Evaluating and Informing the Design of Chatbots

Mohit Jain[†*], **Pratyush Kumar**[†], **Ramachandra Kota**[†‡], **Shwetak N. Patel**[*]
[†]IBM Research, India. mohitjain@in.ibm.com, pratyushkpanda@gmail.com
[‡]Realtor.com, Vancouver, Canada. ramachandra.kota@move.com
[*]Computer Science & Engineering, University of Washington, Seattle, USA. shwetak@cs.washington.edu

ABSTRACT
Text messaging-based conversational agents (CAs), popularly called *chatbots*, received significant attention in the last two years. 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. Hence, understanding the usage patterns of first-time users can potentially inform and guide the design of future chatbots. In this paper, we report the findings of 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 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. We conclude with implications to evolve the design of chatbots, such as: clarify chatbot capabilities, sustain conversation context, handle dialog failures, and end conversations gracefully.

ACM Classification Keywords
H.5.2. Information Interfaces and Presentation (e.g. HCI): User Interfaces

Author Keywords
Chatbot; Conversational Agent; Messenger; Evaluation.

INTRODUCTION
Today, personal smart devices come pre-loaded with Conversational Agents (CAs) such as Siri, Google Assistant, and Alexa. To the companies designing these CAs, and more importantly to the users interacting with them, the shared anthropomorphic goal is clear: talking to a CA should feel like talking to a fellow human [17, 33]. Based on the interaction modality, CAs can be categorized into speech-based CA (*e.g.*, Siri, Alexa), text-messaging based CA (*e.g.*, Google Assistant, Messenger bots), and multimodal CA. The first CA was a text-messaging based agent, called ELIZA emerging in 1966 [44]. Such turn-based, messaging-based CAs are popularly called *chatbots*. Chatbots received significant attention beginning in 2016 [18], with the idea that users can 'text' intelligent agents of businesses, just as they text their friends and family using their mobile devices. Technology companies have raced to deploy platforms for developing such chatbots with built-in natural-language capabilities (such as Facebook Messenger, IBM Watson Conversation, and api.ai). 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 [23] – for varied use-cases ranging from pizza ordering (Domino's) to shopping (Burberry), from connecting like-minded humans (Chatible) to flight booking (Kayak), 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 messaging interface [21].

In spite of this 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 [25]. Hence it is crucial to understand the interaction pattern of first-time chatbot users to inform and guide the design of future bots. While the HCI community has studied how conversational agents are used in different settings [27, 28, 30, 31, 33, 42], none of them focuses on first-time chatbot users. 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.

Towards this goal, in this paper, 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 [18] and is the second-most popular text-messaging app [40]. 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. Our findings indicate that the participants preferred chatbots which provided either a 'human-like' natural language conversation ability, or an engaging experience that exploited 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.

DIS 2018, June 9–13, 2018, Hong Kong.

http://dx.doi.org/10.1145/3196709.3196735

RELATED WORK

The term 'Conversation Agent' has come to mean a wide variety of systems with varying capabilities and purposes, with the underlying assumption that the agent participates in a human-machine dialog. Licklider's 'Man-machine symbiosis' [32] was one of the earliest discourses from an HCI perspective that visualized humans interacting with machines in a natural manner. Research in conversation agents started with messaging-based chatbots, whose purpose 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.

The first chatbot emerged in 1966 from MIT, called ELIZA [44], which emulated a Rogerian psychotherapist. ELIZA worked on simple declarative rules: if a certain keyword was identified in the user text, it responded with one or more predefined 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 [15, 38, 45, 47]. The notable follow-up chatbots included MegaHAL [26], ALICE [1], and Elizabeth [39]. 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.

In the last decade, conversational agents started focusing more on *utility*, with the goal of accomplishing specific task(s). Anthropomorphism, when it exists, seeks to augment the efficiency of the task-solving process. Nowadays, conversational agents range across several modalities, including speech (such as Siri, Alexa, Cortana), text-messaging (such as Domino's, CNN, Pandorabots, Burberry, *etc.* found on Messenger, Slack, and/or Skype platform), and as multimodal embodied agents. Embodied CAs have a graphical front-end as opposed to a robotic body, and attempt to be human-like by employing non-verbal behaviors, such as gestures and expressions, in addition to speech [13, 29]. Embodied agents are yet to reach the wider population. On the other hand, the ease of development, familiarity of use, and the privacy afforded by purely messaging-based agents has ensured that most development efforts have centered on building conversational agents with no provision for gestures or speech. This paper focuses on such text-messaging based CAs, called chatbots.

