their use of GUI elements (e.g. highlighting words) that some participants found to be visually unappealing.

By directly highlighting keywords in the user’s original utterance, Keyword Highlight Explanation was considered more intuitive in explaining how the underlying algorithm worked. However, the particular design decision of including a repetitive and verbose prompt that described the highlighting – *“I’ve highlighted keywords in your response...”* – was disfavored. Future work should consider removing description after the first few rounds of interaction. In comparison, Keyword Confirmation was more concise and appeared to be polite by verifying first, but it has the drawback of adding additional turns and user effort in order to respond to the confirmation. Some participants noted that in such interactions, the chatbot was only explaining its understanding without providing additional information, while in the Keyword Highlighting strategy the response given may add to opportunistic success. While Out-of-Vocabulary Explanation was perceived to be more explicit about its misunderstanding to help the user rephrase, others felt it appeared less intelligent if it could not understand common words.

### 5.3.3 *Impact of Individual and Task Differences (RQ2)*

In this section, we explore how the individual factors of social orientation toward chatbots, service frustration and efficiency, experience with chatbots, and experience with technology, as well as task variables of repair outcome (success/failure) and context (shopping/banking/travel), impacted preferences for the eight repair strategies. We rely on a statistical modeling approach. For each repair strategy, we selected all paired comparisons in which it appeared ( $N \in [356, 389]$ ), then built a logistic regression model predicting whether it would be the winner or not by including the individual and task factors as inde-

pendent variables. Thus, we ran eight logistic regression models. We report the task-related factors and focus on results that are statistically significant.

### *Social Orientation toward Chatbots*

Social orientation reflects individual differences in the tendency to engage in human-like social interactions with chatbots. It is associated with a difference in mental model, of seeing agents as sociable entities rather than machines [119, 123, 124]. We found that users with higher social orientation are significantly more likely to favor Top strategy ( $\beta = 0.39$ ,  $SE = 0.11$ ,  $p < 0.001$ ) and marginally less so for Keyword Confirmation Explanation ( $\beta = -0.18$ ,  $SE = 0.10$ ,  $p = 0.07$ ) and Options ( $\beta = -0.22$ ,  $SE = 0.13$ ,  $p = 0.08$ ). These results are consistent with the notion that people with high social orientation prefer natural conversational interaction, and may have felt the use of options and keywords to be mechanical. While we identified *naturalness* to be a desirable characteristic of repair strategies, it is likely that to be more preferred by those with a high degree of social orientation toward chatbots.

### *Service Frustration and Efficiency*

Lee et al. found that those with a utilitarian orientation preferred robot repair that provided instrumental value instead of emotional comfort [120]. We found participants with higher service frustration are marginally less likely to favor Keyword Confirmation Explanation ( $\beta = -0.20$ ,  $SE = 0.11$ ,  $p = 0.06$ ), but more so for Keyword Highlight Explanation ( $\beta = 0.19$ ,  $SE = 0.11$ ,  $p = 0.10$ ). The difference between these explanation-based strategies is that the latter outputs a response directly and the former takes an additional turn to explain the understanding. Participants who are less patient with service interactions preferred a strategy that resulted in fewer turns, even while it may have appeared more mechanical and less polite.

### *Experience with Chatbots and Technology*

Participants with more prior experience with chatbots are more likely to favor Confirmation ( $\beta = 0.32$ ,  $SE = 0.15$ ,  $p = 0.03$ ), which intuitively makes sense as confirmations are commonly used in existing chatbot services. Participants with a greater level of technological experience are marginally more likely to favor Out-of-Vocabulary Explanation ( $\beta = 0.29$ ,  $SE = 0.16$ ,  $p = 0.07$ ), indicating that designs that expose details of the underlying algorithms may appeal to more tech-savvy users.

### *Repair Outcome*

When repairs were successful compared to unsuccessful, participants are more likely to favor Top ( $\beta = 0.45$ ,  $SE = 0.22$ ,  $p = 0.04$ ) and Repeat ( $\beta = 1.38$ ,  $SE = 0.22$ ,  $p < 0.001$ ), and less likely to favor Defer ( $\beta = -1.37$ ,  $SE = 0.22$ ,  $p < 0.001$ ) and Keyword Confirmation Explanation ( $\beta = -0.45$ ,  $SE = 0.21$ ,  $p = 0.03$ ). We conclude that simple strategies (Top, Repeat) are more acceptable if a repair can be achieved easily, while more complex repair strategies (Keyword Confirmation) or strategies requiring human intervention (Defer) may be more desirable in more difficult repair situations.

### *Task Context*

We did not find any statistical differences across task context, suggesting that our findings on repair strategy preferences may generalize across different domains.

## **5.4 Discussion**

We first summarize design recommendations for repair strategies of chatbots. We then revisit the theoretical framework and discuss how our results contribute to understanding of grounding in the context of human-agent conversation.

### 5.4.1 *Design recommendations*

#### *Acknowledging Misunderstanding with Forthrightness and Less Redundancy*

Our participants preferred repairs that explicitly acknowledge a breakdown, but complained that the repetitive acknowledgement to be “clutter” and “redundant.” We recommend having alternative messages for acknowledging misunderstandings, while carefully setting the uncertainty threshold so that these acknowledgements do not appear overly frequently. For example, for individuals more tolerant of Top strategies, this threshold can be raised.

#### *Explaining Models Naturally, Aesthetically, and Effortlessly*

We show that explaining the mechanisms of the underlying models is considered helpful for repair, making the chatbot appear intelligent and teaching users better ways of interaction. While UI elements such as highlighting can be a powerful tool, one should carefully consider how to embed them in conversations so that they do not appear to be “mechanical,” “unnatural,” “visually unappealing,” “hard to read” or “confusing” (some participants confused highlighted keywords with hyperlinks). Meanwhile, utilizing algorithmic inference and rich UI elements are ways to reduce user effort. We found that the Keyword Highlight Explanation was perceived as efficient by highlighting on the users’ original utterance, saving a conversational turn. More advanced designs, such as suggesting words to use, may further reduce user effort.

#### *Intelligently Repair with User Control*

We show that repair works best when an agent can proactively suggest the correct action. In reality, to achieve such a level of intelligence requires significant effort in implementation, and even so it may fail at times. In the survey responses, some participants noted that the “None of the Above” option provides an explicit “way out” or “reset button.” One of the canonical golden rules of user interface design is to provide a user with the control to permit a reversal of actions [188]. It is even more important in intelligent systems to always allow

user oversight on system agency. Besides a way to exit, a user may also desire to control the triggering condition of a system repair, even to fine-tune the options (e.g., remove an unlikely option for future interactions).

#### *Adapting to Individuals and Contexts*

We observed that preferences for repair strategies are not universal. While it is useful to identify individual and task-related factors that impact preferences, one may also leverage the interactivity of an agent system to adapt to individuals and contexts through data or feedback-driven approaches.

#### *5.4.2 Repair as a Collaborative Action with Costs*

To guide the design choices of the repairs we studied, we used grounding in communication as a theoretical framework [56], which views conversations a collaborative action. Our results show that participants increasingly preferred strategies where the system provides increasing level of contribution to the repair process. Specifically, we considered three levels of contribution: 1) evidencing a breakdown; 2) providing resources to assist user-repair; 3) actively taking the initiative to repair.

In line with earlier work that built adaptive dialog systems based on grounding activities [41, 150, 151, 203], our empirical results support the point of view that grounding theory is a robust framework that can be applied from human-human to human-agent conversations. Core concepts such as *collective contribution*, *evidence of understanding*, *cost of repair*, are important to consider in designing repair capabilities of agents. However, the types of cost and their weights may change in the new context of agent conversations, resulting potentially different phenomena in choices of repair strategies. For example, we found that system (other)-repair was preferred over self-repair in our results, contradicting with observations from human-human conversations [56, 176]. One reason could be that fault cost (being perceived at fault), which one would try to minimize when talking to another person, is no longer an issue when interacting with an agent. Moreover, the design we presented,

requiring a participant to only click an option, largely reduced formulation (rephrasing) and production (typing) cost compared to all the other repair strategies.

There is a caveat to our study, in that it did not capture all dimensions of cost in actual interactions. While we tried to control for the repair outcome in all conditions, a less capable chatbot may have a low chance of suggesting relevant intents, so a user may spend more effort having to re-try from the beginning for multiple times, than directly engaging in self-repair. This problem is relevant to “start-up cost” and additional “turn-taking cost” that are considered in the original grounding framework, but not captured in our study design.

Cost can also be used to interpret the impact of individual and contextual factors, by considering how they vary the weights of different costs. For example, an individual with high social orientation may consider “loss of naturalness” as an undesired cost, but those low on the orientation may assign little weight to such a cost. This explains why the former group was more likely to appreciate simple, natural repair strategies than the latter.

The notion of different costs can also direct us to consider new designs of repairs. For example, a simple improvement to explanation-based strategies is to allow users to easily retrieve and edit previous utterances, reducing their production (typing) cost. Based on the idea of reducing turn-taking cost, another improved design is “type-ahead repair,” by suggesting a potential breakdown and explanation *before* the user sends out the message.

By considering the dimensions of costs and benefits as the underlying mechanism, and how a specific design embodies them, one may start having a theory-guided framework to understand and predict user preferences for various designs of repair and broader conversational capabilities. While grounding theory enumerates a comprehensive list of costs regarding human communications, our work calls for further empirical investigation to establish a theoretical framework of grounding for human-agent communications.

### 5.4.3 *Limitations*

The results of this study are promising in delineating the best repair strategies for human-agent repairs. However, we acknowledge some limitations. First, for a lack of statistical

significance, we could not make strong conclusions for how some of the lesser-ranked repairs fare against each other (Top, Repeat, Confirmation, Defer) given their larger p-values. However, by answering research questions guided by the theoretical framework, we believe that we paint an accurate high-level picture of preferred repairs in human-agent breakdowns. Second, limited by using a scenario-based experimental study, our work could not account for how user preferences for repair strategies are affected by nuances in system performance, such as confidence level and performance of the explanation methods. Future work should explore these questions with a real chatbot system. The study was limited by the fact that we only tested scenarios with a one-turn request-response task. Future studies can benefit from evaluating different kinds of user tasks, such as multi-turn conversations. Our study is also limited by our sample of Mechanical Turk workers. Due to the linguistic nature of our task, we desired to have fluent English speakers participate. However, our final sample was biased toward college educated males. Future work is needed to understand how repair strategy preferences differ across languages and cultures, which may have different expectations or norms for how humans ought to interact with conversational agents.

## Chapter 6

### **FARMCHAT: A CONVERSATIONAL AGENT TO ANSWER FARMER QUERIES**

While recently the HCI community has developed an increasing interest in conversational agents and demonstrated their benefits [93, 95, 123, 124, 127, 183], most prior works focus on literate, technologically-advanced users. With a low adoption rate of only 16% for current chatbots, to make chatbots universal, it needs to be accessible to new demographics. In this chapter, we report our experience of designing a conversational agent, called FarmChat, to meet the information needs of low literate farmers in rural India. Farmers constitute 54.6% of the Indian population, but earn only 13.9% of the national GDP. This gross mismatch can be alleviated by improving farmers' access to information and expert advice (*e.g.*, knowing which seeds to sow and how to treat pests can significantly impact yield). We developed FarmChat and conducted an evaluative study with 34 farmers near Ranchi in India, focusing on assessing the usability of the system, acceptability of the information provided, and understanding the user population's unique preferences, needs, and challenges in using the technology. We performed a comparative study with two different modalities: audio-only and audio+text. Our results provide a detailed understanding on how literacy level, digital literacy, and other factors impact users' preferences for the interaction modality of conversational agents. We found that a conversational agent has the potential to effectively meet the information needs of farmers at scale. More broadly, our results could inform future work on designing conversational agents for user populations with limited literacy and technology experience.

## 6.1 *Why Chatbots for Farmers?*

According to the 2011 census, 54.6% of the Indian population is engaged in agriculture [58], but earning only 14.6% of the country's GDP [201]. To address this gross mismatch, the Indian government aims to double farmer incomes in the next five years [145]. It is well believed that access to information and expert advice is crucial in achieving this goal [36, 75, 78, 154]. Such information includes the choice of seeds to sow, how to combat specific crop diseases, weather forecast based advisory, and optimal harvesting times. However, farmers in rural India often have limited access to such information [75, 78, 154, 187]. Even when the information is available, farmers are often unable to consume it due to illiteracy, as India has the lowest adult literacy rate in the world [198].

Several solutions have been proposed to solve the limited information access issue faced by farmers in the developing world. This includes forums for asking questions to experts and peers [78, 154], peer education using participatory video [75], interactive voice response (IVR) systems [64, 163, 187], and social networks for farmers [111, 133]. Since 2004, the Indian government has been operating the Kisan Call Centre (KCC). The KCC [78] is a toll-free call-centre to answer farmers' queries in 22 local languages daily between 6 am to 10 pm. However, it is difficult for the manually operated call-center to keep up with the massive demand. In June 2014 alone, 1.11 million calls were received by the KCC, out of which over 450,000 (~40%) went unanswered [31]. Thus, it remains an open problem to build a system that satisfies the information needs of rural farmers. Systems meant to serve this population must be usable and acceptable by people with limited literacy, highly scalable, available around-the-clock, responsive, and have a manageable overhead for agricultural experts (referred as agri-experts).

As a potential solution, in this work, we introduce an automated conversational agent, or chatbot, to provide farming related information through natural speech interactions. A chatbot offers several benefits that can potentially satisfy the above-mentioned requirements. First, speech is the most familiar mode of interaction that requires little learning or liter-

acy. In fact, audio-based interactions are considered the preferred—sometimes the only usable—interactions for illiterate users [130, 132]. Interactions with an agent should enable farmers to formulate queries as if they were talking to another person. Second, conversational agent systems offer direct information access without the need to navigate complex information paths as often required by graphic user interfaces (GUI), and simplicity is considered a primary design requirement for low-literate users [130]. Finally, from a system point of view, a chatbot is a scalable solution that can be accessed by any user at any time. Moreover, agri-experts can review user inquiries to the chatbot periodically and then continuously expand the chatbot knowledge base without high maintenance cost.

## **6.2 Formative Findings**

Two sources of knowledge informed the development of FarmChat: farmers’ information inquiries with the Kisan Call Center (KCC) and findings from a formative study with local farmers and agri-experts.

The Government of India has made all logs of calls to the KCC from January 2015 to September 2017 publicly available. In total, this corpus contains data for 8,012,856 calls. Each call log has 11 fields, including the date and time of the call, location, crop (one of the 306 crop types), query, and the answer provided by the KCC agri-expert.

For implementing FarmChat, our system was restricted to potatoes since most farmers around Ranchi were engaged in potato farming during the study period. Moreover, farmers are more keen to gain knowledge about the crop they are currently farming [75]. There were 85,852 calls related to potato crop in the KCC dataset. We performed topic modeling using LDA [37] on these calls and found that the top 5 queries for potato farming were pest and disease (52,070 queries), weather (11,628), best practices (5,648), fertilizer use (4,049), and seeds (3,646); these calls constituted 89.7% of the total potato farming calls. The majority of pest and disease queries (17,668) were about the late blight disease. FarmChat covers these main areas of questions with curated answers.

