ChatrEx: Designing Explainable Chatbot Interfaces for Enhancing Usefulness, Transparency, and Trust

by

Anjali Khurana

B.Tech., PEC University of Technology, 2018

Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of
Master of Science

in the
School of Computing Science
Faculty of Applied Sciences

SIMON FRASER UNIVERSITY
Summer 2021

Declaration of Committee

Name:	**Anjali Khurana**
Degree:	**Master of Science (Computing Science)**
Thesis title:	**ChatrEx: Designing Explainable Chatbot Interfaces for Enhancing Usefulness, Transparency, and Trust**
Committee:	**Chair:** Greg Baker University Lecturer, Computing Science
	Parmit Chilana Supervisor Assistant Professor, Computing Science
	Sheelagh Carpendale Committee Member Professor, Computing Science
	Saba Alimadadi Examiner Assistant Professor, Computing Science

Ethics Statement

The author, whose name appears on the title page of this work, has obtained, for the research described in this work, either:

a. human research ethics approval from the Simon Fraser University Office of Research Ethics

or

b. advance approval of the animal care protocol from the University Animal Care Committee of Simon Fraser University

or has conducted the research

c. as a co-investigator, collaborator, or research assistant in a research project approved in advance.

A copy of the approval letter has been filed with the Theses Office of the University Library at the time of submission of this thesis or project.

The original application for approval and letter of approval are filed with the relevant offices. Inquiries may be directed to those authorities.

Simon Fraser University Library
Burnaby, British Columbia, Canada

Update Spring 2016

Abstract

When breakdowns occur during a human-chatbot conversation, the lack of transparency and the "black-box" nature of task-oriented chatbots can make it difficult for end users to understand what went wrong and why. Inspired by recent HCI research on explainable AI solutions, we explored the design space of explainable chatbot interfaces through ChatrEx. We followed the iterative design and prototyping approach and designed two novel in-application chatbot interfaces (ChatrEx-VINC and ChatrEx-VST) that provide visual example-based step-by-step explanations about the underlying working of a chatbot during a breakdown. ChatrEx-VINC provides visual example-based step-by-step explanations in-context of the chat window whereas ChatrEx-VST provides explanations as a visual tour overlaid on the application interface. Our formative study with 11 participants elicited informal user feedback to help us iterate on our design ideas at each of the design and ideation phases and we implemented our final designs as web-based interactive chatbots for complex spreadsheet tasks. We conducted an observational study with 14 participants to compare our designs with current state-of-the-art chatbot interfaces and assessed their strengths and weaknesses. We found that visual explanations in both ChatrEx-VINC and ChatrEx-VST enhanced users' understanding of the reasons for a conversational breakdown and improved users' perceptions of usefulness, transparency, and trust. We identify several opportunities for future HCI research to exploit explainable chatbot interfaces and better support human-chatbot interaction.

Keywords: chatbots; visual explanations; in-app help; conversational breakdowns; human-chatbot interaction; Human-AI interaction

Statement of Contributions

This thesis includes ideas and content (including figures and tables) that have been accepted but not yet published, in a peer-reviewed publication for which I was the lead author. The conference paper from which I have adapted content is the following:

Khurana, A., Harjandi, P., Chilana, P. (2021) ChatrEx: Designing Explainable Chatbot Interfaces for Enhancing Usefulness, Transparency, and Trust. Proceedings of the IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC 21) *[to appear]*

Dedication

It is my genuine gratefulness and warmest regard that I dedicate this work to my parents, *Lovely Khurana and Joginder Pal Khurana* whose love, encouragement, and blessings have driven me to work hard persistently and strengthen me to achieve such success and honor. I am always thankful to them for inspiring me every day to not only dream, but also work diligently to fulfill them.

Especially my mom, I am grateful to her for always believing in me and being my backbone, for it wouldn't have been possible without her constant support.

To my kid brother, *Rohit Khurana*, for always being supportive and cheering me up.

Acknowledgements

Above all, I would like to express my sincere gratitude to my supervisor, Dr. Parmit Chilana, for her continued support, guidance, motivation, and encouragement. I am grateful to her for providing me this opportunity and placing faith in me, without which none of this research would have been possible.