Evaluating Conversational Agents

Recent works [28, 30, 31, 33, 42] evaluated CAs from an HCI perspective. Luger and Sellen [33] 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 CAs turned out to be enabling hands-free interactions that save time [33]. One of their central findings is participants complaining about the inaccuracies in speech-to-text conversion. Similarly, Jiang *et al.* [28] evaluated different tasks in Cortana, a speech-based CA, and found major issues with the quality of speech recognition and agent's intent classification.

In contrast to the speech-modality of the studied CAs, chatbots have text-modality with very different user expectations and interaction patterns, which is the focus of this paper. Thies *et al.* [42] 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 [42]. 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.* [30, 31] 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.

Our work differs significantly from the above as 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 [28, 33] studies were 'regular' users of speech-based CA, participants in Liao *et al.* [30, 31] studies were new office joining employees, and participants in Thies *et al.* [42] 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. While Liao *et al.* [31] is the closest to our work; they study a single CA - an HR bot installed on a company-wide IM tool. In contrast, we study 8 different chatbots across domains on the Messenger platform, and exclusively work with first-time chatbot users.

In summary, the literature on CAs has largely focused on the AI and natural language challenges. The few studies from an HCI perspective primarily focus on speech-based CAs, and do not evaluate the experience of using the currently available text-based chatbots in the industry. Particularly, there is no existing study that attempts to characterize the first-time user experience of interacting with chatbots. With the recent prolific deployment of chatbots, it becomes imperative to conduct a study to understand this experience.

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.

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 [40] 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 [19]. 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 1.

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.

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* [5], *Call of Duty* [6], *chatShopper* [7], *CNN* [8], *Hi Poncho* [9], *Pandorabots* [10], *Swelly* [11], *Trivia Blast* [12]) 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 [28, 36], 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 [24]. All the interviews were conducted in either the participants' office or home. The interviews were voice-recorded after receiving permission from the participants, 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.

QUANTITATIVE DATA ANALYSIS

In this section, we present the quantitative analysis of the chat logs. The analysis shows that the participants were engaged

Table 1: Description of the selected eight chatbots for the user study.

Chatbot	**Domain**	**Description** (in their own words, or *from https://chatbottle.co)
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 & 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.

throughout the study, as hinted by their high level of participation. In total, participants and chatbots exchanged 9968 messages interacting for 25 hours (Figure 1a, 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 [43]. 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.

Total Interaction Time

Over the course of the study, the time spent by a participant interacting with the chatbots was 93.5±53.9 mins, and the number of sessions was 23.7±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). This hints that the participants spent similar amounts of time with each of the chatbot as the study facilitator instructed. Figure 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 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 [41]. 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 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 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 2a) and participants' messages (Figure 2b). *Pandorabots* lacked interactive elements (Figure 2) as it is a completely text-based chatbot, while *Trivia Blast* (Figure 2a) was predominantly click-based. In Messenger, human input was limited to text messages or button clicks (Figure 2b). As no key loggers were used, only the button presses that resulted in input text (e.g., Figure 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 5, 'auto-suggestion buttons') or buttons that lead to external websites (*e.g.*, Figure 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

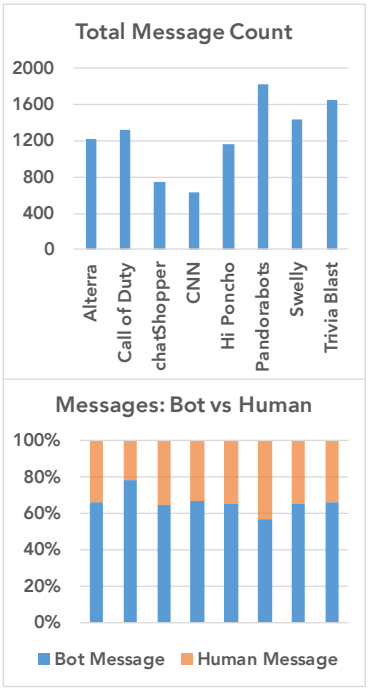

Figure 1: (a) *Top*: Total chat messages exchanged; (b) *Bottom*: Division of those messages among chatbots and participants.