The KCC dataset does not contain the complete dialogues between the farmer and the

KCC expert, but rather a limited summary of the question and the answer provided. To fill this gap and validate that the common questions identified from the KCC dataset apply to the local situations around Ranchi, we conducted semi-structured interviews with 14 farmers (9 male, 5 female) and 2 male agri-experts, in September 2017. We worked closely with a local agriculture NGO (Non Governmental Organization), where the two agri-experts were employed. They helped us recruit the farmers and obtain their consent for participation, following their own internal ethics policies. The farmers were from three different villages situated within 50 miles of Ranchi. Five farmers were *literate* (can speak, read and write Hindi), three were *semi-literate* (can speak and read Hindi), and six were *illiterate* (can only speak Hindi). The definition of different literacy levels have been adapted from previous works [73,115]. Six of the literate and semi-literate farmers owned a smartphone with Internet access. Both agri-experts had a graduate degree in agriculture and more than 15 years of farming advisory experience. Though five of the farmers and both the agri-experts had heard about the KCC, only two of them have tried calling the service and none of their calls were answered. One of the researchers conducted the interviews. The interviews were conducted in Hindi and took 20 minutes. All sessions were audio-recorded, and were transcribed and translated to English later. Both the farmers and the agri-experts participated in the study voluntarily without compensation. From the interviews, we tried to understand: (1) What are the information needs of these farmers? (2) What are their current approaches to seek that information? (3) What are the concerns and limitations of these approaches?

These questions are intended to inform the potential usage patterns of FarmChat. We performed a thematic analysis [24] on the interviews data to identify themes related to the above three questions.

### 6.2.1 Information Needs

The farmers (denoted as  $F$ ) and agri-experts (denoted as  $AE$ ) provided us with similar questions as the ones we found in the KCC dataset. Based on both sources, we identified four major areas requiring information support:

**Plant Protection:** In the KCC dataset, 60.6% of the potato farming calls were related to remedies for protecting plants against pests and diseases. Similarly, agri-experts stated that a majority of farmers seek suggestions on which medicine to spray for a particular crop disease. None of the farmers we interviewed were aware of any disease name. Usually, farmers describe crop diseases by their visible symptoms to the agri-expert; with a few back-and-forth questions, the agri-expert hypothesizes the issue and recommends medicine with dosage information. This example conversation was provided by AE<sub>1</sub> – F: “*The leaves have big brown spots, what should I do?*”, AE: “*How was the weather in the past few days?*”, F: “*It has been foggy.*”, AE: “(Must be the late blight disease.) *Spray Ridomil ...*”. This is analogous to visiting a doctor with medical symptoms. In the design of FarmChat, we follow a similar Q&A conversational style.

**Weather:** In the KCC dataset, 39.4% of the overall calls were about weather-related questions; 13.5% of potato farming questions were about weather. Farmers eagerly sought weather information, as rains can wash away expensive sprayed pesticides and weather conditions determine the best time to harvest crops. A few commonly-asked questions by farmers were: “*Is it going to rain tomorrow?*” - F<sub>3</sub>, “*Should I spray pesticide today?*”- F<sub>8</sub>. The farmers we interviewed primarily relied on local regional television channels for weather information.

**Best Practices:** Information related to best practices can help increase yield in terms of the quantity or quality of potatoes. Common questions were: “*Till what height should I put water?*”- F<sub>1</sub>, “*After how many days, should I harvest?*”- F<sub>12</sub>. These best practices questions comprise of 6.6% of the potato farming calls in the KCC dataset. Agri-experts also stated that farmers consistently asked them tips to increase yield and, consequently, income. For instance, “*On harvest, I got small potatoes. What did I do wrong?*”- AE<sub>2</sub>. Such a question would lead to a longer conversation, as the agri-expert lists different possibilities that might have resulted in small-sized potatoes, *e.g.*, seed size, fertilizer used, irrigation. FarmChat is designed to follow similar Q&A conversations.

**Unbiased Recommendations on Products:** Apart from best practice questions,

farmers wanted recommendations from agri-experts on products they should purchase. Questions such as “*Which fertilizer to put and how many times?*” - F<sub>3</sub> and “*Which seeds are the best for red potatoes?*” - AE<sub>2</sub> were commonly asked. They prefer to ask these questions to agri-experts instead of local shopkeepers, believing that agri-experts would provide unbiased and trustworthy response; they feared that shopkeepers may be motivated by the profit margin of products.

### 6.2.2 Current Information Sources, Challenges and Design Requirements

With the formative study, we focused on understanding farmers’ current sources to access farming related information and identifying their limitations, in order to inform design requirements for FarmChat. In Table 6.1, we summarize the main information sources and challenges in using them. Based on them, we identify the following design requirements and how we intend to address them in FarmChat.

**Specificity:** Farmers often request information on highly specific problems such as diagnosing issues of their crop, which is hard to satisfy by current automated information sources, including Google search and agri-apps. Farmers often have to turn to agri-experts but their availability is limited and it is difficult for AEs to address farmers’ inquiries in a timely fashion. Chatbot enables more interactive search through multi-turn conversations [159]. We utilize that in FarmChat by asking follow-up questions to narrow down the information queries if needed. With the input from agri-experts, we attempted to follow their way of diagnostic discussions with farmers.

**Localization:** A main area of farmers’ information needs focuses on local information such as weather and harvesting time. This again cannot be satisfied by most automated information sources or television. In designing FarmChat answers, we relied on local agri-experts to customize the information based on local conditions.

**Trust:** Trust is a theme that repeatedly appeared in the interviews. Most of the farmers were highly skeptical of the advice given by friends, shopkeepers, and agri-apps. Given the importance of agriculture information to the farmers’ livelihood, it is critical and also

Source	Description	Count	Quotes and Challenges
Self	Prior knowledge	14	<i>“Just do what I think is right”</i> - F <sub>9</sub> ; <i>“gut feeling”</i> - F <sub>2</sub> .
Friends	Others with prior experience	7	<i>“Ask my friends who have previously seen the same disease”</i> - F <sub>3</sub> ; <i>“Because of jealousy, others give wrong advice. There is a lot of ill feeling... I never trust them.”</i> - F <sub>13</sub> [Distrust].
Shopkeeper	Shopkeepers selling agricultural products	10	<i>“They just want money so sell their unsold stuff.”</i> - F <sub>9</sub> ; <i>“Sometimes he sell stuff which is not required. Only 1 medicine can solve it, but he will give 2 medicines.”</i> - F <sub>13</sub> [Distrust].
Television	For daily weather updates	14	Not localized to their particular region [Localization].
Agri-Advisory	Receives automated calls 2-3 times a week with agri-tips from the Government	8	<i>“very generic... not relevant for my current problem”</i> - F <sub>2</sub> ; <i>“The advices are mostly for North Bihar which has more fertile land. Here we have rocky land, so all those advice fails.”</i> - F <sub>7</sub> [Specificity, Localization].
KCC, Agri-Expert	Call KCC or agri-experts, employed/trained by the local agri-NGO	11	<i>“They (KCC) don't pick the call, ever.”</i> - F <sub>3</sub> ; trust the responses from agri-experts but hesitate to bother them too often; advice received over the phone was hard for farmers to remember as they can not write down the responses due to illiteracy [Availability, Persistence].
WhatsApp with Agri-Experts	Hesitant to interrupt agri-experts, so message them on WhatsApp with images of the diseased crop	5	<i>“delayed responses”</i> - F <sub>2</sub> ; <i>“It does not work. For the image of infected crop, we need to ask a few follow-up questions... Whatsapp is like a 1-way communication channel with high lag... as they are illiterate.”</i> - AE <sub>1</sub> [Availability].
Google, Apps	Farmers with smartphones search on Google or agri-apps (such as MyAgriGuru, Gramophone)	6	<i>“You ask one thing (to Google) and get thousand responses. What to do!”</i> - F <sub>11</sub> ; <i>“With Google, I do not trust the response.”</i> - F <sub>3</sub> ; <i>“the apps are developed by big agricultural corporates... to advertise their own products”</i> - F <sub>8</sub> [Trust, Specificity].

Table 6.1: Current information sources and their challenges summarized in the square brackets.

challenging to build trust. The literature suggests that trust implies both competency and unbiased viewpoints [74], which we attempted to achieve through iterative development of the knowledge base with pilot studies, and acquiring contents from agri-experts of local NGOs. Whether farmers could trust the information provided by FarmChat is an open question that we look to answer in our study.

**Persistence:** While using most audio-based systems, such as automated agri-advisory, the KCC, agri-experts, *etc.*, due to low literacy, farmers cannot write down the recommendations given to them, resulting in repeated inquiries for the same information. The audio-based chatbot allows easy re-access of information. The Audio+Text version offers additional benefit of persistence, where the user can easily access previous messages.

**Availability:** Agri-experts and the KCC are not available throughout the day. Farmers also feel it is socially unacceptable to contact experts too frequently, however responses on WhatsApp cannot satisfy immediate information needs. Google search and agri-apps are thus favored by some farmers (4/14) to access information at any time. The chatbot solves the problem by allowing access at any time anywhere from one’s mobile device.

In summary, we designed the proposed novel chatbot solution, FarmChat, to respond in real-time, be available 24x7, provide specific responses to custom queries in the local language, and give localized answers provided by the local agri-experts ensuring the information content received by farmers to be unbiased and trustworthy.

### **6.3 FarmChat System Design**

In this section, we will describe the developed FarmChat system (Figure 6.1). We will start with providing the system’s overview, followed by describing its three major components: the two versions of the user interface (Audio-only and Audio+Text), the generated knowledge base derived from the KCC dataset and inputs from agri-experts, and the conversational intelligence combining intents, entities, and dialogue.

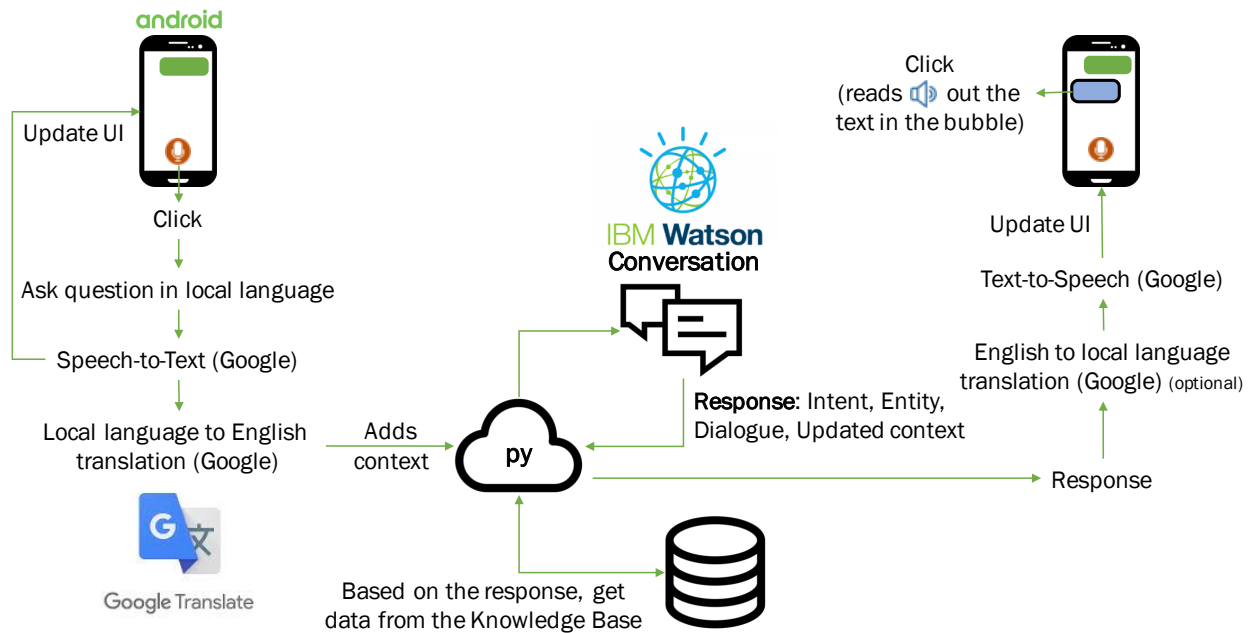


Figure 6.1: System Design of FarmChat

### 6.3.1 System Overview

To interact with the FarmChat mobile app, the user clicks the red microphone button and speaks after hearing a ‘beep’. Once the app detects a long silence, it stops listening. After every user speech input, the screen displays a waiting icon to acknowledge that the system has begun to process the user’s input and is retrieving a response. In the current version, FarmChat supports Hindi. We chose Hindi because 41.1% of Indians are native speakers of Hindi [79], and all the farmers in our study region know Hindi. The phone app passes the received speech input through Google’s Speech-to-Text transcription service (Hindi speech to Hindi text) and then through Google’s language translation service to translate the Hindi text to English (Figure 6.1). The translated English text, along with the current context of the conversation, is then passed to a back-end application built with Python Flask. The Python application passes the English transcription and context to IBM’s Watson Conversation service, which identifies the intent and entity in the text based on a pre-defined language

model. The IBM Watson responds with an updated context and response in English text, based on a pre-defined dialogue flow. The details of the dialogue flow will be discussed in Section 6.3.4. The response from IBM Watson is received by the Python application and may consist of a few empty fields. Those empty fields are populated by retrieving data from the domain-specific knowledge base database. For example, when a user asks “*can see black spots on the leaves, what to do?*”, the IBM Watson understands that the user needs information related to the ‘cure’ of ‘late blight disease’ for ‘potato’. Information for the cure is then retrieved from the knowledge base. Separating the dialogue flow logic and knowledge base allows agri-experts to make easy edits to the knowledge base without impacting the conversation model. The final response text with updated context is sent to the FarmChat phone app. If the received response is in English, the system translates the response to Hindi using the Google translation service. FarmChat then reads the text to the user, using Google Text-to-Speech, thus completing the conversation.

### 6.3.2 User Interface

Most rural farmers in India are illiterate or semi-literate [115]. Recent work suggests supporting semi-literate users differently in interface design from illiterate users [73]. The appearance of text might negatively impact the illiterate users [132], while text for semi-literate users offer faster and unambiguous mode of interaction [73]. More importantly, the additional modality of text-based output allows persistent access to previous messages, which is identified to be our key design requirement. In this work, we examine two variants of FarmChat: Audio-only (input: speech; output: audio) and Audio+Text (input: speech, button; output: audio, text, image).

#### *Audio-only FarmChat*

The Audio-only FarmChat interface (Figure 6.2) consists of only two buttons: (1) a red ‘microphone’ button that the user needs to click to provide speech input, and (2) a blue ‘play’ button, which enables the user to listen to the chatbot’s last response again. Users

can click on the blue button any number of times to repeat the most recent response. While the bot is playing the last response, the blue button can be clicked again to pause the response. After the user’s speech input is received, the interface shows a waiting icon and does not allow the user to click any of the buttons. This was done based on the results from a pilot study with 4 farmers that uncovered major design issues. We found participants had a tendency to talk further while waiting for the bot’s response, perhaps presuming that the bot did not understand the previous input. Once the Audio-only FarmChat app receives the bot’s response, it removes the waiting icon, and speaks out the response.

#### *Audio+Text FarmChat*

The Audio+Text FarmChat user interface (Figure 6.3) closely resembles a typical text-messaging interface, wherein the user input and bot response are presented in message bubbles. There are two major differences from a typical text-messaging interface: (1) Audio+Text FarmChat can only receive speech and button click input, not text; (2) The text/image output can be processed as audio, *i.e.*, clicking on a message bubble in Audio+Text FarmChat results in it being read aloud through Text-to-Speech.

Similar to Audio-only FarmChat, the user needs to click the red microphone button and then speak in order to provide speech input. Using Speech-to-Text, the system converts the Hindi speech input into Hindi text, which appears in a new green-colored bubble (aligned right) on the messaging interface. The response text from the bot appears in a new blue-colored bubble (aligned left) on the interface, and it is read aloud using the Text-to-Speech service. If multiple bubbles get added in the same response, they are read aloud in order. Users can scroll up to access previous messages; tapping any message bubble reads it aloud to the user. The message bubble currently being read gets highlighted with a black border around it. The user can pause the audio play by tapping the same bubble again. The audio automatically pauses when the user presses the microphone button to provide new input.