I would also like to thank Dr. Sheelagh Carpendale for her valuable feedback on this research- it has been a privilege to learn from her and being a part of Ixlab. Thank you to my thesis committee member, Dr. Saba Alimadadi, for her important insights to refine this thesis.

I am thankful for all the support and help given by my labmates- Rimika, Parsa, David, Laton, Amir, Narges, Foroozan, and other fellow Ixlab members. Thanks for providing your advice and making this experience so fun and lively.

Finally, I would like to thank my family- my parents and brother for their unconditional love and always being there for me, and my friends- Khushwant, Nimarta for always cheering me up and all your discussions which have been my stressbusters.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

The current era is experiencing growing interest [8] in conversational agents or chatbots which are Artificial Intelligence (AI) software designed to have conversations with or act as a personal assistant to humans. The rise of chatbots has fundamentally automated standard task completion where most tasks are now accomplished through an interaction between people and conversational agents, such as Siri (Apple, 2011), Cortana (Microsoft, 2015), Google Now (2012), Alexa (Amazon, 2015). The spike in the potential of these chatbots has spurred us to believe that these spoken dialogue interfaces are a critical way to reach customers [9]. As recognized by the major technology companies such as Google, Facebook, and Microsoft, chatbots are the next popular technology [8] and future gateways to many key services [42].

With the transition of the interfaces from websites and apps to AI systems such as bots at present [9], the conversation has become a new mode of Human-Computer Interaction (HCI) [42]. Virtual assistants and chatbots are increasingly being used to automatically recognize and respond to end users' needs and automate complex tasks in a variety of different contexts [26, 32] using AI. For example, customer service bots are being used to reduce operating costs in many industries [24, 32] while conversational assistants, such as Siri and Alexa, are available to millions of end users across various devices to help them complete personal tasks [26, 44]. Despite the promise of virtual assistants, in fact, many ends up being completely abandoned by users after their initial interaction and lack of perceived success [3, 30, 32]. In particular, the class of virtual assistants that have yet to reach mass adoption are chatbots embedded inside software applications to support usage of application features.

In-application virtual assistants and task-oriented chatbots embedded inside software applications offer several opportunities to automate various tasks and support the use of complex application features. But, despite the promise of these chatbots, many users feel

annoyed and even abandon these assistants after repeated unsuccessful interactions [64]. For example, *Clippy* was introduced in the Microsoft Office suite as early as 1996 [45] to assist users in performing various word processing tasks, only to be removed four years later based on negative user feedback.

In the not-too-distant future, it is expected that 85% of human interactions will be handled through chatbots [26]. However, the success of the chatbots not only depends on how good the software is but also on their interface, thus the research in AI has taken a human-centered approach. There are many reasons why users may give up on using a chatbot. For example, there is a fine line between providing help to users while not interrupting or annoying them, which remains a challenge that has to be overcome [64].

Recent progress in machine learning (ML) and Natural Language Processing (NLP) has contributed to improving chatbot functionality manyfold at the underlying algorithmic level. However, complexities of natural language interactions [3, 51] and limited training sets and poor conversational understanding [2] remain to be key obstacles in fully realizing the potential of human-chatbot interaction. For example, when interacting with task-oriented chatbots, a key challenge for users is dealing with conversational dead-ends or breakdowns [5, 37, 36]. In fact, during a conversational breakdown, as many as 70% of users may opt to quit the task or completely abandon the chatbot, while others may try to rephrase their queries with little or no success [51].

A breakdown usually occurs when a chatbot fails to understand the user's intent in a query [42] and the user does not know what to do next. In fact, the chatbot often appears as a "black-box" to the user, making it difficult to understand why something did not work, what actions are actually possible, and how to recover from the breakdown. From a user interaction perspective, another major concern influencing the adoption of chatbots is the lack of transparency that is inherent in human-chatbot interaction [42, 65]. This lack of transparency, in turn, impacts the users' perceptions of usefulness and trust in the system [65, 25, 49] especially when the chatbot makes inexplicable errors. To tackle transparency and trust concerns in AI-based systems, recent research has recognized the need to incorporate explainability features or explanations, giving rise to a new class of Explainable AI(XAI) solutions [10, 25, 38]. However, XAI design solutions are yet to be fully explored in task-oriented chatbot products [38].