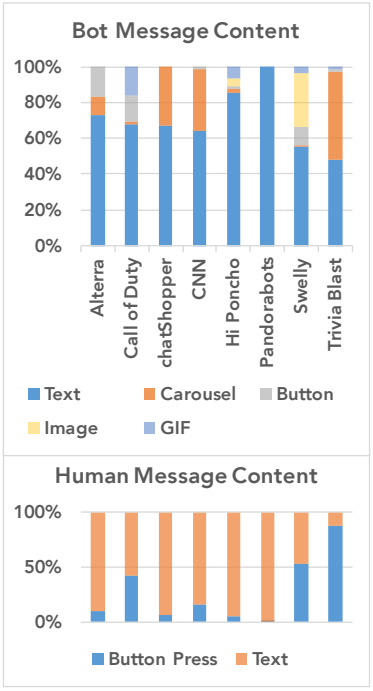

Figure 2: (a) *Top*: Content of the chatbot messages; (b) *Bottom*: Content of the participant (human) messages.

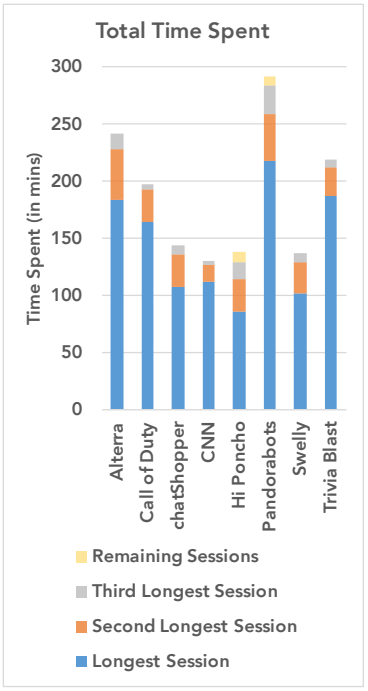

Figure 3: Interaction Time spent by the participant per chatbot.

than *Call of Duty*, *Swelly* and *Trivia Blast* (all $p<0.01$). This indicates that *Alterra* (23.5±19.5 characters/message) and *CNN* (21.7±26.4) needed longer text input as in a search query, while *Call of Duty* (13±10.1), *Swelly* (9.2±7.6) and *Trivia Blast* (12.7±9.9) were predominantly click-based.

Chatbot messages comprise of several interactive and rich media elements (Figure 2a). 70.1% of messages comprises of only text. The character count per chatbot message response was highest for *CNN* (84.3±60.2), closely followed by *Hi Poncho* (79.4±55.1) and *Call of Duty* (73.8±38.1), and lowest for *Pandorabots* (41.9±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 4), constituting of 14.5% messages (Figure 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 4). Carousels were extensively used to show news in *CNN*, quiz questions in *Trivia Blast*, and shopping items in *chatShopper*.

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.

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 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*

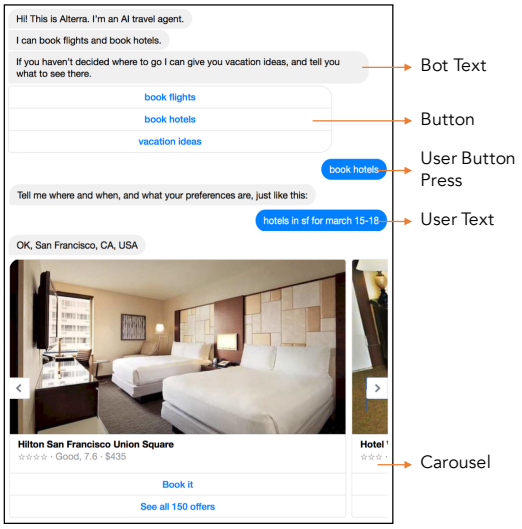

Figure 4: *Alterra* chatbot showing different UI elements

Figure 5: *CNN* chatbot showing different UI elements

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*". However, participants mentioned that flight-booking websites are better than *Alterra*, as "*I can quickly browse through hundreds of flights*" - $P_{2,Alterra}$. 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_{15} 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_1 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?*" - $P_{3,Call\ of\ Duty}$. This is consistent with Norman's theory of "human error" [34]. 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*"-$P_{13,CNN}$. 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.