Audio+Text FarmChat use images in two ways – for asking multiple choice question, and for explaining a farming concept. As shown in Figure 6.3, for example, the user asks

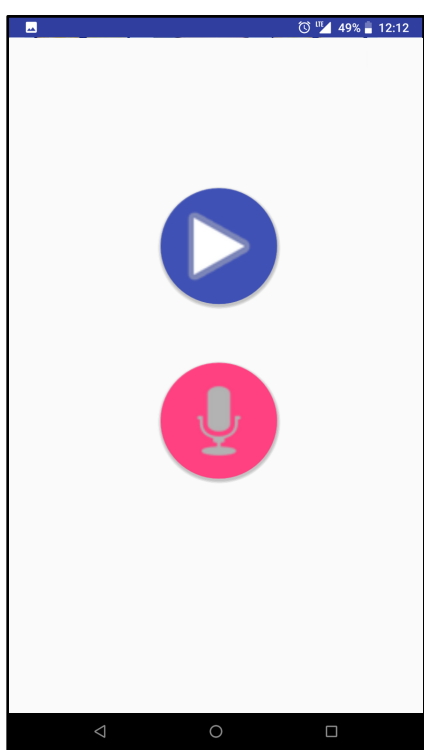


Figure 6.2: User Interface of Audio-only FarmChat

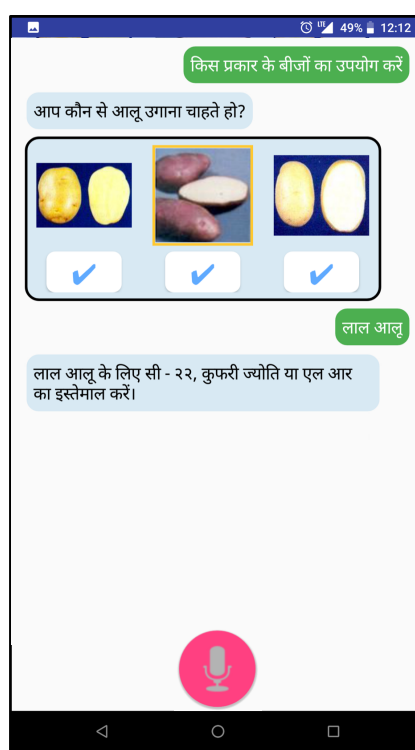


Figure 6.3: User Interface of Audio+Text FarmChat

FarmChat ‘*what kind of seeds should I use*’, to which the bot responds, ‘*which potatoes do you want to grow?*’ with three images as options ‘*white potato*’, ‘*red potato*’, and ‘*potato for chips*’. The user can either press the microphone button and say the response to the question aloud or click one of the three tick-mark buttons to select a particular option. The user can also click the image to hear a description of the selected image. The response from the bot may also contain images to explain certain concepts. For example, regarding the question ‘*how much water is ideal for irrigation?*’, the bot responds with an image explaining that ‘*give water up to two-thirds of the ridge’s height*’.

### 6.3.3 Knowledge Base

We developed a knowledge base for potato farming using the KCC dataset and information collected from formative interviews with farmers and agri-experts. For each of the identified topics in Section 6.2.1, we asked the two agri-experts (who participated in the Formative Study) to provide examples of typical farmer questions, the follow-up questions that they would ask in order to understand the problem, and the final advice they would provide. All such conversations were added to the IBM Watson Conversation dialogue flow, and the informational advice was included in the FarmChat knowledge base. In the current version, the knowledge base is a SQL database consisting of four tables, one for each of the topics identified in Section 6.2.1. Each table has a crop type, multiple tags, and the knowledge from agri-experts, *e.g.*, one of the rows in the Recommendation table has ‘*potato*’ as crop type, ‘*seeds*’ and ‘*red*’ as tags, and ‘*C-22, Kufri Jyothi, and LR*’ as the recommendation associated with the combination of these tags and crop type. The knowledge base can be easily extended to add more crops and/or more data related to each crop.

The knowledge base was iteratively populated. During the pilot study, we learned that a few standard Hindi words are not well understood by the local farmers. For instance, given the instruction “*for irrigation, give water up to two-thirds of the ridge’s height*”, many farmers did not understand the phrase “two-thirds”. We modified the knowledge base using local terms with the agri-experts help, adding examples to make it easier for the farmers to

understand. In the above example, for instance, we added “... *for 6 feet ridge’s height, fill it till 4 feet of water.*”

#### 6.3.4 *Conversational Intelligence: Intent, Entity and Dialogue*

During our formative interviews with farmers and agri-experts, we learned that the farmers are often not familiar with disease names and are most likely to describe the visual symptoms. Supporting such interactions requires developing a robust conversational back-end.

Conversation systems identify *intents* and *entities* from user’s input to understand the meaning of the user text. The user’s *intent* is the current goal or purpose of their interaction with the chatbot. The *entities* add value to that purpose and narrow it further to make it specific. For any chatbot, the intent and entity types are defined by the chatbot designer. As the conversation involves multiple back-and-forth rounds between the user and the chatbot, the conversation system maintains *context* to keep track of the discussion. The context comprises of a selection of intents and entities from previous utterances. Without context, a user’s new input would be processed oblivious of their previous inputs. Based on the intent, entity and context, the response is generated as per the *Dialogue flow* detailed by the chatbot designer. As an example, here is a typical conversation with FarmChat:

Human: *can see black spots on the leaves, what to do?*

Bot: *tell me how was the weather for the last 5 days?*

Human: *mostly rainy*

Bot: *spray Ridomil ...*

From the first message by the user, FarmChat recognizes that the intent is ‘*plant protection*’, with the entity being ‘*black*’ spots. Black spots can be a symptom for several diseases. To accurately diagnose the problem, FarmChat asks a follow-up question taken from the Dialogue flow. As the conversation proceeds, the chatbot maintains the context of plant protection so that the ‘*rainy*’ intent corroborates with black spots to identify ‘*late blight*’ as

the disease. Overall, we defined 38 intents, 12 entities, and the dialogue flow in the IBM Watson Conversation.

## 6.4 *FarmChat Comparative Evaluation*

We conducted a user study (in February 2018) to understand the usability of the FarmChat system.

### 6.4.1 *Participants*

Thirty-four farmers (15 female, 19 male) with average age of  $40.5 \pm 14.3$  years and farming experience of  $17.9 \pm 11.5$  years participated in the study. They were from six different villages situated within 50-100 miles of Ranchi, India. A local NGO staff helped to recruit the participants. The NGO staff also obtained consent of the farmers for participation, using their own internal ethics policies. The NGO actively collaborated with the researchers on this project; neither the NGO nor its staff received any funds from us.

All participants spoke Hindi and Khortha, the local language. Eleven of the participants were illiterate, nine were semi-literate and fourteen were literate (Table 6.2). Thirteen participants owned a touch-screen smartphone. Only one of them used the smartphone just to receive incoming calls. The remaining 12 participants were typical smartphone users, using the phone for browsing Internet, watching videos, and extensive WhatsApp usage. We consider these participants to be *digitally-literate*. Among the 22 digitally-illiterate participants, one owned a smartphone and twelve had access to a smartphone at home with their son or husband, but they only used it to receive incoming calls. Nine out of the 12 digitally-literate participants were agri-entrepreneurs (Table 6.2). *Agri-entrepreneurs* are farmers who have undergone a 35-day agriculture-related training program conducted by the local NGO. As part of the training, each agri-entrepreneur receives a Samsung smartphone (costing 60 USD). Overall, nine participants knew about the KCC. Of those participants, only two had called the KCC, and only one had received a response once. None of the participants had any prior experience with a chatbot. Participation was voluntary without financial compensation.

#### 6.4.2 Procedure

We conducted a task-based user study to compare the two interfaces using a within-subject design. The ordering of the interfaces was randomized across participants to counter ordering effects. Before the tasks, participants went through a training task to learn how to use each interface. The training comprised of three dialogues: (1) Exchanging greetings: FarmChat asks the participant *“how are you?”* and waits for the participant to respond. (2) Gender information: FarmChat asks about the participant’s gender. For Audio+Text, it shows multiple choice image-based options. (3) Family size: FarmChat asks the participant the number of people in his/her family and expects a number in response. The training was the same and required for both interfaces. To help participants remember the primary mode of interaction, they were told: *“to listen, press blue; to speak, press red”*.

After completing the training successfully, participants were asked to perform three tasks: a structured task, a semi-structured task, and an unstructured task (in that order). The structured task was used in the beginning to help familiarize participants with the interface. For the structured task, participants were shown paper-printed color images of symptoms related to four common potato pests/diseases. Participants were asked if they have seen any of these pests/diseases in their own field or in their neighbor’s recently. If yes, the participants could ask FarmChat questions about it. This was done to ensure that the participants were able to answer follow-up questions by FarmChat about the particular pest/disease. Each question related to pests/diseases requires at least three dialogue exchanges in order to ascertain the remedy. For the semi-structured task, participants were shown paper-printed color images of four major farming practices: buying input seeds, seeding, irrigation, and harvesting (including bad yield). Participants were required to ask at least one question related to farming practices to FarmChat. Finally, for the unstructured task, participants were encouraged to ask any potato farming-related questions on their mind. This task was aimed to help us identify the knowledge gap within FarmChat. Throughout all of the tasks, one of the researchers helped participants when needed, to ensure that all tasks were

completed.

Apart from these tasks, demographic information was collected at the beginning of the study. After completing all the tasks with a particular interface, participants were asked to rate their usage experience by answering eight questions on a five-point Likert scale – enjoyment in using the system, ease of use, difficulty in learning, support they would need in the future to use it, satisfaction in the input interpretation, trust in the response, speed, and willingness to use it in the future. Note: Except the questions regarding difficulty in learning and future support, higher scores indicated a better experience. These questions were followed by a semi-structured interview, which delved deeper into specific features that participants liked or disliked with suggestions for improvement. At the end, participants were asked about their preference between the two interfaces.

During our pilot study with 4 participants from the same demographic, we encountered several errors with the Speech-to-Text and language translation services. Hence we decided to complement these services with a wizard-of-oz approach. Specifically, one of the researchers acted as a wizard, and updated the participants’ transcribed and translated text minimally to correct any errors from the Speech-to-Text and language translation services before passing those messages to the Python-based back-end application. It is important to note that the wizard did not interfere with the conversational back-end (*e.g.*, intent classification). This is in contrast to a complete wizard-of-oz system where the wizard is a proxy for an automated conversational system (*e.g.* [183]).

The two researchers – one conducting the study and another acting as wizard – were accompanied by a local NGO staff member during the study. All interactions between the researchers and participants were in Hindi. All sessions were audio-recorded, and were later transcribed and translated to English. Participants’ interactions on the phone and the wizard’s interactions on the computer were logged. The study took place in the participants’ home or farm land. On average, each participant took 60 minutes.

### 6.4.3 Hardware

All the participants performed the tasks on a Lenovo K8 Plus phone (screen: 5.2 inch, price: 150 USD). The current FarmChat system requires access to Internet, and the Internet data requirements were ideally fulfilled by a 4G SIM card on the phone. Moreover, the wizard’s MacBook Air laptop also used the phone’s Internet using hot-spot.

## 6.5 Results

In total, participants (denoted as  $P$ ) provided 626 inputs to FarmChat. Inputs entail only speech in Audio-only FarmChat, and both speech and button clicks in Audio+Text FarmChat. Inputs can be in the form of questions, answers, or comments. General statistics are summarized in Table 6.3. We conducted paired t-tests between the two interfaces on various parameters, and did not find any significant difference between the two interfaces. Figure 6.4 shows the Likert scale ratings for FarmChat. To compute this, we used the ratings given to the FarmChat interface that was *preferred* by the particular participant.

For evaluating the FarmChat system and understanding the opportunities and challenges of using conversational systems for rural Indian farmers, we focus on three main aspects: (1) the acceptability of FarmChat as an information system to satisfy farmers’ information needs, (2) the usability in interacting with conversational interfaces, and (3) the preference between the two variants of conversational interfaces – Audio+Text versus Audio-only – and how it differs for different user populations (literate versus semi-literate versus illiterate users, digitally-illiterate versus digitally-literate users, farmers vs agri-entrepreneurs). To explore these, we rely on demographic information, log data, Likert-scale ratings, user study notes by the study facilitator, and audio transcriptions of the user study and post-study interviews. Two authors individually reviewed the audio transcriptions and used open coding to extract themes [24]. Codes were harmonized after two iterations of review and discussion, resulting in the final set of themes in each of the above mentioned aspects.

<b>LITERACY</b>	<b>Illiterate</b>	<b>Semi-literate</b>	<b>Literate</b>
Number of participants	11	9	14
Gender (M,F)	3,8	7,2	9,5
Age (m±std)	43.9±15.9	43.1±16.2	36.1±11.2
Years in school (m±std)	1.3±1.3	5.6±1.7	10.3±3.0
Digitally-literate	0	1	11
Agri-entrepreneur	0	0	9
<b>Preference (A-o,A+T)</b>	11,0	5,4	2,12
<b>SMARTPHONE</b>	<b>No</b>	<b>Access</b>	<b>Owns</b>
Number of participants	9	12	13
Gender (M,F)	6,3	5,7	8,5
Age (m±std)	45.5±16.4	41.8±15.2	35.8±11.2
Years in school (m±std)	2.7±3.0	4.3±5.1	9.2±3.9
Literacy (I,SL,L)	4,5,0	6,3,3	1,1,11
Agri-entrepreneur	0	0	9
<b>Preference (A-o,A+T)</b>	9,0	6,6	3,10
<b>PROFESSION</b>	<b>Only farmer</b>	<b>Agri-entrepreneur</b>	
Number of participants	25	9	
Gender (M,F)	13,12	6,3	
Age (m±std)	42.4±14.7	35.3±12.3	
Years in school (m±std)	4.2±4.5	10.1±3.4	
Literacy (I,SL,L)	11,9,5	0,0,9	
Digitally-literate	3	9	
<b>Preference (A-o,A+T)</b>	18,7	0,9	

Table 6.2: Demographic data for the user study (refer Section 6.4.1). Note: The same 34 participants are categorized with respect to their literacy level, smartphone usage and profession. (A-o: Audio-only; A+T: Audio+Text; I: Illiterate; SL: Semi-literate; L: Literate; m=mean; std: standard deviation)