In this thesis, we explore the design of in-application task-oriented chatbots that can explain the underlying steps of a task and where and why they failed during a conversational breakdown. We take inspiration from recent research in XAI which recognizes the need to incorporate explainability features or explanations for improving transparency and trust [38, 10, 25]. The goal of our approach was not only to acknowledge the occurrence of a

breakdown (as has been explored in recent work [5]) but also to design novel mechanisms that can enhance user understanding of what caused the breakdown and where exactly the breakdown occurred. Our overarching goal was: **how can we design an in-application task-oriented chatbot that can explain the underlying steps of a task and where and why it failed?**

We propose a novel class of explainable chatbot interfaces (ChatrEx) that visually explain a chatbot's high-level operations and causes of a breakdown. We explore two variations of ChatrEx that either provide visual explanations in-context of the chatbot (*ChatrEx-VINC*, Figure 4.2), or as a visual tour overlaid on the application interface (*ChatrEx-VST*, Figure 4.5). We followed the *research through design* approach [67] to iteratively design these chatbot interfaces across different stages of ideation and prototyping: we built low-fidelity prototypes in the form of paper mock-ups *(Stage 1: exploratory design stage)*; followed by image-based PowerPoint mock-ups *(Stage 2: detailed prototypes)*; and finally progressed to Axure-based medium-fidelity prototypes*(Stage 3: Partially interactive prototypes)*. We also used informal user feedback with 11 participants to help us iterate on design choices for our final prototypes. Finally, we implemented the design of these chatbot interfaces as an add-on for Google Sheets, an online spreadsheet application.

To evaluate our two explainable chatbot designs (ChatrEx-VINC and ChatrEx-VST), we compared them to an existing explanation design based on keyword highlighting [5] and a baseline chatbot that provided no explanations. We conducted an observational usability study with 14 participants and assessed their perceptions of usefulness, transparency, and trust across these four chatbots. Overall, we found that there was a significant difference in how participants ranked each of the chatbot designs–in particular, ChatrEx-VST and ChatrEx-VINC were consistently ranked higher across all of our key measures. The visual example-based explanations made the chatbot's functionality and decisions more transparent and useful, and in turn, improved users' perceived trust in these chatbots.

Our main contributions in this thesis are: (1) the design and implementation of two novel in-application task-oriented chatbot interfaces that provide visual example-based explanations to illustrate the underlying working of a chatbot and helps users recognize the causes of a breakdown; (2) empirical insights into the strengths and weaknesses of these explainable chatbot interfaces in terms of usefulness, transparency, and trust during the situation of breakdown.

This thesis is structured as follows: Chapter 2 outlines the related work on the user perceptions of task-oriented chatbots, design and evaluation of in-application chatbots, as well as explainable AI systems so far explored in this domain. We discuss the struggles that users have faced with these task-oriented and in-application assistants, the recent studies

addressing them using explainable AI, and how our work differs from them. Chapter 3 explains the literature review that informed design requirements for in-application chatbot interfaces, culminating them into key design goals. Chapter 4 provides details on the system design and implementation of the two variations of ChatrEx: ChatrEx-VST and ChatrEx-VINC. Chapter 5 describes the evaluation of ChatrEx web-based chatbot interfaces and their comparison with the state-of-the-art chatbots. This chapter also sheds light on the key findings of user perception of these chatbots in terms of usefulness, transparency, and trust. In Chapter 6 and 7, we reflect on the limitations of the approaches in the aforementioned chatbot interfaces and discuss other implications from this research and potential future work.

Chapter 2

Related Work

In this chapter, we provide a survey of existing research in the field of Human-AI interaction with a particular focus on designing explainable solutions for task-oriented chatbots. We build upon these works in our thesis and our research is also contextualized by the wide range of prior literature on user perceptions of task-oriented chatbots and explainable AI systems.