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*" - $P_{9,Trivia\ Blast}$. 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 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*" - $P_{9,Pandorabots}$. "*It answers like my spouse*" - $P_{1,Pandorabots}$. Five participants tried to "*break the* (intelligence of) *Pandorabots*", similar to [42], which might be one of the reasons for maximal interaction with *Pandorabots* (Figure 1a, 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*" - $P_{5,chatShopper}$.

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_9 asked *chatShopper* for "*eye-liner*" and it responded with socks instead. P_9 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_8 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_1 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_{10} 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*" - $P_{14,\text{chatShopper}}$. 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_3 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 [14], 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_{12} appreciated questions asked by *Alterra* to refine the flight search, while P_5 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_5.

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.

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 [35]. 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*" - $P_{15,\text{Call of Duty}}$. A few participants expected the chatbots to be more personal. "*She was not addressing me by my name... very impersonal.*" - $P_{15,\text{Pandorabots}}$. 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_6 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_{15} 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 [30, 31], 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?*" - $P_{4,\text{Pandorabots}}$. 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.*" - $P_{9,\text{Call of Duty}}$. 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.

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 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 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 5). P_5 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 [33], 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 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_1 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. P5 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 [42] 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_{10} 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_{15}, "*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.

Chatbot Ratings

Participants were asked to rate each chatbot on six different metrics on a 5-point Likert scale (Figure 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±0.8) and *Pandorabots* (4.2±1.3) were rated significantly higher than *Call of Duty* (2.7±1.3), *CNN* (2.4±0.9) and *Swelly* (2.7±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±1.2. Also *Call of Duty* (3.7±1.2) and *Swelly* (3.7±0.9) were found to be the most *frustrating*, and were significantly worse than *Pandorabots* (2.4±1.3) and *Trivia Blast* (2.5±0.9), with p<0.01. Regarding *using the bot in future*, the general opinion was unfavorable, still *Pandorabots* (3.5±1.4) was voted higher than *Call of Duty* (1.8±1.3) and *Swelly* (1.8±1.1) (p<0.01). In *Overall* rating, *Pandorabots* (3.9±1.2) was rated significantly higher than *Call of Duty* (2.4±1.3), *CNN* (2.5±1.1) and *Swelly* (2.5±1.1) (p<0.01).

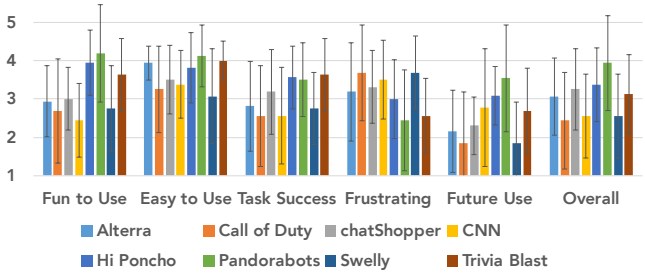

Figure 6: Likert-scale rating by the participants (with standard deviation shown by error bars)

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±1.5). *Pandorabots* (ranking=3.5±2.7) was a close second with 11, and *Trivia Blast* (ranking=3.7±2.3) was third with 8. The worst ranked was *CNN* (ranking=6.3±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 conservation 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.

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 language 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 [33], 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 [28, 33] and chatbots [30, 31, 42]. Existing work emphasizes the speech modality with features such as ease of hands-free interaction [33] and inaccuracies in speech-to-text conversion [28, 33]. 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 [30, 31, 42] 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.

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.* [33] 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 [42], 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 [33]. 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 [16]. 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 [37]) 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 [30, 31, 33, 42, 46], 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*).

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 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 [33], 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 [22].

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, *UReport Global* is a civil reporting chatbot, and *Swelly* crowdsources votes on A-vs-B questions. Logging data as part of HCI-relevant diary studies [20] 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.

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.

CONCLUSION

We define chatbots as text-based, turn-based, task-fulfilling programs, embedded within existing platforms. Our study, involving 16 participants interacting with 8 pre-selected chatbots for the first-time, over three days, spanning almost 10,000 messages, revealed that expectations of 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. The comments for the high-rated chatbots provided directions for improvements. Clarifying a chatbot's capabilities, supporting context resolution for dialog efficiency, managing dialogue failures, engaging in small talk, and ending conversation gracefully, are some of the guidelines for chatbot designers. We expect the results from our work to inform and guide the design of future chatbots.