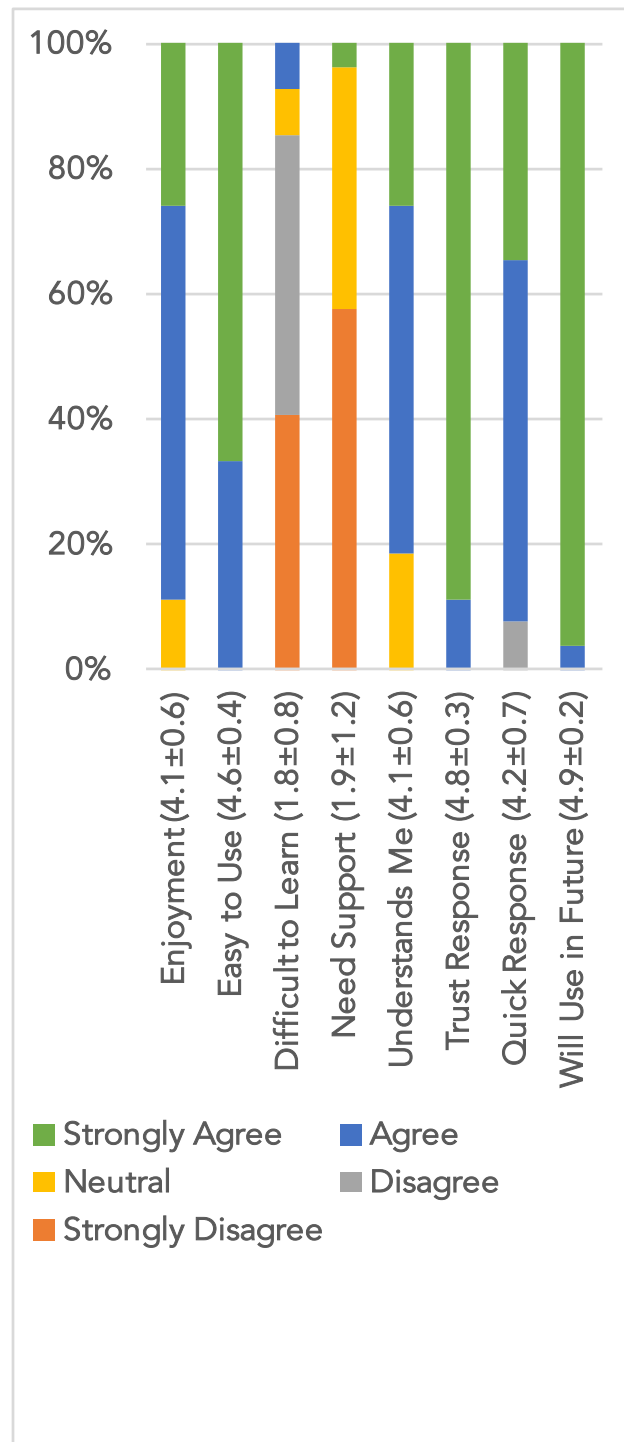


Figure 6.4: Likert-scale rating considering ratings for the interface preferred by the participant

### 6.5.1 Information Acceptability

Overall, we found FarmChat to be generally acceptable by the farmers as an information source to satisfy their farming information needs. All participants expressed willingness to continue using FarmChat in the future (Figure 6.4). The major reasons that farmers enjoyed using FarmChat were immediate responses to their queries and constant access to farming-related knowledge. For example: *“Information is the key... If I know more, I will earn more!”* - P<sub>10</sub>, *“It gave me new knowledge... like treating the seed initially will help... It even told me medicines.”* - P<sub>16</sub>. In particular, responses that included a medicine name and quantity were re-read and replayed most often ( $2.4 \pm 1.5$  times for Audio-only). This may be because they wanted to memorize the names or the hard medicine names were not clear to them in the first attempt. Four participants even asked for FarmChat to be installed on their personal device: *“Give me this on my phone. I will use it. I will learn from it, and even teach others.”* - P<sub>12</sub>. A few participants (3/34) also discussed the potential value of having continuous access to FarmChat beyond in-situ information needs: *“when I am free, I can ask for which fertilizer to use, how much... I can gain new knowledge.”* - P<sub>26</sub>. This suggests that compared to existing human information sources like the KCC and agri-experts, a chatbot system has the potential to better serve farmers’ needs for continuous learning. In the following, we summarize themes that emerged from the user study on information assistance provided by FarmChat.

#### *Precise and Localized Answers*

Specificity and localization were identified as keys to the information needs of farmers. With the help of agri-experts, we carefully tailored the system responses to local conditions. Participants appreciated such information contents: *“Gave precise answers for the disease related questions”* - P<sub>15</sub>, *“I know that these medicines are available locally.”* - P<sub>4</sub>. However, needs for even more fine-grained localization and personalization still emerged. *E.g.*, P<sub>4</sub>, who praised the localized responses complained *“the big seeds it is advising is not available here”*. Also,



Figure 6.5: FarmChat speech-based interaction

*“my land is rocky, the soil is not sandy. Hence for harvesting it requires adding water... she is suggesting to stop water a week before harvesting, which may work well for other’s field, not mine.” - P<sub>4</sub>.*

### *Trust*

Trust is another key design requirement. In general, participants trusted the responses provided by FarmChat (Figure 6.4). Six participants even asked the facilitator to write down the recommended medicines, seeds variety, and/or fertilizer with their quantities for them to refer later. Participants often formed trust in FarmChat by validating its responses with their existing knowledge. For example, *“I used the Bevestin medicine for wheat, and I was really happy with the results. This phone (app) is telling me to use Bevestin for potato, so must be telling the right thing.” - P<sub>10</sub>*, *“Its giving all correct information... some NGO people... a while ago, gave us similar information.” - P<sub>3</sub>*. The researchers were accompanied by a local NGO staff during the study. This assumed endorsement might have also made the FarmChat more trustworthy for the rural farmers, as observed before in prior work [42, 52].

<b>Data Type</b>	<b>Audio-only</b>	<b>Audio+Text</b>
Time spent/participant (mins)	14.8±8.5	12.7±6.6
Total number of speech, button click input	275, –	304, 47
Number of speech input/participant	8.1±2.2	9.2±4.5
Number of button click input/participant	–	1.4±1.1
Length of speech input (words)	6.1±4.8	6.8±4.5
Length of speech input (sec)	7.1±1.3	7.4±1.6
% of bot’s response replayed	51.7±32.2	46.8±33.1*
% mic button clicked but no input	15.4±14.1	18.5±13.0

Table 6.3: Results from log data, mean±std. No significant difference was obtained between the two FarmChat interfaces. (\*Cannot be calculated accurately, as a majority of literate participants read the response multiple times instead of tapping the chat bubble. The reported number is for the sake of completeness.)

### *Over-expectation*

We found participants’ tendency to overestimate the capabilities of FarmChat, which led to some level of dissatisfaction in usage. Lack of clear affordability is a known challenge for conversational interface [127], but the novelty effect for the farmer population seemed to exacerbate the problem. Although FarmChat had high success rate (overall only 21 out of 238 questions went unanswered), some participants (4/34) were still disappointed when one of their questions was not answered and said: “*Why are you making excuses?*” - P<sub>25</sub>, “*it should have the answer to all questions*” - P<sub>24</sub>. Moreover, many farmers (15/34) got excited, and in spite of being told it could only help with potato-farming related queries, they started asking questions about other crops (*e.g.*, onions, tomatoes). In particular, farmers wanted to learn about “new” crops like lettuce and broccoli because no one in their village had grown them before.

### *6.5.2 Chatbot Usability*

Although it was the first time for all participants to interact with a chatbot, they generally found the system to be usable, with the Likert ratings on usability being 4.6±0.4 and difficulty

of learning being  $1.8 \pm 0.8$ . From the user study, the following themes emerged as highlights of the usability and related issues of FarmChat.

### *Anthropomorphism and Familiar Phone Interaction*

FarmChat by design was not anthropomorphic, as we did not introduce any human-like features such as name or character. In spite of that, we found our participants to have a high tendency to anthropomorphize the bot. A few participants (6/34) referred to the bot as “*didi*”, which means elder sister in the local language, since the bot had a female voice. Some participants (4/34) said: “*ok*”, “*good*”, “*yes*”, after every sentence said by the bot, as if they were talking to a human. Also, participants were very polite in their interactions. Questions usually began with “*please listen ...*”, “*can you please tell me ...*”, and ended the conversation with “*thank you for the help*”, as if talking to an agri-expert. Only P<sub>32</sub> explicitly treated FarmChat as a machine, and commented: “*Why should I respond to its good morning. Can I just ask my question?*”. Also contrary to previous research with technologically-advanced users [124, 127], farmers seemed to be “nicer” – more patient and more forgiving – to the chatbot. Participants even tried to frame their questions in formal Hindi, assuming it could help FarmChat understand them better. At times, the participant knew a particular term only in the local Khortha language, hence he/she asked the facilitator: “*What should I say (for) this in Hindi?*” - P<sub>29</sub>. Our participants were trying to help the bot to understand their queries precisely. This is in sharp contrast to previous work [123], in which participants tried to ‘break’ or challenge the bot.

One reason our users anthropomorphized the agent could be that they resorted to their familiar mode of interaction in making a phone call. The farmers had no concept of a chatbot; hence, they interacted with it as if they were interacting with a human. This familiarity enabled participants to quickly grasp the interactions, as suggested by the ratings for difficulty of learning and support needed to use FarmChat in the future (Figure 6.4). However, mismatches between making phone calls and interacting with a bot posed some challenges. For instance, the participants would start responding to FarmChat’s query as

soon as it ends, without pressing the red button. They assumed that FarmChat was always listening, similar to a person on the other side of the phone line. Instead of increasing the phone’s speaker sound, P<sub>17</sub> said “*I can’t listen properly, speak loudly*”, assuming it was a human.

### *Speech as Input*

Participants were pleasantly surprised that FarmChat was able to understand their complex questions in Hindi (rating  $4.1 \pm 0.6$ ). “*The question I am asking, it is able to understand well. Most times, only after a single attempt.*” - P<sub>26</sub>. Note that this was partly due to the wizard’s role in fixing errors in speech processing; 36.3% speech inputs were fixed by the wizard, as computed from the wizard log files. Speech as input failed for a few of the illiterate participants (4/11) who were not able to speak Hindi fluently. As they have never attended school, they learned Hindi from their neighbors, friends, *etc.* Their spoken Hindi was not grammatically complete and was strongly influenced by the local Khortha language. “*My Hindi won’t be understood by the bot, as its mostly Khortha.*” - P<sub>29</sub>. For such participants, the wizard edited 72.7% of the speech input.

Even when other interaction modalities were available, we found that participants still preferred to rely on speech input. For example, when buttons were available to provide precise input in Audio+Text, participants used speech  $28.2 \pm 39.7\%$  of the time to provide input instead of clicking the button. Moreover, instead of clicking the text-box (Audio+Text) or the replay button (Audio-only), participants occasionally asked FarmChat to “*repeat what you said*” - P<sub>8</sub>. One potential explanation is that switching between touch and speech input was cumbersome.

### *Responses by Speech*

A majority of the participants (18/34) appreciated the fast responses given by the bot in speech: “*very quick response, no wait*” - P<sub>26</sub>. The average response time was  $9.2 \pm 2.8$  sec, which includes Google transcription and translation time ( $0.9 \pm 0.2$ ), wizard time ( $5.7 \pm 4.5$

during edits,  $2.1 \pm 1.2$  without any edit), Watson Conversation service with Python Flask response time ( $3.1 \pm 1.1$ ), and network delays. Using button clicks for input significantly lowered the response time to  $1.7 \pm 1.0$  sec since neither the wizard nor transcription/translation services were used. Participants were generally satisfied with FarmChat's response time (Figure 6.4). Echoing the finding on persistence from our formative study, another highly appreciated feature was the capability to replay speech output, as participants at times had trouble understanding or memorizing the bot's responses— "*if I missed hearing something... I can easily play it again*" - P<sub>23</sub>. Users clicked the bubble/button for replay frequently.

### *Interaction Order as a Usability Challenge*

For a majority of illiterate and digitally-illiterate participants, the ordering of pressing the red microphone button, waiting for the beep sound, and then speaking was challenging to follow. Note that this order is required by the current speech input technologies. Many times, participants did not press the microphone button at all or started speaking before the beep sound. At times, the mic button was pressed but no speech input was received (Audio+Text:  $18.5 \pm 13.0\%$ , Audio-only:  $15.4 \pm 14.1\%$ ). This often happened when the participants were thinking of what to say next, but the app stopped listening after it detected a long silence. Moreover, the interaction convention in Audio+Text interface was challenging for those who were unfamiliar with messaging interface. For example, such users were confused whether to scroll up or down to see the previous chat bubbles and had trouble understanding the addition of new chat bubbles. This again highlights the challenge of designing for population, which has limited experience with technology.

### *6.5.3 Interface Preference*

We compared user responses to Audio-only and Audio+Text FarmChat interfaces. Our results suggest that users' preferences were highly dependent on the participants' literacy level, digital-literacy level, and other individual factors like profession, physical and environmental factors.

### *Literacy Level*

All the illiterate participants preferred the Audio-only interface. Among the literate and semi-literate participants, 16 participants preferred Audio+Text, while others preferred Audio-only (Table 6.2).

**Reading vs Listening:** Literate and semi-literate farmers preferred Audio+Text for several benefits of text. First, the Audio+Text FarmChat enabled them to quickly access relevant information in the bot’s response, such as medicine names and quantities; in the Audio-only interface, users had to listen to the whole response again. Second, participants found it easier to memorize “*crucial*” information (*e.g.*, seed name, fertilizer quantity) after reading it, compared to information that was only heard: “*can not remember just by listening*” - P<sub>26</sub>. This has been supported by prior human memory research [35]. Third, reading allowed the participants to take in information at their own pace. “*Listening is easier, but she speaks very fast. I can read slowly in the first one (Audio+Text)*” - P<sub>32</sub>. Fourth, the facilitator noticed that three participants took pride in the fact that they could read, with one saying “*it (Audio-only) is for illiterate people to use... I can read easily!*” - P<sub>32</sub>. This is also supported by the behavioral data, as (semi-)literate participants chose to listen again much less frequently compared to illiterate participants ( $0.7 \pm 0.3$  versus  $2.2 \pm 0.7$  times). This finding echoes a previous study, showing that a few participants (3/11) preferred the textual interface even though they could not read, to avoid the stigmatized perception of illiteracy [110]. However, seven of the (semi-)literate participants preferred Audio-only FarmChat because they found it to be less mentally demanding (“*nothing to read, so no tension.*” - P<sub>10</sub>), they had limited reading skills (“*I can’t read fast enough*” - P<sub>14</sub>), and/or other environmental and physical reasons (discussed in Section 6.5.3).

**Persistent Presence of Messages in Audio+Text:** A key advantage of the Audio+Text interface is that previous conversations are persistently shown on the interface and can be accessed later by scrolling up to the message. This was appreciated by the semi-literate participants. “*I will not remember the medicine names, its quantity. In this*

(Audio+Text), *I can open again, whenever needed, and find the details.*” - P<sub>6</sub>. On the other hand, persistently showing all information has its disadvantages – *“There is so much competition among farmers. If it is on my phone, other farmers may also read and learn. With the other one (Audio-only), I will only listen.”* – P<sub>10</sub>.

**Complexity of Audio+Text:** The researchers noticed that most illiterate participants struggled with the Audio+Text interface, but learned to use the Audio-only interface easily. First, illiterate participants found the Audio-only interface to be less complex, and easy to learn and use: *“Two buttons is easy. No tension... no confusion.”* - P<sub>9</sub>, *“less things to put my mind on”* - P<sub>21</sub>. Second, illiterate users were scared and nervous when confronted with lots of written text, which reduced their confidence level. *“I can’t do this.. its too hard. I won’t be able to do it.”* - P<sub>34</sub>. Third, as they could not read any of the text in the speech bubble, they easily lost their current context of interaction and were unsure of which speech bubble to click next. *“Touching in this (Audio+Text) is confusing as (there were) multiple things to touch, which to touch when... It is hard to remember which bubbles I have already heard, and which I have not heard.”* - P<sub>11</sub>.

**Numerically Illiterate:** The researchers learned that a majority of literate and semi-literate users (11/23) were not able to read Hindi numerals. This is in contrast to previous findings that most illiterate people in India are numerically-literate [132]. One of the main reasons for the disparity is that English numerals are more prevalent in India than Hindi numerals. In FarmChat, numbers are crucial for specifying quantities like medicine dosages.

**Usefulness of Images:** The Audio+Text interface uses images both as part of bot’s questions and to support certain answers. Images in the button click input was appreciated because it allowed participants to visually understand the multiple options. However, participants sometimes tried clicking the image itself for selection rather than the buttons, leading to confusion. The images in bot’s response not only helped the participants better understand the responses, but also allowed them to locate previous threads. Illiterate users highly appreciated images of the medicine package, as they are often concerned about being sold the wrong medicines by the shopkeepers.