2.1 User perceptions of task-oriented chatbots

Since the emergence of the first chatbot called ELIZA [62] in 1966, significant research has been taking place in this area. However, there have not been many solutions adopted by the industry. The AI industry focuses on the 'task-based interactive bots' that can allow users to accomplish the tasks using them [30] seamlessly and efficiently. Despite the growing industry efforts and the advancements in AI, there has not been much evolution in the user interfaces of these chatbots [30].

While many recognize the necessity to incorporate explainability features in these AI systems (XAI), how to address the real-world user needs for understanding AI remains an open question [38]. To make an explainable AI effective, one should know how the users understand as they may have their mental model on the capabilities and trustworthiness of the system [58]. There can be a lot at stake if a user's mental model overestimates the intelligence of a system or even underestimate the user control over a system[58].

Previous user studies with task-oriented chatbots have contributed insights into how users perceive human-chatbot interaction and some of the struggles that they face. For example, Luger and Sellen [42] pointed out various limitations such as trust issues and lack of more meaningful system feedback that users faced during human-chatbot interaction

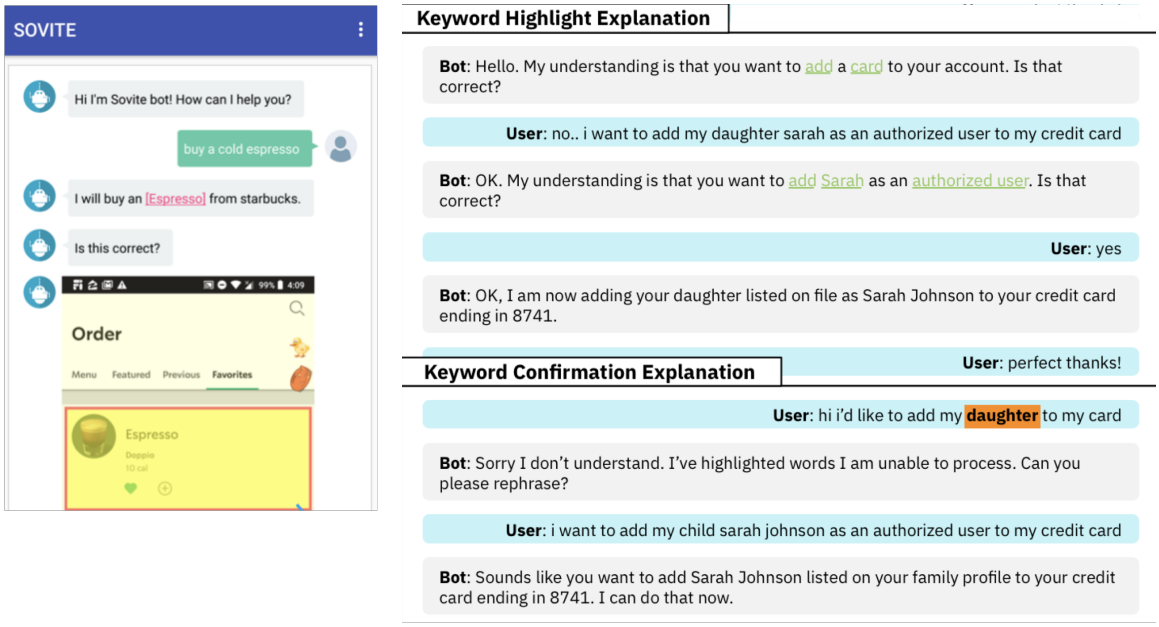


Figure 2.1: (a) SOVITE (System for Optimizing Voice Interfaces to Tackle Errors)[37] (b) Keyword highlighting-based breakdown repair strategy [5]

with intelligent assistants such as Siri and Cortana. Further, they highlighted a gap between user expectation and system operation because users found it difficult to understand the capability of the chatbot and how the chatbot could actually accomplish a task.