REFERENCES

1. 2002. A.L.I.C.E. Foundation website. (2002). Retrieved January 4, 2017 from http://alicebot.org
2. 2013. Mitsuku. (2013). Retrieved January 4, 2017 from http://www.mitsuku.com
3. 2013. Rose. (2013). Retrieved January 4, 2017 from http://brilligunderstanding.com/rosedemo.html
4. 2016. Facebook Messenger bots. (2016). Retrieved Dec 1, 2016 from https://chatbottle.co/bots/messenger
5. 2017. Facebook Messenger Alterra. (2017). Retrieved January 30, 2017 from https://www.messenger.com/t/alterra.cc
6. 2017. Facebook Messenger Call of Duty. (2017). Retrieved January 30, 2017 from https://www.messenger.com/t/CallofDuty
7. 2017. Facebook Messenger chatShopper. (2017). Retrieved January 30, 2017 from https://www.messenger.com/t/chatShopper
8. 2017. Facebook Messenger CNN. (2017). Retrieved January 30, 2017 from https://www.messenger.com/t/cnn
9. 2017. Facebook Messenger Hi Poncho. (2017). Retrieved January 30, 2017 from https://www.messenger.com/t/hiponcho
10. 2017. Facebook Messenger Pandorabots. (2017). Retrieved January 30, 2017 from https://www.messenger.com/t/chatbots.io
11. 2017. Facebook Messenger Swelly. (2017). Retrieved January 30, 2017 from https://www.messenger.com/t/swell.bot
12. 2017. Facebook Messenger Trivia Blast. (2017). Retrieved January 30, 2017 from https://www.messenger.com/t/triviablast1
13. Timothy W. Bickmore and Justine Cassell. 2005. Social dialongue with embodied conversational agents. In *Advances in natural multimodal dialogue systems*. Springer, 23–54.
14. Timothy W. Bickmore and Rosalind W. Picard. 2005. Establishing and Maintaining Long-term Human-computer Relationships. *ACM Trans. Comput.-Hum. Interact.* 12, 2 (June 2005), 293–327. DOI: http://dx.doi.org/10.1145/1067860.1067867
15. Dan Bohus and Alexander I. Rudnicky. 2003. Ravenclaw: dialog management using hierarchical task decomposition and an expectation agenda.. In *INTERSPEECH*. ISCA.
16. Susan Brennan. 1990. Conversation as Direct Manipulation: An Iconoclastic View. *The Art of Human-Computer Interface Design* (1990).
17. Justine Cassell. 2000. *Embodied conversational agents*. MIT press.
18. Kathleen Chaykowski. 2016. More Than 11,000 Bots Are Now On Facebook Messenger. (2016). Retrieved Dec 28, 2016 from http://www.forbes.com/sites/kathleenchaykowski/2016/07/01/more-than-11000-bots-are-now-on-facebook-messenger/
19. O' Brien Chris. 2016. Facebook Messenger chief says platform's 34,000 chatbots are finally improving user experience. (2016). Retrieved February 7, 2017 from http://venturebeat.com/2016/11/11/facebook-messenger-chief-says-platforms-34000/-chatbots-are-finally-improving-user-experience/
20. Mary Czerwinski, Eric Horvitz, and Susan Wilhite. 2004. A Diary Study of Task Switching and Interruptions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '04)*. ACM, New York, NY, USA, 175–182. DOI: http://dx.doi.org/10.1145/985692.985715
21. Craig Elimeliah. 2016. Why chatbots are replacing apps. (2016). Retrieved January 20, 2017 from http://venturebeat.com/2016/08/02/why-chatbots-are-replacing-apps/
22. Facebook. 2017. Discover. (2017). Retrieved May 31, 2017 from https://developers.facebook.com/docs/messenger-platform/discover
23. Matt Grech. 2017. The Current State of Chatbots in 2017. (2017). Retrieved Jan 5, 2018 from https://getvoip.com/blog/2017/04/21/the-current-state-of-chatbots-in-2017/
24. Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Advances in psychology* 52 (1988), 139–183.
25. Orange Hive. 2017. First time bot users deserve good bots. (2017). Retrieved Jan 5, 2018 from https://unfiltered.orangehive.de/first-time-bot-users-deserve-good-bots/
26. Jason L Hutchens. 1996. 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).
27. Mohit Jain, Ramachandra Kota, Pratyush Kumar, and Shwetak Patel. 2018. Convey: Exploring the Use of a Context View for Chatbots. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, 6.
28. Jiepu Jiang, Ahmed Hassan Awadallah, Rosie Jones, Umut Ozertem, Imed Zitouni, Ranjitha Gurunath Kulkarni, and Omar Zia Khan. 2015. Automatic Online Evaluation of Intelligent Assistants. In *Proceedings of the 24th International Conference on World Wide Web (WWW '15)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 506–516. DOI: http://dx.doi.org/10.1145/2736277.2741669