### *Digital-Literacy Level:*

Twelve participants were digitally-literate (Table 6.2). Twelve other participants had occasional access to a smartphone at home and were nervous to participate in a study involving one: “*My son doesn’t let me use his phone as he thinks I might break it. He tells me just to click green button to (pick the call and) talk.*” - P<sub>3</sub>. Self-efficacy, both in general and regarding technology specifically, impacted attitude and usage patterns. A majority of the digitally-illiterate users (16/22) preferred Audio-only interface.

**Only two buttons and No scrolling:** Since the Audio-only interface has only two buttons, it was “*easy to learn and use*” – P<sub>3</sub> by digitally-illiterate users. The Audio+Text interface was complex and confusing for them, since each speech bubble was a button. Moreover, due to limited phone screen size, the buttons in the Audio+Text interface were comparatively smaller than that of the Audio-only interface, which made them more likely to miss buttons. Finally, scrolling was not intuitive. The log data shows that none of the digitally-illiterate users scrolled even once in the Audio+Text interface, while participants using smartphone regularly scrolled  $4.5 \pm 2.4$  times.

### *Other Individual Factors*

We also found the following factors affected the interface preference:

**Environmental and Physical:** A few participants (6/34) determined their interface preferences based on environmental or physical factors. Semi-literate P<sub>7</sub> chose Audio-only because “*in the field, in sunlight, I can not read (text on the smartphone display)*”, while P<sub>18</sub> preferred Audio+Text because he was not able to hear properly due to the noise from a nearby tractor. P<sub>14</sub> liked Audio+Text because he has hearing problems, while P<sub>3</sub> and P<sub>17</sub> chose Audio-only because they had eyesight problems. One of the older participants (P<sub>28</sub>) preferred the Audio-only interface because he had shaky hands that prevented him from pressing smaller buttons.

**Profession:** All our participants were farmers. Nine of them were also agri-entrepreneurs,

along with being farmers. All agri-entrepreneurs recommended Audio-only for the farmers in their village, but preferred Audio+Text FarmChat for their personal use in helping others. *“After listening, I might not remember everything. With written stuff (in Audio+Text), I can quickly look it up... only the relevant part and tell others... rather than listening to the whole thing again (in Audio-only). Also, it will certainly reduce my chances of making errors.”* - P<sub>27</sub>, *“I don’t want to listen in front of others and advise them.”* - P<sub>26</sub>. This point is interesting as they considered the interaction modality as part of their self-representation; they did not want to be seen as taking help from their phone. In contrast, agri-entrepreneurs thought Audio-only would suit the other farmers better as they tend to have low literacy and low digital-literacy level. *“Most people here are illiterate”* - P<sub>24</sub>, *“I can easily give this (Audio-only) to others, and provide training on how to use on their own. With just two buttons, they can easily use it.”* - P<sub>12</sub>.

## **6.6 Discussion**

In this work, we propose the use of a conversational agent, implemented using cloud-based scalable services, as a way to help address the information needs of rural farmers. From the user study, we found positive evidence for introducing chatbot technology as a usable solution for serving the information needs of Indian farmers. With FarmChat, it was possible to provide satisfying answers for common questions that farmers encounter in their farming activities. With the help of local NGOs to edit and endorse content, farmers showed trust in the information given by FarmChat. The resemblance with human-human interactions made it easy to use, even for illiterate and digitally-illiterate users. We discuss a few key points as design implications below.

### 6.6.1 Design Implications

#### *Future Conversational Interface*

The study results suggest that the two versions of interfaces—Audio-only and Audio+Text—each have their own advantages, and user preferences depend on several individual factors. Therefore, we suggest a hybrid interface that incorporates positive features from both variants. The hybrid interface would show two buttons, like the Audio-only interface preferred by illiterate and/or digitally-illiterate users. An extra button would allow users to toggle the view and provide a conversation history in the bubble chat format, adopting from the Audio+Text interface favored by semi-literate and literate users. To complement the current limitations in processing speech input, the image-based multiple-choice mechanism to provide precise input can be shown below the two buttons in the hybrid interface when appropriate. It can also show images in the output below the two buttons to aid users in understanding the bot’s responses.

We also suggest the following improvements for interface elements: (a) instead of the red microphone button, a green-colored call button can be used as users related FarmChat directly to making phone calls, (b) domain-specific terms such as medicine names are less familiar to farmers but critical to their work. This results in farmers often reading and listening to such responses multiple times. In the future, the app should emphasize these special terms by highlighting them and/or repeating them in speech. Finally, as we found participants struggling with Hindi numbers, a hybrid language with the numbers in English should be used.

#### *Speech as Input*

First-time smartphone users struggled with providing speech as input. One major issue was that the period of detecting silence was short; while participants paused to think, FarmChat stopped listening, which was frustrating for the participants. There are several solutions to this. FarmChat could listen for longer *initially* by extending the expected silence time at

first and decreasing it over time. Based on the average length of speech input, we suggest 7-8 seconds would be optimal. Alternatively, the chatbot could remind users to press a button to reply, highlight the button after the bot’s response, or be in always listening mode. In the next version of FarmChat, we plan to use a custom solution wherein the user needs to speak with a button being pressed, similar to a walkie-talkie. On releasing the button, the speech would be submitted for processing. This would not only solve the premature submission issue, but would also help with the interaction ordering problem experienced by the illiterate participants.

Hardware buttons (e.g., volume controls) could be enabled to provide inputs. All these possible solutions need to be tested in future deployments.

### *Anthropomorphism*

In spite of the lack of any anthropomorphic features in FarmChat, participants assumed that they were interacting with a human. In the future, it would be interesting to explore the use of an agri-expert character to understand how anthropomorphism impacts the conversation. Participants also wanted to use the app in the future in their free time to gain more knowledge. This would move from a ‘search’ to a ‘browse’ type of information seeking, which hints that a role of more general virtual ‘companion’ is possible.

### *No Reliance on a Wizard*

According to the 2001 Census of India, India has 22 official languages, 122 major languages, and 1599 other languages [79]. In rural India, the local languages have a major influence on spoken Hindi, including accent and word adoption. This results in poor accuracy by Google translation and transcription services. For instance, Google translates the spoken phrase ‘*faala maar diya*’ to ‘slapped’ (in English), however it actually means ‘*hit by fog*’. The reason for this error is that ‘*faala*’ is not a Hindi word, but is adopted from the local Khortha language. Similar errors have been previously reported [61]. Background noise from machinery (including tractors and water-pumping motors) and village festivities further

adversely affected the speech processing performance. In our user study, we opted for a wizard to fix such issues to understand the usability and information acceptability of a chatbot-based solution for farmers. In the future, we aim to develop a completely automated system. One solution is to develop a custom Speech-to-Text and translation service, possibly on a limited set of keywords, for the specific geography that we are targeting (similar to [61]). Another solution would be to use the recently proposed crowd-powered conversational assistant to automate itself over time [95]. This might be challenging as we may not find enough crowd workers with the knowledge of the local language.

### *Image-based Diagnostics*

Along with speech, FarmChat can be enabled to provide image-based input. This is similar to existing plant pathology apps [17], wherein a farmer is supposed to click a picture of the infected crop and the app predicts the disease name and recommends medicine. However, our agri-entrepreneurs and agri-experts found the accuracy of such apps to be really low. In our case, we can combine the speech-based description of the symptoms with the image-based features, which could improve diagnostic accuracy.

### *Hyper-local Information*

The information provided to the farmers needs to be localized, considering climate conditions, soil type, and market availability. To ensure such information is in place, FarmChat needs to collect it from local partners and domain experts, and add to its knowledge base. Moreover, FarmChat needs to support more crops, specifically ‘new’ crops. Adopting from the core idea of Digital Green [75], we can develop a knowledge base from farmers who are experts in one such ‘new’ crop and disseminate that information using FarmChat to farmers in need of that information. Chatbot offers a scalable solution to iteratively expand and customize the knowledge base with expert input and disseminate the knowledge to wide audience.

### 6.6.2 Limitations

We acknowledge several limitations of this work. First, the results should be interpreted based on some specificities of the study, including the relatively small sample size, users' first encounter with a chatbot, and the task scenarios used. The positive user responses and other conclusions should be validated by future, preferably longitudinal studies. Second, the knowledge base was developed with the help of only two local agri-experts, due to the limited resources of the NGO we worked with. The performance of the system could be further improved by adding more agri-experts. Third, besides the limited capabilities of speech-to-text and language translation technologies, which we fixed with a human wizard, FarmChat could also fail due to limited access to Internet in rural India as the core technologies are offered as cloud services. Fourth, we need to conduct system-level analysis, such as scalability testing, power measurement, *etc.* in future work before scaling the proposed solution to masses. Finally, although participants showed high trust in the initial encounter, errors and inaccurate information in daily usage can lead to distrust of the system. We also note the potential risk of disseminating inaccurate information at scale with chatbot technologies. So the knowledge base should be carefully developed and reviewed by subject experts.

## Chapter 7

### CONCLUSION AND FUTURE DIRECTIONS

Throughout this dissertation, we provided support for the thesis statement: **The acceptability of chatbots can be increased by making them more transparent and broadening their applicability to new demographics.**

Chatbots are turn-based, task-fulfilling programs, available in speech and text modalities. Our initial formative study in Chapter 3, involving 16 participants interacting with 8 pre-selected Messenger chatbots for the first-time, over multiple days, spanning almost 10,000 messages, revealed that expectations of chatbot users were not met. Participants were either disappointed or frustrated with mediocre natural language capabilities and the limited set of features offered by the chatbots, to the effect that several of them did not see themselves re-using most chatbots. At an abstract level, the key problem with current chatbots is that its internal working is not ‘transparent’ to the user.

Based on these findings, we aimed to solve the major identified issues with current chatbots, and proposed solutions based on user interface principles.

First, to help users have an understanding of the mental-state of the chatbot during the conversation (helping users and chatbot be on the same page) while sustaining the familiarity of the text-based messaging interface, we proposed adding a context view called *Convey* to the top of the chatbot interface, described in Chapter 4. Moreover, *Convey* provides a summary of the conversation persistently, and adds the benefits of a form-based user interface by enabling entry of precise input through the interactive elements. The results from a 16 participant user study demonstrated that participants perceived chatbot with *Convey* to be faster and easier to use. *Convey* is generalizable to chatbots in any domain, and in future, we expect *Convey* to be integrated and offered by many chatbot-hosting platforms.

Second, dialog failures are one of the key reasons for users abandoning a chatbot. It is not transparent to the users the reason(s) behind conversation failure and how to recover from that breakdown. To design repair strategies for breakdowns of conversational agents, in Chapter 5, we consider key issues based on grounding theory in communication: evidence of breakdown, self- versus other- repair, and cost of repair. We provide a list of eight strategies that capture variances in these dimensions, including a group of novel repair strategies that explain the understanding mechanisms of the underlying model. We conduct a scenario-based study to compare user preferences for these repair strategies, and perform detailed analysis on the reasons behind and individual differences. Our results empirically validate theory-driven guidelines that recommend three levels of contribution from the agent to the collaborative action of repair: acknowledging potential breakdown, providing resource to assist user repair, and proactively suggesting solutions. As a starting point, we encourage future work to develop a unified framework that guides the choice of repair strategies for different individuals and contexts.

Third, to make chatbots accessible to a broader demography, specifically people in the developing world, we proposed FarmChat in Chapter 6 which combines conversational and language technologies to naturally converse with farmers in answering their farming-related queries. Building a scalable, always available and real-time responsive system that satisfies the information needs of rural farmers is an open problem, and one that has significant impact in an agriculture-dominated economy like India. The conversational intelligence of FarmChat was informed by analysis of large corpus of farmer call center logs and guided by agricultural experts who work closely with farmers. Our study with 34 potato farmers in rural India indicated that it is possible to provide satisfying information support to the farmers through chatbot, who attributed trust to the chatbot and expressed willingness to adopt the technology. We also compared the effectiveness of two interface modalities: Audio-only and Audio+Text. This study indicated that although text-based output allows for repeated consumption of the same information, participants expressed different preferences due to literacy, digital-literacy, and other environmental and physical factors. The positive

feedback of the farmers indicates that conversational intelligence as a technology delivered through the ubiquitous smartphone can be an effective tool to improve information access in a rural context for people with limited literacy and technology experience.

This work establishes transparency as a major issue with current chatbots, and provides solutions including *Convey* to reduce mismatch between the chatbot’s context and the user’s perception of the chatbot’s context, and *Resilient Chatbots* to make users aware of reason(s) behind conversation failure and suggest recovery strategies. Moreover, current chatbots are mainly targeted towards a narrow base of technically-advanced users. To increase its acceptability to a wider demography, this work proposes *FarmChat* making chatbots accessible to low literate non-technical users.

## 7.1 Future Directions

In this section, we will discuss a few limitations of our work and possible future directions. (Note: For short-term future work, please refer to the Discussion sections of Chapter 3, 4 and 5).

First, a major part of this dissertation – messenger study, *Convey*, and *Resilient Chatbots* – focus on text messaging-based chatbots. On the other hand, voice-based chatbots in the form of Alexa, Siri, have gained massive popularity overtaking text-based chatbots. With respect to interaction, voice and text-based chatbots are very different from each other. However, the underlying issues that we uncovered for text-based chatbots from our formative study with first-time Facebook Messenger users – mismatch between the chatbot’s context and the user’s perception of the chatbot’s context, and user’s unawareness of reason(s) behind conversation failure and how to recover from that breakdown – are applicable to any bot, including voice-based chatbots. Similarly, some of the learning from this work can be extended to voice-based chatbots, such as providing a voice-based summary of the conversation to the user (similar to *Convey*) after a few round of message exchanges to ensure that the user and the chatbot have the same mental model, providing options to restrict interaction with the user within the chatbots’ capabilities, and providing explanations (instead of text

highlighting, the bot can announce which term(s) it was not able to understand) to help the user recover from breakdowns. In future, it would be interesting to explore and understand how these context related features impact interaction with voice-based chatbots.

Second, our formative study work with first-time chatbot users highlight several limitations of current chatbots. In this thesis, we focus on the *transparency* issue and provide two solutions. Future work can explore various other broad identified challenges, including trust, gender, localization, and relation (master-slave versus companion) with chatbots. Each of these aspects require further understanding of the problem in greater depth, before proposing novel solutions impacting both the interface and intelligence of future chatbots.

Third, this thesis focuses on solving the identified problems using user interface principles, enabled by rule-based systems with wizard-of-oz wherever required. AI-based solutions are needed to achieve the proposed end-to-end systems. For instance, automatically generating the summary of the chatbot conversation to enable Convey, developing accent agnostic speech-to-text engine to support FarmChat, or automatically identifying keywords resulting in breakdown, are unsolved hard AI problems requiring further investigation in future.

Finally, all the user studies in this thesis – for the formative study, Convey, Resilient Chatbots and FarmChat – were conducted in a (semi-)supervised setting with a relatively small sample size. Real-world deployment in unsupervised setting with a larger set of users is crucial to further evaluate the proposed solutions.

Thus, this work opens future work opportunities in a variety of domains, including usability research, voice-based chatbots, and artificial intelligence.