Another study [49] raised concerns with the lack of effective system status and transparency among the chatbots. They addressed how chatbots, such as Alexa, were a "black box" for users when they faced an error or a breakdown. Consequently, being unaware of the system status and capabilities, users were more likely to lose trust and less likely to continue using these chatbots after experiencing a breakdown, especially when engaged in complex tasks [42]. Another factor affecting the users' trust and desire to use chatbots includes the conversational breakdowns [42]. Typically, there are two situations when the breakdown occurs within the chatbot: (i) Disambiguation queries when the chatbot misunderstand the user's intent, (ii) Infeasible or out-of-domain queries when the chatbot is incapable of accomplishing the task [51]. Most of the recent work has focused on providing solutions for the former case while the proposed study intends to cover the latter tasks as well [31].

The introduction of explainability features can be a solution to make black-box ML systems more transparent to the user [10, 25, 38]. However, in terms of conveying a task-oriented chatbot's understanding during a breakdown, only a few examples exist. Recently Li et. al [37] explored multi-modal strategies in the context of the existing mobile app Graphical

User Interface (GUIs) for fixing Natural Language Understanding (NLU) breakdowns and command disambiguations [36]. In particular, one of their system solution (Figure 2.1.a) i.e., SOVITE (System for Optimizing Voice Interfaces to Tackle Errors) system allowed the users to discover the conversational breakdowns, identify their causes and finally fix these errors. Although these solutions focused more on supporting interactive repair strategies during conversational breakdowns, they demonstrated an effective use of app GUIs to help with grounding.

Our work goes further to address the gap of improving users' perception of transparency and trust for chatbots that are embedded in feature-rich applications, such as spreadsheets. Our novel ChatrEx designs that can visually explain the underlying working of a chatbot using the UI components as referents, allow users to learn about the chatbot's competencies and limitations even if they are not familiar with the application functionality.

2.2 Design and evaluation of in-application chatbots

In-application task-oriented chatbots embedded inside software applications were envisioned to help end users be more efficient with software tasks [22]. However, early versions of these chatbots, unfortunately, saw high rates of user abandonment [22]. Perhaps the most well-known failure has been that of the Office Assistant named "Clippy" which was introduced in the Office suite in November 1996 [45] to assist the user in performing the tasks. Clippy received widespread negative user feedback such that four years later it had to be removed by Microsoft from the later versions. [7, 43, 22].

Since then, there have been many research efforts to advance the work in creating more helpful and efficient automated chatbots in applications. For example, *Calendar.help*, was introduced as a personal assistant to provide fast and efficient scheduling via email, but, ultimately, it was unable to handle a lot of the complex calendaring tasks on its own [18]. The opacity of these systems is known to be a key challenge as users struggle to understand what inputs and outputs are actually possible. More recently, Glass et al. [25] assessed the factors impacting the trust and understandability of Cognitive Assistant that Learns and Organizes (CALO) system, a personalized assistant for office-related tasks, and similarly found that users perceived the system to be too "opaque" and difficult to comprehend. In fact, the lack of transparency was mentioned to be one of the most crucial reasons responsible for affecting trust among users. While some works suggest using explanation-based systems [25] to augment chatbots and make them easier to understand, it is yet to be explored how to structure and design such explanations.

Our paper complements these existing works by designing novel explainable interfaces for in-application task-oriented chatbots that can improve transparency and trust among users.

2.3 Explainable AI (XAI) systems

Almost a few decades ago, the Explainable AI concepts were first introduced within expert systems [20, 14], and since then, XAI has become an extensively growing field for making the ML systems and their decisions comprehensible to the users[28, 40, 55]. Recently, there has been a big push in AI and HCI research to design Explainable AI solutions to make these complex ML algorithms more understandable for end users. [11, 66, 33, 35, 39, 12].

Notably, many of these studies have focused on explaining the underlying algorithms through different explanation methods. Typically, the taxonomy of XAI as shown in the Figure 2.2 specified that these explanation methods comprises four categories: The first explanation method includes *Global explanations* that focus on explaining the entire model using global features. Next, *Local explanations* that explain a prediction or an individual outcome using local features. In contrast, *Inspect Counterfactual explanations* aims to explain the features that influence the change in output or prediction[38]. Recently, the taxonomy included a distinct method named as *Example-based explanations*. These example-based explanations explain the prediction using examples similar to or different from the instance.