29. Stefan Kopp, Lars Gesellensetter, Nicole C. Krämer, and Ipke Wachsmuth. 2005. Lecture Notes in Computer Science. Springer-Verlag, London, UK, UK, Chapter A Conversational Agent As Museum Guide: Design and Evaluation of a Real-world Application, 329–343.
30. Q. Vera Liao, Matthew Davis, Werner Geyer, Michael Muller, and N. Sadat Shami. 2016. 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)*. ACM, New York, NY, USA, 264–275. DOI:http://dx.doi.org/10.1145/2901790.2901842
31. Vera Q. Liao, Muhammed Masud Hussain, Praveen Chandar, Matthew Davis, Marco Crasso, Dakuo Wang, Michael Muller, Sadat N. Shami, and Werner Geyer. 2018. 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)*. ACM, New York, NY, USA, 13.
32. J. C. R. Licklider. 1960. *IRE Transactions on Human Factors in Electronics* HFE-1 (March 1960), 4–11.
33. Ewa Luger and Abigail Sellen. 2016. "Like Having a Really Bad PA": The Gulf Between User Expectation and Experience of Conversational Agents. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 5286–5297.
34. Donald A. Norman. 2002. *The Design of Everyday Things*. Basic Books, Inc., New York, NY, USA.
35. Amy Ogan, Samantha Finkelstein, Elijah Mayfield, Claudia D'Adamo, Noboru Matsuda, and Justine Cassell. 2012. "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)*. ACM, New York, NY, USA, 39–48. DOI:http://dx.doi.org/10.1145/2207676.2207684
36. Susan Robinson, Antonio Roque, and David R. Traum. 2010. 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. http://people.ict.usc.edu/~traum/Papers/Robinson-LREC2010.pdf
37. Susan Robinson, David R. Traum, Midhun Ittycheriah, and Joe Henderer. 2008. 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). http://people.ict.usc.edu/~traum/Papers/Blackwell-LREC08.pdf
38. Ronald Rosenfeld, Dan Olsen, and Alex Rudnicky. 2001. Universal Speech Interfaces. *interactions* 8, 6 (Oct. 2001), 34–44. DOI:http://dx.doi.org/10.1145/384076.384085
39. Bayan Abu Shawar and Eric Atwell. 2002. A comparison between ALICE and Elizabeth chatbot systems. (2002).
40. Statista. 2017. Most popular global mobile messenger apps as of January 2017. (2017). Retrieved February 7, 2017 from https://www.statista.com/statistics/258749/most-popular-global-mobile-messenger-apps/
41. N. Suzuki, K. Ishii, and M. Okada. 1998. Talking Eye: autonomous creature as accomplice for human. In *Proceedings. 3rd Asia Pacific Computer Human Interaction (Cat. No.98EX110)*. 409–414. DOI: http://dx.doi.org/10.1109/APCHI.1998.704479
42. Indrani M Thies, Nandita Menon, Sneha Magapu, Manisha Subramony, and Jacki O'Neill. 2017. 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, 20.
43. 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. 2001. DARPA communicator dialog travel planning systems: the june 2000 data collection. In *INTERSPEECH*.
44. Joseph Weizenbaum. 1966. ELIZA - A computer program for the study of natural language communication between man and machine. *Commun. ACM* 9, 1 (1966), 36–45.
45. Tsung-Hsien Wen, Milica Gasic, Nikola Mrksic, Pei-hao Su, David Vandyke, and Steve J. Young. 2015. Semantically Conditioned LSTM-based Natural Language Generation for Spoken Dialogue Systems. *CoRR* abs/1508.01745 (2015). http://arxiv.org/abs/1508.01745
46. Yorick Wilks. 2010. *Close Engagements with Artificial Companions: Key Social, Psychological, Ethical, and Design Issues*. John Benjamins Publishing Company, Amsterdam.
47. Steve Young. 1996. A review of large-vocabulary continuous-speech. *IEEE Signal Processing Magazine* 13, 5 (Sept 1996), 45–. DOI: http://dx.doi.org/10.1109/79.536824