## **7.2 Conclusion**

To conclude, we showcase a modified FarmChat app (Figure 7.1), which combines the best of Convey and Resilient Chatbot learning together to enhance the usability of FarmChat. The chatbot has a Convey window at the top displaying the context of the conversation, which gets dynamically updated as the conversation proceeds. It also provides intuitive interactions on the context values (*e.g.*, updating the inferred intent, leaf or stem, as shown

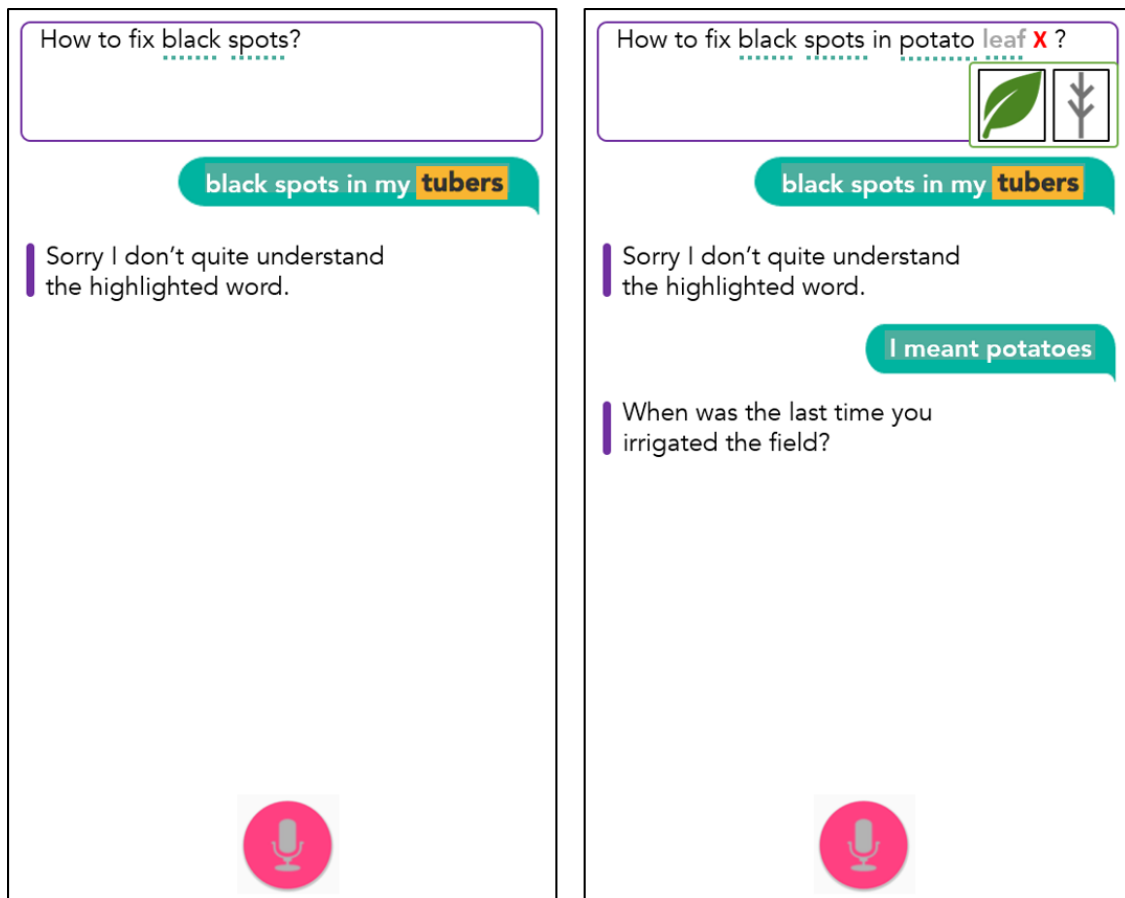


Figure 7.1: Combining learning from the messenger study, Convey, Resilient Chatbot and FarmChat to propose a final chatbot design.

in Figure 7.1), enabling users to modify them in a simple and efficient manner. In case the chatbot is not able to understand a term, it highlights it using the ‘Out-of-vocabulary Explanation’ method (*e.g.*, the term ‘tubers’ in Figure 7.1), and ask the user to rephrase it. Adding these transparency feature has the potential to futher increase the usability and acceptability of FarmChat.

Thus, in this thesis, we showed that the acceptability of chatbots can be increased by making them more transparent and broadening their applicability to new demographics

## BIBLIOGRAPHY

- [1] A.l.i.c.e. foundation website, 2002.
- [2] Mitsuku, 2013.
- [3] Rose, 2013.
- [4] Facebook messenger bots, 2016.
- [5] jabong, 2016.
- [6] Chatbots, 2017.
- [7] Facebook messenger alterra, 2017.
- [8] Facebook messenger call of duty, 2017.
- [9] Facebook messenger chatshopper, 2017.
- [10] Facebook messenger cnn, 2017.
- [11] Facebook messenger hi poncho, 2017.
- [12] Facebook messenger pandorabots, 2017.
- [13] Facebook messenger swelly, 2017.
- [14] Facebook messenger trivia blast, 2017.
- [15] Ibm watson conversation, 2017.
- [16] Natural language toolkit, 2017.
- [17] Plantix: Grow smart, 2017.
- [18] Voice driven web apps: Introduction to the web speech api, 2017.
- [19] Wordnet: A lexical database for english, 2017.

- [20] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y. Lim, and Mohan Kankanhalli. Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, CHI '18, pages 582:1–582:18, New York, NY, USA, 2018. ACM.
- [21] Alan Agresti. Categorical data analysis, volume 482. John Wiley & Sons, 2003.
- [22] Applied AI. Epic chatbot / conversational bot failures (2018 update), 2016.
- [23] Ahmed Al Maimani and Anne Roudaut. Frozen suit: designing a changeable stiffness suit and its application to haptic games. In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, pages 2440–2448. ACM, 2017.
- [24] Jodi Aronson. A pragmatic view of thematic analysis. The Qualitative Report, (1):1989–1991, 1994.
- [25] Solon Barocas, Sorelle Friedler, Moritz Hardt, Joshua Kroll, Suresh Venkatasubramanian, and Hanna Wallach. The fat-ml workshop series on fairness, accountability, and transparency in machine learning. 2018.
- [26] John Barrett, Yilan Jiang, and Harry Wang. 360 view: Mobility & the app economy. 2016.
- [27] Bob Batacharia, Daniel Levy, Roberta Catizone, Alexander Krotov, and Yorick Wilks. Converse: a conversational companion. In Machine conversations, pages 205–215. Springer, 1999.
- [28] Christian Becker, Stefan Kopp, and Ipke Wachsmuth. Simulating the emotion dynamics of a multimodal conversational agent. In Tutorial and Research Workshop on Affective Dialogue Systems, pages 154–165. Springer, 2004.
- [29] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. A neural probabilistic language model. Journal of Machine Learning Research, 3(Feb):1137–1155, 2003.
- [30] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the royal statistical society. Series B (Methodological), pages 289–300, 1995.
- [31] Sayantan Bera. Farm distress calls hit record high but many go unanswered, 2014.

- [32] Timothy W. Bickmore and Justine Cassell. Social dialogue with embodied conversational agents. In Advances in natural multimodal dialogue systems, pages 23–54. Springer, 2005.
- [33] Timothy W. Bickmore and Rosalind W. Picard. Towards caring machines. In CHI '04 Extended Abstracts on Human Factors in Computing Systems, CHI EA '04, pages 1489–1492, New York, NY, USA, 2004. ACM.
- [34] Timothy W. Bickmore and Rosalind W. Picard. Establishing and maintaining long-term human-computer relationships. ACM Trans. Comput.-Hum. Interact., 12(2):293–327, June 2005.
- [35] James Bigelow and Amy Poremba. Achilles' ear? inferior human short-term and recognition memory in the auditory modality. PLoS One, (2):1–8, 2014.
- [36] Christopher Blattman, Robert Jensen, and Raul Roman. Assessing the need and potential of community networking for development in rural india special issue: Icts and community networking. The Information Society, 19(5):349–364, 2003.
- [37] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. J. Mach. Learn. Res., 3:993–1022, March 2003.
- [38] Dan Bohus and Alexander I. Rudnicky. Ravenclaw: dialog management using hierarchical task decomposition and an expectation agenda. In INTERSPEECH. ISCA, 2003.
- [39] Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. Biometrika, 39(3/4):324–345, 1952.
- [40] Susan Brennan. Conversation as direct manipulation: An iconoclastic view. The Art of Human-Computer Interface Design, 1990.
- [41] Susan E Brennan. The grounding problem in conversations with and through computers. Social and cognitive approaches to interpersonal communication, pages 201–225, 1998.
- [42] E. Brewer, M. Demmer, B. Du, M. Ho, M. Kam, S. Nedeveschi, J. Pal, R. Patra, S. Surana, and K. Fall. The case for technology in developing regions. Computer, 38(6):25–38, May 2005.
- [43] Bonnie Brinton, Martin Fujiki, Diane Frome Loeb, and Erika Winkler. Development of conversational repair strategies in response to requests for clarification. Journal of Speech, Language, and Hearing Research, 29(1):75–81, 1986.

- [44] Janet E Cahn and Susan E Brennan. A psychological model of grounding and repair in dialog. In Proc. Fall 1999 AAAI Symposium on Psychological Models of Communication in Collaborative Systems, 1999.
- [45] Justine Cassell. Embodied conversational agents. MIT press, 2000.
- [46] Justine Cassell and Kristinn R Thorisson. The power of a nod and a glance: Envelope vs. emotional feedback in animated conversational agents. Applied Artificial Intelligence, 13(4-5):519–538, 1999.
- [47] Joyce Y Chai, Lanbo She, Rui Fang, Spencer Ottarson, Cody Littley, Changsong Liu, and Kenneth Hanson. Collaborative effort towards common ground in situated human-robot dialogue. In Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction, pages 33–40. ACM, 2014.
- [48] Vishavjeet Chaudhary and Gursharan Singh. What is the future of agriculture in india?, 2016.
- [49] Kathleen Chaykowski. More than 11,000 bots are now on facebook messenger, 2016.
- [50] Kuan-Ta Chen, Chen-Chi Wu, Yu-Chun Chang, and Chin-Laung Lei. A crowdsourcable qoe evaluation framework for multimedia content. In Proceedings of the 17th ACM international conference on Multimedia, pages 491–500. ACM, 2009.
- [51] Mei-Ling Chen and Hao-Chuan Wang. How personal experience and technical knowledge affect using conversational agents. In Proceedings of the 23rd International Conference on Intelligent User Interfaces Companion, IUI '18 Companion, pages 53:1–53:2, New York, NY, USA, 2018. ACM.
- [52] Karen G. Cheng, Francisco Ernesto, and Khai N. Truong. Participant and interviewer attitudes toward handheld computers in the context of hiv/aids programs in sub-saharan africa. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '08, pages 763–766, New York, NY, USA, 2008. ACM.
- [53] Sylvain Choisel and Florian Wickelmaier. Evaluation of multichannel reproduced sound: Scaling auditory attributes underlying listener preference. The Journal of the Acoustical Society of America, 121(1):388–400, 2007.
- [54] O' Brien Chris. Facebook messenger chief says platform's 34,000 chatbots are finally improving user experience, 2016.

- [55] Victor Chue, Juan Sebastian Rios, and Donghee Yvette Wohn. A hot bot: Testing effect of temperature on feelings of closeness when interacting with a chatbot. In Proceedings of the 2018 ACM Conference on Supporting Groupwork, pages 136–139. ACM, 2018.
- [56] Herbert H Clark, Susan E Brennan, et al. Grounding in communication. Perspectives on socially shared cognition, 13(1991):127–149, 1991.
- [57] Josh Constine and Sarah Perez. Facebook messenger now allows payments in its 30,000 chat bots, 2016.
- [58] Cooperation and Department of Agriculture Farmers Welfare. Annual report. pages 1–188, 2017.
- [59] Kevin Corti and Alex Gillespie. Co-constructing intersubjectivity with artificial conversational agents: people are more likely to initiate repairs of misunderstandings with agents represented as human. Computers in Human Behavior, 58:431–442, 2016.
- [60] Duncan Cramer and Dennis Laurence Howitt. The Sage dictionary of statistics: a practical resource for students in the social sciences. Sage, 2004.
- [61] Sebastien Cuendet, Indrani Medhi, Kalika Bali, and Edward Cutrell. Videokheti: Making video content accessible to low-literate and novice users. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '13, pages 2833–2842, New York, NY, USA, 2013. ACM.
- [62] Mary Czerwinski, Eric Horvitz, and Susan Wilhite. A diary study of task switching and interruptions. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '04, pages 175–182, New York, NY, USA, 2004. ACM.
- [63] Herbert Aron David. The method of paired comparisons, volume 12. London, 1963.
- [64] Andy Dearden, Paul Matthews, and Haider Rizvi. Kheti: mobile multimedia in an agricultural co-operative. Pers Ubiquit Comput, pages 597–607, 2011.
- [65] Doris M. Dehn and Susanne van Mulken. The impact of animated interface agents: A review of empirical research. Int. J. Hum.-Comput. Stud., 52(1):1–22, January 2000.
- [66] Alan Dix, Tiziana Catarci, Benjamin Habegger, Yannis Ioannidis, Azrina Kamaruddin, Akrivi Katifori, Giorgos Lepouras, Antonella Poggi, and Devina Ramduny-Ellis. Intelligent context-sensitive interactions on desktop and the web. In Proceedings of the International Workshop in Conjunction with AVI 2006 on Context in Advanced Interfaces, CAI '06, pages 23–27, New York, NY, USA, 2006. ACM.

- [67] Julie S. Downs, Mandy B. Holbrook, Steve Sheng, and Lorrie Faith Cranor. Are your participants gaming the system?: Screening mechanical turk workers. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '10, pages 2399–2402, New York, NY, USA, 2010. ACM.
- [68] Pawel Dybala, Michal Ptaszynski, Rafal Rzepka, and Kenji Araki. Evaluating subjective aspects of hci on an example of a non-task oriented conversational system. International journal on artificial intelligence tools, 19(06):819–856, 2010.
- [69] Craig Elimeliah. Why chatbots are replacing apps, 2016.
- [70] Satu Elo and Helvi Kyngäs. The qualitative content analysis process. Journal of advanced nursing, 62(1):107–115, 2008.
- [71] Sara Engelhardt, Emmeli Hansson, and Iolanda Leite. Better faulty than sorry: Investigating social recovery strategies to minimize the impact of failure in human-robot interaction. In 1st Workshop on Conversational Interruptions in Human-Agent Interactions, WCIHAI 2017, Stockholm, Sweden, 27 August 2017, volume 1943, pages 19–27. CEUR-WS, 2017.
- [72] Facebook. Discover, 2017.
- [73] Leah Findlater, Ravin Balakrishnan, and Kentaro Toyama. Comparing semiliterate and illiterate users' ability to transition from audio+text to text-only interaction. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '09, pages 1751–1760, New York, NY, USA, 2009. ACM.
- [74] BJ Fogg and Hsiang Tseng. The elements of computer credibility. In Proceedings of the SIGCHI conference on Human Factors in Computing Systems, pages 80–87. ACM, 1999.
- [75] Rikin Gandhi, Rajesh Veeraraghavan, Kentaro Toyama, and Vanaja Ramprasad. Digital green: Participatory video for agricultural extension. In 2007 International Conference on Information and Communication Technologies and Development, pages 1–10, Dec 2007.
- [76] Dave Gomboc, Steve Solomon, Mark G Core, H Chad Lane, and Michael Van Lent. Design recommendations to support automated explanation and tutoring. Proc. of BRIMS, 2005.
- [77] Department of School Education Literacy Govt of India. Adult education, 2016.