However, end users who do not have any knowledge or experience with ML struggle to understand many of these in-depth algorithm-specific explanations [34, 59]. Still, it has been shown that such explanations can play a key role in enhancing transparency and trust for AI systems [33, 35, 39]. To our knowledge, prior work has not explored the potential of XAI design solutions in the context of improving user interaction with in-application task-oriented chatbots, as is the goal of our paper.

The closest work to ours is perhaps the recent work on keyword highlighting [5] as shown in Figure 2.1.b. Their proposed keyword highlight and confirmation explanation design tries to explain the underlying intent of the user's input in a query and highlights parts that the chatbot did and did not understand respectively. Although this level of highlighting was useful as a repair strategy, we argue that for more complex tasks and applications, these highlighting-based explanations are not sufficient enough to explain the underlying working of the chatbot. To provide transparency and more in-depth reasons of the breakdown ("What cause"), it is equally important to allow users the exposure of chatbot's inner workings and provide a window into its competencies and limitations [13]. ChatrEx takes inspiration from these existing works to expand and explore the design space of explainable chatbots

Category of Methods	Explanation Method	Definition
Explain the model (**Global**)	Global feature importance	Describe the weights of features used by the model (including visualization that shows the weights of features)
	Decision tree approximation	Approximate the model to an interpretable decision-tree
	Rule extraction	Approximate the model to a set of rules, e.g., if-then rules
Explain a prediction (**Local**)	Local feature importance and saliency method	Show how features of the instance contribute to the model's prediction (including causes in parts of an image or text)
	Local rules or trees	Describe the rules or a decision-tree path that the instance fits to guarantee the prediction
Inspect counterfactual	Feature influence or relevance method	Show how the prediction changes corresponding to changes of a feature (often in a visualization format)
	Contrastive or counterfactual features	Describe the feature(s) that will change the prediction if perturbed, absent or present
Example based	Prototypical or representative examples	Provide example(s) similar to the instance and with the same record as the prediction
	Counterfactual example	Provide example(s) with small differences from the instance but with a different record from the prediction

Figure 2.2: Taxonomy of Explainable AI (XAI) methods [38]

while contributing novel visual explanation designs for chatbots embedded in feature-rich applications.

Chapter 3

Exploring the design of in-application explainable chatbot interfaces: Motivation and Design Goals

In this thesis, we explore the design space of in-application explainable chatbot interfaces (ChatrEx) that can explain a chatbot's underlying functionality during a breakdown. Our main goal was to improve users' perceptions of transparency, trust, and usefulness when working with in-application chatbots. In this chapter, we will discuss how we derived design requirements and design goals for in-application explainable chatbot interfaces.

3.1 Deriving Design Requirements

Based on the related work and current state-of-the-art in task-oriented chatbots, we investigated two categories as described below: *How to structure the explanation* and *How to enhance the explanation with examples and visuals* for deriving the design requirements.

3.1.1 Structuring the explanation with intent and entity

Recent studies have shown that users' trust can be influenced by the layout and the comprehensibility of the explanations [66]. In addition, explanations that tend to be concise showed potential to augment the adoption of AI systems. Thus, it is imperative to structure the explanation that provides the appropriate and to-the-point information of "What needs to be explained". To perform the action requested in a user's query, a typical task-oriented

chatbot first identifies the *intent* and the *entity*. The intent refers to the final objective of the user's query, while the entity includes the remaining information from the query to add parameters and to make the objective more specific [30, 5, 23]. For example, consider this chatbot query in a spreadsheet application: "Create a graph that shows square root of column C data." While the intent would be *creating a graph*, the entity would be the functions or operations such as square root and data (i.e., Column C). Most of the critical conversational breakdown occurs when the chatbot fails to correctly comprehend the intended meaning of the user's query. In explaining the internal working of a chatbot, it is imperative to structure the explanation such that it provides clear and concise information about the intent and the entity.