- [78] Ministry of Agriculture Govt of India. Kisan call centre, 2004.
- [79] Ministry of Home Affairs Govt of India. Data on language, 2016.
- [80] Bettina Graf, Maike Krüger, Felix Müller, Alexander Ruhland, and Andrea Zech. Nom-bot: Simplify food tracking. In Proceedings of the 14th International Conference on Mobile and Ubiquitous Multimedia, MUM '15, pages 360–363, New York, NY, USA, 2015. ACM.
- [81] Matt Grech. The current state of chatbots in 2017, 2017.
- [82] A. S. Grover, M. Plauch, E. Barnard, and C. Kuun. Hiv health information access using spoken dialogue systems: Touchtone vs. speech. In 2009 International Conference on Information and Communication Technologies and Development (ICTD), pages 95–107, April 2009.
- [83] David Gunning. Explainable artificial intelligence (xai). Defense Advanced Research Projects Agency (DARPA), nd Web, 2017.
- [84] Joakim Gustafson, Johan Boye, Morgan Fredriksson, Lasse Johanneson, and Jürgen Königsmann. Providing computer game characters with conversational abilities. In International Workshop on Intelligent Virtual Agents, pages 37–51. Springer, 2005.
- [85] Christopher W Hart, James L Heskett, and Jr WE Sasser. The profitable art of service recovery. Harvard business review, 68(4):148–156, 1990.
- [86] Sandra G Hart and Lowell E Staveland. Development of nasa-tlx (task load index): Results of empirical and theoretical research. Advances in psychology, 52:139–183, 1988.
- [87] Mikko Hartikainen, Esa-Pekka Salonen, and Markku Turunen. Subjective evaluation of spoken dialogue systems using ser vqual method. In INTERSPEECH, 2004.
- [88] Mariam Hassib, Daniel Buschek, Paweł W. Wozniak, and Florian Alt. Heartchat: Heart rate augmented mobile chat to support empathy and awareness. In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, CHI '17, pages 2239–2251, New York, NY, USA, 2017. ACM.
- [89] Jennifer Hill, W. Randolph Ford, and Ingrid G. Farreras. Real conversations with artificial intelligence: A comparison between human-human online conversations and human-chatbot conversations. Computers in Human Behavior, 49:245–250, 2015.

- [90] Orange Hive. First time bot users deserve good bots, 2017.
- [91] Kate S Hone and Robert Graham. Towards a tool for the subjective assessment of speech system interfaces (sassi). Natural Language Engineering, 6(3&4):287–303, 2000.
- [92] Eric Horvitz. Principles of mixed-initiative user interfaces. In Proceedings of the SIGCHI conference on Human Factors in Computing Systems, pages 159–166. ACM, 1999.
- [93] Tianran Hu, Anbang Xu, Zhe Liu, Quanzeng You, Yufan Guo, Vibha Sinha, Jiebo Luo, and Rama Akkiraju. Touch your heart: A tone-aware chatbot for customer care on social media. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, CHI '18, pages 415:1–415:12, New York, NY, USA, 2018. ACM.
- [94] Ting-Hao K Huang, Walter S Lasecki, Amos Azaria, and Jeffrey P Bigham. “is there anything else i can help you with?”: Challenges in deploying an on-demand crowd-powered conversational agent. In Proceedings of the Fourth AAAI Conference on Human Computation and Crowdsourcing, HCOMP '16, 2016.
- [95] Ting-Hao (Kenneth) Huang, Joseph Chee Chang, and Jeffrey P. Bigham. Evorus: A crowd-powered conversational assistant built to automate itself over time. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '18, New York, NY, USA, 2018. ACM.
- [96] Jason L Hutchens. How to pass the turing test by cheating. School of Electrical, Electronic and Computer Engineering research report TR97-05. Perth: University of Western Australia, 1996.
- [97] IANS. Nearly 70 percent of indian farms are very small, census shows, 2015.
- [98] IBM. Watson ai assistant, 2018.
- [99] M. Igarria and J. Iivari. The effects of self-efficacy on computer usage. Omega, 23(6):587 – 605, 1995.
- [100] Facebook Inc. Facebook messenger platform, 2018.
- [101] BI Intelligence. Messaging apps are now bigger than social networks, 2016.
- [102] Ido A Iurgel and Manuel Ziegler. Ask&answer: An educational game where it pays to endear your capricious virtual companion. In International Workshop on Intelligent Virtual Agents, pages 15–24. Springer, 2005.

- [103] Joris H. Janssen. A three-component framework for empathic technologies to augment human interaction. Journal on Multimodal User Interfaces, 6(3):143–161, Nov 2012.
- [104] Jacqueline J. Goodnow Jerome S. Bruner and George A. Austin. A study of thinking. Science Editions, 1967.
- [105] Jiepu Jiang, Ahmed Hassan Awadallah, Rosie Jones, Umut Ozertem, Imed Zitouni, Ranjitha Gurunath Kulkarni, and Omar Zia Khan. Automatic online evaluation of intelligent assistants. In Proceedings of the 24th International Conference on World Wide Web, WWW '15, pages 506–516, Republic and Canton of Geneva, Switzerland, 2015. International World Wide Web Conferences Steering Committee.
- [106] Candace Kamm, Marilyn A Walker, and Diane Litman. Evaluating spoken language systems. In Proc. of AVIOS. Citeseer, 1999.
- [107] Manolya Kavakli, Manning Li, and Tarashankar Rudra. Towards the development of a virtual counselor to tackle students' exam stress. Journal of Integrated Design and Process Science, 16(1):5–26, 2012.
- [108] Konstantinos Kazakos, Siddhartha Asthana, Madeline Balaam, Mona Duggal, Amey Holden, Limalemla Jamir, Nanda Kishore Kannuri, Saurabh Kumar, Amarendar Reddy Manindla, Subhashini Arcot Manikam, GVS Murthy, Papreen Nahar, Peter Phillimore, Shreyaswi Sathyanath, Pushpendra Singh, Meenu Singh, Pete Wright, Deepika Yadav, and Patrick Olivier. A real-time ivr platform for community radio. In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, CHI '16, pages 343–354, New York, NY, USA, 2016. ACM.
- [109] Vinod Khanal. Farmers in a fix over kisan call centre's tech language, 2015.
- [110] Hendrik Knoche and Jeffrey Huang. Text is not the enemy-how illiterates use their mobile phones. In NUIs for new worlds: new interaction forms and interfaces for mobile applications in developing countries-CHI 2012 workshop, 2012.
- [111] Hendrik Knoche, Pr Sheshagiri Rao, HS Jamadagni, and Jeffrey Huang. Actions and advice in coli: A mobile social network to support agricultural peer learning. In Proceedings of the 17th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct, MobileHCI '15, pages 1191–1198, New York, NY, USA, 2015. ACM.
- [112] Ilker Koksals. These are the most important chatbot metrics to track, 2017.

- [113] David A Kolb. Experiential learning: Experience as the source of learning and development. FT press, 2014.
- [114] Stefan Kopp, Lars Gesellensetter, Nicole C. Krämer, and Ipke Wachsmuth. Lecture notes in computer science. chapter A Conversational Agent As Museum Guide: Design and Evaluation of a Real-world Application, pages 329–343. Springer-Verlag, London, UK, UK, 2005.
- [115] Brij Kothari, Joe Takeda, Ashok Joshi, and Avinash Pandey. Same language subtitling: a butterfly for literacy? International Journal of Lifelong Education, 21(1):55–66, 2002.
- [116] Sonya S Kwak, Yunkyung Kim, Eunho Kim, Christine Shin, and Kwangsu Cho. What makes people empathize with an emotional robot?: The impact of agency and physical embodiment on human empathy for a robot. In RO-MAN, 2013 IEEE, pages 180–185. IEEE, 2013.
- [117] Nancy Larson-Powers and Rose Marie Pangborn. Paired comparison and time-intensity measurements of the sensory properties of beverages and gelatins containing sucrose or synthetic sweeteners. Journal of Food Science, 43(1):41–46, 1978.
- [118] Joonhwan Lee, Soojin Jun, Jodi Forlizzi, and Scott E. Hudson. Using kinetic typography to convey emotion in text-based interpersonal communication. In Proceedings of the 6th Conference on Designing Interactive Systems, DIS '06, pages 41–49, New York, NY, USA, 2006. ACM.
- [119] Min Kyung Lee, Sara Kiesler, and Jodi Forlizzi. Receptionist or information kiosk: how do people talk with a robot? In Proceedings of the 2010 ACM conference on Computer supported cooperative work, pages 31–40. ACM, 2010.
- [120] Min Kyung Lee, Sara Kiesler, Jodi Forlizzi, Siddhartha Srinivasa, and Paul Rybski. Gracefully mitigating breakdowns in robotic services. In Human-Robot Interaction (HRI), 2010 5th ACM/IEEE International Conference on, pages 203–210. IEEE, 2010.
- [121] Yeoreum Lee, Jae-eul Bae, Sona S Kwak, and Myung-Suk Kim. The effect of politeness strategy on human-robot collaborative interaction on malfunction of robot vacuum cleaner. In RSS Workshop on HRI, 2011.
- [122] Linda H Lewis and Carol J Williams. Experiential learning: Past and present. New directions for adult and continuing education, 1994(62):5–16, 1994.
- [123] Q. Vera Liao, Matthew Davis, Werner Geyer, Michael Muller, and N. Sadat Shami. What can you do?: Studying social-agent orientation and agent proactive interactions

- with an agent for employees. In Proceedings of the 2016 ACM Conference on Designing Interactive Systems, DIS '16, pages 264–275, New York, NY, USA, 2016. ACM.
- [124] Vera Q. Liao, Muhammed Masud Hussain, Praveen Chandar, Matthew Davis, Marco Crasso, Dakuo Wang, Michael Muller, Sadat N. Shami, and Werner Geyer. All work and no play? conversations with a question-and-answer chatbot in the wild. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, CHI '18, New York, NY, USA, 2018. ACM.
- [125] J. C. R. Licklider. IRE Transactions on Human Factors in Electronics, HFE-1:4–11, March 1960.
- [126] Diane Litman and Shimei Pan. Empirically evaluating an adaptable spoken dialogue system. In User Modeling, pages 55–64. Springer, 1999.
- [127] Ewa Luger and Abigail Sellen. “like having a really bad pa”: The gulf between user expectation and experience of conversational agents. In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, CHI '16, pages 5286–5297, New York, NY, USA, 2016. ACM.
- [128] Carlos Garcia Martin. Factors that might affect empathy towards robots. 2014.
- [129] Michael Martin. Verstehen: The uses of understanding in the social sciences. Routledge, 2018.
- [130] Indrani Medhi, Somani Patnaik, Emma Brunskill, SN Gautama, William Thies, and Kentaro Toyama. Designing mobile interfaces for novice and low-literacy users. ACM Transactions on Computer-Human Interaction (TOCHI), 18(1):2, 2011.
- [131] Indrani Medhi, Archana Prasad, and Kentaro Toyama. Optimal audio-visual representations for illiterate users of computers. In Proceedings of the 16th International Conference on World Wide Web, WWW '07, pages 873–882, New York, NY, USA, 2007. ACM.
- [132] Indrani Medhi, Aman Sagar, and Kentaro Toyama. Text-free user interfaces for illiterate and semiliterate users. Inf. Technol. Int. Dev., 4(1):37–50, October 2007.
- [133] Indrani Medhi-Thies, Pedro Ferreira, Nakull Gupta, Jacki O’Neill, and Edward Cutrell. Krishipustak: A social networking system for low-literate farmers. In Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing, CSCW '15, pages 1670–1681, New York, NY, USA, 2015. ACM.

- [134] Rob van der Meulen. Gartner says 6.4 billion connected 'things' will be in use in 2016, up 30 percent from 2015, 2016.
- [135] Preeti Mudliar, Jonathan Donner, and William Thies. Emergent practices around cgnnet swara, voice forum for citizen journalism in rural india. In Proceedings of the Fifth International Conference on Information and Communication Technologies and Development, ICTD '12, pages 159–168, New York, NY, USA, 2012. ACM.
- [136] Bilge Mutlu, Allison Terrell, and Chien-Ming Huang. Coordination mechanisms in human-robot collaboration. In Proceedings of the Workshop on Collaborative Manipulation, 8th ACM/IEEE International Conference on Human-Robot Interaction, pages 1–6. Citeseer, 2013.
- [137] Chelsea Myers, Anushay Furqan, Jessica Nebolsky, Karina Caro, and Jichen Zhu. Patterns for how users overcome obstacles in voice user interfaces. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, page 6. ACM, 2018.
- [138] Andrew Ng. Woebot: Ai for mental health. 2017.
- [139] Andreea I Niculescu, Kheng Hui Yeo, Luis F D'Haro, Seokhwan Kim, Ridong Jiang, and Rafael E Banchs. Design and evaluation of a conversational agent for the touristic domain. In Asia-Pacific Signal and Information Processing Association, 2014 Annual Summit and Conference (APSIPA), pages 1–10. IEEE, 2014.
- [140] Donald A. Norman. The Design of Everyday Things. Basic Books, Inc., New York, NY, USA, 2002.
- [141] Joseph D Novak. Meaningful learning: The essential factor for conceptual change in limited or inappropriate propositional hierarchies leading to empowerment of learners. Science education, 86(4):548–571, 2002.
- [142] US Dept of Agriculture. India's agricultural exports climb to record high, 2014.
- [143] Government of India. mkisan, 2017.
- [144] Press Trust of India. Agriculture's share in gdp declines to 13.7
- [145] Press Trust of India. Govt to prepare roadmap for doubling farmers income by 2022, 2017.

- [146] Amy Ogan, Samantha Finkelstein, Elijah Mayfield, Claudia D'Adamo, Noboru Matsuda, and Justine Cassell. "oh dear stacy!": Social interaction, elaboration, and learning with teachable agents. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '12, pages 39–48, New York, NY, USA, 2012. ACM.
- [147] Hiroaki Ogata and Yoneo Yano. Context-aware support for computer-supported ubiquitous learning. In Proceedings of the 2nd IEEE International Workshop on Wireless and Mobile Technologies in Education (WMTE'04), WMTE '04, pages 27–, Washington, DC, USA, 2004. IEEE Computer Society.
- [148] Kweku Opoku-Agyemang, Bhaumik Shah, and Tapan S. Parikh. Scaling up peer education with farmers in india. In Proceedings of the Ninth International Conference on Information and Communication Technologies and Development, ICTD '17, pages 15:1–15:10, New York, NY, USA, 2017. ACM.
- [149] Tim Paek. Empirical methods for evaluating dialog systems. In Proceedings of the workshop on Evaluation for Language and Dialogue Systems-Volume 9, page 2. Association for Computational Linguistics, 2001.
- [150] Tim Paek and Eric Horvitz. Uncertainty, utility, and misunderstanding: A decision-theoretic perspective on grounding in conversational systems. In AAAI Fall Symposium on Psychological Models of Communication, North, 1999.
- [151] Tim Paek and Eric Horvitz. Grounding criterion: Toward a formal theory of grounding. Technical report, MSR Technical Report, 2000.
- [152] Ana Paiva, Iolanda Leite, Hana Boukricha, and Ipke Wachsmuth. Empathy in virtual agents and robots: A survey. ACM Trans. Interact. Intell. Syst., 7(3):11:1–11:40, September 2017.
- [153] Neil Patel, Sheetal Agarwal, Nitendra Rajput, Amit Nanavati, Paresh Dave, and Tapan S Parikh. A comparative study of speech and dialed input voice interfaces in rural india. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pages 51–54. ACM, 2009.
- [154] Neil Patel, Deepti Chittamuru, Anupam Jain, Paresh Dave, and Tapan S. Parikh. Aavaaj otalo: A field study of an interactive voice forum for small farmers in rural india. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '10, pages 733–742, New York, NY, USA, 2010. ACM.
- [155] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In EMNLP, volume 14, pages 1532–43, 2014.

- [156] Dennis Perzanowski, Alan C Schultz, William Adams, Elaine Marsh, and Magda Bugajska. Building a multimodal human-robot interface. IEEE intelligent systems, 16(1):16–21, 2001.
- [157] Roberto Pieraccini, Krishna Dayanidhi, Jonathan Bloom, Jean-Gui Dahan, Michael Phillips, Bryan R. Goodman, and K. Venkatesh Prasad. Multimodal conversational systems for automobiles. Commun. ACM, 47(1):47–49, January 2004.
- [158] Martin Porcheron, Joel E Fischer, Stuart Reeves, and Sarah Sharples. Voice interfaces in everyday life. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, page 640. ACM, 2018.
- [159] Filip Radlinski and Nick Craswell. A theoretical framework for conversational search. In Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval, pages 117–126. ACM, 2017.
- [160] Byron Reeves and Clifford Nass. The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places. Cambridge University Press, New York, NY, USA, 1996.
- [161] MIT Technology Review. 10 breakthrough technologies, 2016.
- [162] MIT Technology Review. The biggest technology failures of 2016, 2016.
- [163] Waleed Riaz, Haris Durrani, Suleman Shahid, and Agha Ali Raza. Ict intervention for agriculture development: Designing an ivr system for farmers in pakistan. In Proceedings of the Ninth International Conference on Information and Communication Technologies and Development, ICTD '17, pages 33:1–33:5, New York, NY, USA, 2017. ACM.
- [164] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. “why should i trust you?”: Explaining the predictions of any classifier. In Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16, pages 1135–1144, New York, NY, USA, 2016. ACM.
- [165] L. D. Riek, T. Rabinowitch, B. Chakrabarti, and P. Robinson. Empathizing with robots: Fellow feeling along the anthropomorphic spectrum. In 2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, pages 1–6, Sept 2009.
- [166] Torsten Ringberg, Gaby Odekerken-Schröder, and Glenn L Christensen. A cultural models approach to service recovery. Journal of Marketing, 71(3):194–214, 2007.

- [167] Albert Rizzo, Belinda Lange, John G Buckwalter, Eric Forbell, Julia Kim, Kenji Sagae, Josh Williams, JoAnn Difede, Barbara O Rothbaum, Greg Reger, et al. Simcoach: an intelligent virtual human system for providing healthcare information and support. International Journal on Disability and Human Development, 10(4):277–281, 2011.
- [168] Susan Robinson, Antonio Roque, and David R. Traum. Dialogues in context: An objective user-oriented evaluation approach for virtual human dialogue. In 7th International Conference on Language Resources and Evaluation (LREC), Valletta, Malta, May 19–21, 2010 2010.
- [169] Susan Robinson, David R. Traum, Midhun Ittycheriah, and Joe Henderer. What would you ask a conversational agent? observations of human-agent dialogues in a museum setting. In Language Resources and Evaluation Conference (LREC), Marrakech (Morocco), May 2008 2008.
- [170] Ronald Rosenfeld, Dan Olsen, and Alex Rudnicky. Universal speech interfaces. interactions, 8(6):34–44, October 2001.
- [171] Astrid Marieke Rosenthal-von der Pütten, Frank P. Schulte, Sabrina C. Eimler, Laura Hoffmann, Sabrina Sobieraj, Stefan Maderwald, Nicole C. Krämer, and Matthias Brand. Neural correlates of empathy towards robots. In Proceedings of the 8th ACM/IEEE International Conference on Human-robot Interaction, HRI '13, pages 215–216, Piscataway, NJ, USA, 2013. IEEE Press.
- [172] Katri Saarikivi. The rise of empathy-enabling technology. 2017.
- [173] Tommy Sandbank, Michal Shmueli-Scheuer, Jonathan Herzig, David Konopnicki, John Richards, and David Piorkowski. Detecting egregious conversations between customers and virtual agents. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), volume 1, pages 1802–1811, 2018.
- [174] Ari-Heikki Sarjanoja, Minna Isomursu, Pekka Isomursu, and Jonna Häkkinä. Integrating collaborative context information with social media: A study of user perceptions. In Proceedings of the 25th Australian Computer-Human Interaction Conference: Augmentation, Application, Innovation, Collaboration, OzCHI '13, pages 437–446, New York, NY, USA, 2013. ACM.
- [175] Ayse Pinar Saygin, Ilyas Cicekli, and Varol Akman. Turing test: 50 years later. In The Turing Test, pages 23–78. Springer, 2003.

- [176] Emanuel A Schegloff, Gail Jefferson, and Harvey Sacks. The preference for self-correction in the organization of repair in conversation. Language, 53(2):361–382, 1977.
- [177] Robert P Schumaker, Ying Liu, Mark Ginsburg, and Hsinchun Chen. Evaluating mass knowledge acquisition using the alice chatterbot: The az-alice dialog system. International journal of human-computer studies, 64(11):1132–1140, 2006.
- [178] Alex Sciuto, Arnita Saini, Jodi Forlizzi, and Jason I. Hong. ”hey alexa, what’s up?”: A mixed-methods studies of in-home conversational agent usage. In Proceedings of the 2018 Designing Interactive Systems Conference, DIS ’18, pages 857–868, New York, NY, USA, 2018. ACM.
- [179] Phil Scott, Hilary Asoko, and J Leach. Student conceptions and conceptual learning. Handbook of research on science education, pages 31–56, 2007.
- [180] Stela H Seo, Denise Geiskkovitch, Masayuki Nakane, Corey King, and James E Young. Poor thing! would you feel sorry for a simulated robot?: A comparison of empathy toward a physical and a simulated robot. In Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction, pages 125–132. ACM, 2015.
- [181] Alexander Serenko. A model of user adoption of interface agents for email notification. Interact. Comput., 20(4-5):461–472, September 2008.
- [182] Marcos Serrano, Anne Roudaut, and Pourang Irani. Visual composition of graphical elements on non-rectangular displays. In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, pages 4405–4416. ACM, 2017.
- [183] Ameneh Shamekhi, Q. Vera Liao, Dakuo Wang, Rachel K. E. Bellamy, and Thomas Erickson. Face value? exploring the effects of embodiment for a group facilitation agent. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI ’18, New York, NY, USA, 2018. ACM.
- [184] Bayan Abu Shawar and Eric Atwell. A comparison between alice and elizabeth chatbot systems. 2002.
- [185] Bayan Abu Shawar and Eric Atwell. Chatbots: are they really useful? In Ldv Forum, volume 22, pages 29–49, 2007.
- [186] J. Sherwani, S. Palijo, S. Mirza, T. Ahmed, N. Ali, and R. Rosenfeld. Speech vs. touch-tone: Telephony interfaces for information access by low literate users. In 2009 International Conference on Information and Communication Technologies and Development (ICTD), pages 447–457, April 2009.

- [187] Sujit Shinde, Divya Piplani, Karthik Srinivasan, Dineshkumar Singh, Rahul Sharma, and Preetam Mohnaty. mkrishi: Simplification of ivr based services for rural community. In Proceedings of the India HCI 2014 Conference on Human Computer Interaction, IndiaHCI '14, pages 154:154–154:159, New York, NY, USA, 2014. ACM.
- [188] Ben Shneiderman. Designing the user interface: strategies for effective human-computer interaction. Pearson Education India, 2010.
- [189] Ben Shneiderman and Pattie Maes. Direct manipulation vs. interface agents. interactions, 4(6):42–61, 1997.
- [190] Ben Shneiderman, Catherine Plaisant, Maxine Cohen, Steven Jacobs, Niklas Elmqvist, and Nicholas Diakopoulos. Grand challenges for hci researchers. interactions, 23(5):24–25, August 2016.
- [191] K Alan Shore. Quantum computing for computer scientists, ns yanofsky and ma manucci.
- [192] Amy K Smith, Ruth N Bolton, and Janet Wagner. A model of customer satisfaction with service encounters involving failure and recovery. Journal of marketing research, pages 356–372, 1999.
- [193] Vasant Srinivasan and Leila Takayama. Help me please: Robot politeness strategies for soliciting help from humans. In Proceedings of the 2016 CHI conference on human factors in computing systems, pages 4945–4955. ACM, 2016.
- [194] Statista. Most popular global mobile messenger apps as of january 2017, 2017.
- [195] Simone Stumpf, Vidya Rajaram, Lida Li, Margaret Burnett, Thomas Dietterich, Erin Sullivan, Russell Drummond, and Jonathan Herlocker. Toward harnessing user feedback for machine learning. In Proceedings of the 12th international conference on Intelligent user interfaces, pages 82–91. ACM, 2007.
- [196] N. Suzuki, K. Ishii, and M. Okada. Talking eye: autonomous creature as accomplice for human. In Proceedings. 3rd Asia Pacific Computer Human Interaction (Cat. No.98EX110), pages 409–414, Jul 1998.
- [197] Julian Szymanski, Tomasz Sarnatowicz, and Wlodzislaw Duch. Towards avatars with artificial minds: Role of semantic memory. Journal of Ubiquitous Computing and Intelligence, 2007.
- [198] Didem Tali. India’s rural farmers struggle to read and write, 2015.

- [199] Indrani M Thies, Nandita Menon, Sneha Magapu, Manisha Subramony, and Jacki O’Neill. How do you want your chatbot? an exploratory wizard-of-oz study with young, urban indians. In Proceedings of the International Conference on Human-Computer Interaction (HCI), INTERACT ’17. IFIP, 2017.
- [200] Robert Thornberg. Informed grounded theory. Scandinavian Journal of Educational Research, 56(3):243–259, 2012.
- [201] Statistics Times. Sector-wise contribution of gdp of india, 2017.
- [202] David Traum and Jeff Rickel. Embodied agents for multi-party dialogue in immersive virtual worlds. In Proceedings of the first international joint conference on Autonomous agents and multiagent systems: part 2, pages 766–773. ACM, 2002.
- [203] David R Traum. Computational models of grounding in collaborative systems. In Psychological Models of Communication in Collaborative Systems-Papers from the AAI Fall Symposium, pages 124–131, 1999.
- [204] Heather Turner, David Firth, et al. Bradley-terry models in r: the bradleyterry2 package. Journal of Statistical Software, 48(9), 2012.
- [205] Francisco AM Valério, Tatiane G Guimarães, Raquel O Prates, and Heloisa Candello. Here’s what i can do: Chatbots’ strategies to convey their features to users. In Proceedings of the XVI Brazilian Symposium on Human Factors in Computing Systems, page 28. ACM, 2017.
- [206] Oriol Vinyals and Quoc Le. A neural conversational model. arXiv preprint arXiv:1506.05869, 2015.
- [207] Luis von Ahn and Laura Dabbish. Designing games with a purpose. Commun. ACM, 51(8):58–67, August 2008.
- [208] Marilyn A. Walker, John S. Aberdeen, Julie E. Boland, Elizabeth Owen Bratt, John S. Garofolo, Lynette Hirschman, Audrey N. Le, Sungbok Lee, Shrikanth S. Narayanan, Kishore Papineni, Bryan L. Pellom, Joseph Polifroni, Alexandros Potamianos, P. Prabhu, Alexander I. Rudnicky, Gregory A. Sanders, Stephanie Seneff, David Stallard, and Steve Whittaker. Darpa communicator dialog travel planning systems: the june 2000 data collection. In INTER\_SPEECH, 2001.
- [209] Marilyn A Walker, Diane J Litman, Candace A Kamm, and Alicia Abella. Paradise: A framework for evaluating spoken dialogue agents. In Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics, pages 271–280. Association for Computational Linguistics, 1997.

- [210] Eric W Weisstein. Bonferroni correction. 2004.
- [211] Joseph Weizenbaum. Eliza - a computer program for the study of natural language communication between man and machine. Communications of the ACM, 9(1):36–45, 1966.
- [212] Tsung-Hsien Wen, Milica Gasic, Nikola Mrksic, Lina Maria Rojas-Barahona, Pei-Hao Su, Stefan Ultes, David Vandyke, and Steve J. Young. A network-based end-to-end trainable task-oriented dialogue system. CoRR, abs/1604.04562, 2016.
- [213] Tsung-Hsien Wen, Milica Gasic, Nikola Mrksic, Pei-hao Su, David Vandyke, and Steve J. Young. Semantically conditioned lstm-based natural language generation for spoken dialogue systems. CoRR, abs/1508.01745, 2015.
- [214] Wikipedia contributors. Concept learning — Wikipedia, the free encyclopedia, 2018. [Online; accessed 5-October-2018].
- [215] Yorick Wilks. Close Engagements with Artificial Companions: Key Social, Psychological, Ethical, and Design Issues. John Benjamins Publishing Company, Amsterdam, 2010.
- [216] Jason D Williams, Nopal B Niraula, Pradeep Dasigi, Aparna Lakshmiratan, Carlos Garcia, Jurado Suarez, Mouni Reddy, and Geoff Zweig. Rapidly scaling dialog systems with interactive learning. 2015.
- [217] Peter Wright and John McCarthy. Empathy and experience in hci. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '08, pages 637–646, New York, NY, USA, 2008. ACM.
- [218] Anbang Xu, Zhe Liu, Yufan Guo, Vibha Sinha, and Rama Akkiraju. A new chatbot for customer service on social media. In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, CHI '17, pages 3506–3510, New York, NY, USA, 2017. ACM.
- [219] Holly A Yanco, Jill L Drury, et al. Classifying human-robot interaction: an updated taxonomy. In SMC (3), pages 2841–2846, 2004.
- [220] Rayoung Yang and Mark W. Newman. Learning from a learning thermostat: Lessons for intelligent systems for the home. In Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing, UbiComp '13, pages 93–102, New York, NY, USA, 2013. ACM.

- [221] Dani Yaniv. Dynamics of creativity and empathy in role reversal: Contributions from neuroscience. Review of General Psychology, 16(1):70, 2012.
- [222] Mariya Yao. 5 metrics every chatbot developer needs to track, 2016.
- [223] Tom Yeh, Tsung-Hsiang Chang, Bo Xie, Greg Walsh, Ivan Watkins, Krist Wongsuphasawat, Man Huang, Larry S. Davis, and Benjamin B. Bederson. Creating contextual help for guis using screenshots. In Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology, UIST '11, pages 145–154, New York, NY, USA, 2011. ACM.
- [224] Steve Young. A review of large-vocabulary continuous-speech. IEEE Signal Processing Magazine, 13(5):45–, Sept 1996.
- [225] Zhou Yu, Leah Nicolich-Henkin, Alan W Black, and Alexander Rudnicky. A wizard-of-oz study on a non-task-oriented dialog systems that reacts to user engagement. In Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 55–63, 2016.
- [226] Zhou Yu, Ziyu Xu, Alan W Black, and Alexander Rudnicky. Strategy and policy learning for non-task-oriented conversational systems. In Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 404–412, 2016.

## VITA

Mohit Jain is a Senior Researcher in the Technologies for Emerging Markets Group at Microsoft Research India. Prior to that, he was a Senior Research Engineer in the Cognitive IoT team at IBM Research India. Along with, he was pursuing a PhD in Computer Science & Engineering from the University of Washington, working with Dr. Shwetak Patel in the Ubiquitous Computing lab. Before that, he graduated with a Masters in Computer Science from the University of Toronto working with Dr. Ravin Balakrishnan. His PhD work is at the intersection of HCI and chatbots, focusing on the intelligibility, accessibility and security of chatbots. His research primarily explores the introduction of emerging technologies to new populations, by developing ubiquitous computing systems, designing interaction techniques, conducting usability evaluations, and collecting and analyzing data. He has worked in the domain of Health, Sustainability, Education and HCI4D (HCI for Development). Over the course of 10+ years of his research work, he has published 25+ research papers at top-tier computer science journals and conferences, filed 15+ patents, received several awards (including 1 best paper award, 1 best paper nomination, a Wolfond Fellowship, an Aegis Graham Bell award, invited to attend the Heidelberg Laureate Forum, 8 IBM Accomplishment awards, 2 IBM Outstanding Technical Achievement awards, won DFS Tech Chatbots Challenge 2017 and Microsoft Imagine Cup 2009), and contributed to multiple client deals with his work at IBM Research.

To get in touch, please email: [mohitj@uw.edu](mailto:mohitj@uw.edu), [mohja@microsoft.com](mailto:mohja@microsoft.com) or [mohitrbl@gmail.com](mailto:mohitrbl@gmail.com).