A common challenge for XAI solutions is to reconcile the significance of explaining decisions versus competencies of the AI system [29]. Typically, XAI systems are expected to explain the decision process (i.e., reasons for the system's action), especially when the system goes wrong. [4]. While it is helpful for users when a system (i.e., chatbot) acknowledges the decision (i.e., breakdown) [5], it is equally significant to help users comprehend the competencies or capabilities of AI systems [29]. We hypothesize that explaining the competencies and limitations of the chatbot using the identified intent and entity will not only aid users to recognize the breakdown but also improve transparency. Furthermore, within the breakdown decision, an indication of where the problem occurred and its possible causes would help the users more clearly understand the cause of the breakdown and repair their queries [37].

3.1.2 Enhancing the explanation with examples and visuals

Examples have been shown to be effective [54, 53] for explaining AI predictions without overwhelming users with internal algorithmic logic [50, 56, 61]. Particularly, users find *high level* and *simple* explanations to be more useful and easier to interpret [15, 52, 48, 46]. In fact, the XAI Taxonomy recommends the method of "example-based explanations" [38] that provides explanations in the form of normative or comparative examples of the instance [11, 16, 21] of the instance to potentially improve understanding of end users with limited expertise for these algorithms. We draw inspiration from recent studies that have attempted to explore "Example-based explanations" to effectively explain complex concepts [11, 66].

Specifically, the former study [11] investigated two kinds of explanations: Normative and comparative example-based explanations. Normative explanations display the most similar training examples from the target classes for enhancing system understanding. In contrast, comparative explanations highlight similarities or differences between a user's input and the alternative classes as limitations, which can be useful for representing breakdowns related

to disambiguation and infeasibility [11]. When normative and comparative explanations are used to demonstrate the capability and limitations of complex systems, they are found to be more effective in improving users' trust [11, 66].

Another consideration for designing explanations is whether to present them verbally [59] or visually [19]. Recent studies suggest that verbal prompts tend to become "visually unappealing" and "difficult to read" [5] whereas visual explanations increase transparency and users' trust in the automated systems [11, 66, 6]. Next, the internal working of the chatbot processes the user's query in several steps [23], therefore if we explain the working of chatbot analogous to the flow just as ML works in steps, in the form of modular step-by-step explanations, it would be more comprehensible and useful. Finally, chatbots that only appear when called upon can be less intrusive [22] and may be perceived to be more useful [10], thus a separate feature within the chatbot may appear less cluttered and more focused to the user.

3.2 Design Goals for Explainable Chatbot Interfaces

Based on the above considerations, we synthesized five key design goals for building in-application chatbots that can explain their underlying functionality during a breakdown:

1. **DG1: Explain the chatbot's functionality in terms of entity and intent.**

   Because chatbot identifies intent and entity, the system should explain the chatbot's functionality in terms of intent and entity so that users can better understand the high-level underlying working of the chatbot.

2. **DG2: Illustrate competencies of the chatbot and reasons why a breakdown occurred**

   The explanations should be designed to provide information on what chatbot could and could not comprehend to help users assess breakdown and competencies. Within the breakdown decision, the explanations should indicate the exact reason of the chatbot's failure by elucidating "Where" and "What cause" the breakdown with respect to intent and entity identified in the users' query.

3. **DG3: Provide normative and comparative example-based explanations**

   The system should leverage the normative and comparative example-based explanations for explaining both the competencies and breakdown reasons, respectively.

4. **DG4: Provide visual step-by-step explanations**

The system should take advantage of the visual explanations presented step-by-step to make them appealing, relatable, and better comprehensible.

5. **DG5: Allow users to have freedom and control in navigating the explanations**

   The system should provide control to the users to access the explanations when required. Also, the system should include UI controls such as "next", "previous", or "exit" to give users the freedom to navigate explanations as required thus making it less annoying.

3.3 Summary

In summary, we learned the importance of appropriate and to-the-point structuring of the explanations for explaining the chatbot's functionality. We also learned that the explanations could be enhanced with visuals and examples to make them more intuitive and useful. Informed by these design implications from the recent literature, we synthesized the design goals for building our novel in-application explainable chatbot system in the next chapter.

Chapter 4

Explainable Chatbot Interfaces (ChatrEx): System Design and Implementation

Informed by the design goals in the previous chapter *(Chapter: 3)*, we designed and implemented novel web-based chatbot interfaces that simulate breakdowns and their corresponding explanations. In this chapter, we present our system design and implementation for in-application explainable chatbot interfaces (ChatrEx). We also describe the iterative prototyping process that we followed across multiple ideation phases to design and evaluate low and medium fidelity prototypes for ChatrEx.

We began with low-fidelity paper prototypes, followed by image-based mock-ups using PowerPoint, and medium-fidelity prototypes using the *Axure* prototyping software. We solicited informal user feedback at each stage to help us iterate on our ideas before we finalized our web-based interactive chatbots. After conducting several rounds of brainstorming and following iterative design and prototyping approaches [67] (explained in detail in the Section 4.4), we arrived at two designs for ChatrEx: *ChatrEx-VST*(Figure 4.5) and *ChatrEx-VINC* (Figure 4.3, 4.4). These designs present two different ways for a chatbot to explain its underlying functionality during a breakdown and why it failed, including reasons related to disambiguation and infeasibility. We selected *Google Sheets* as the underlying application as it has several complex spreadsheet features that would allow us to devise tasks for chatbot assistance. [60]. Next, we present a brief overview of how a user can initiate a conversation with ChatrEx followed by the detailed description of the explanation design for the two novel chatbot interfaces, ChatrEx-VINC and ChatrEx-VST.

Our ChatrEx-VINC and ChatrEx-VST chatbots represent two different kinds of visual explanations, as described below. In both cases, users can issue text-based queries to the

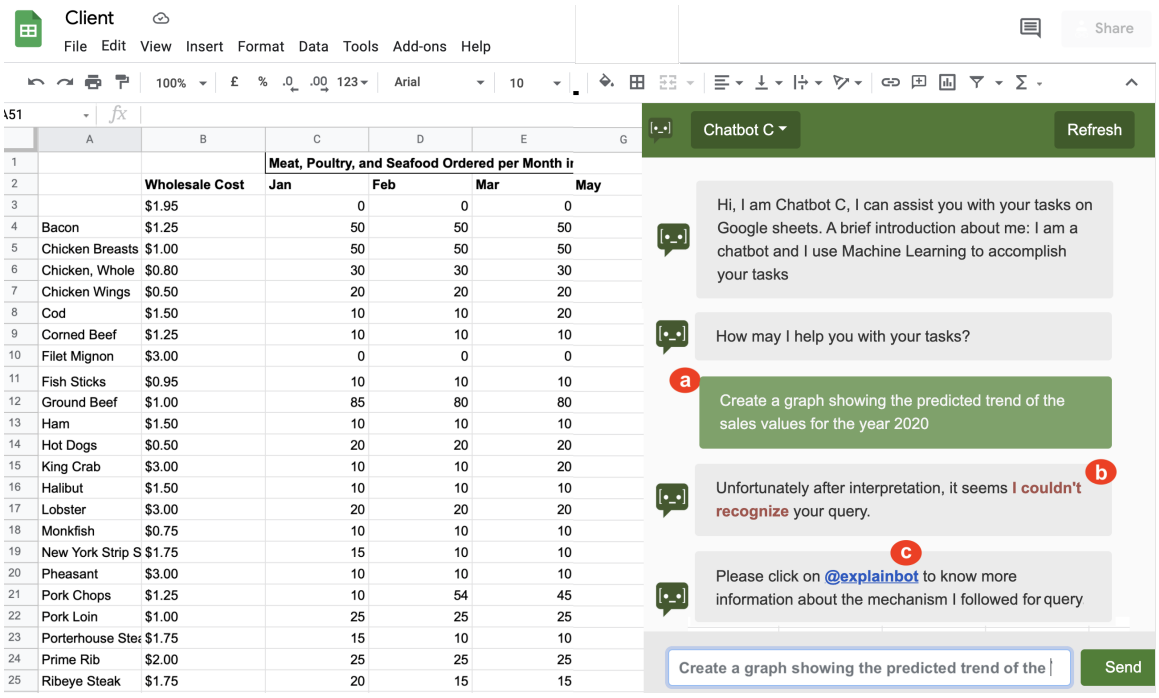
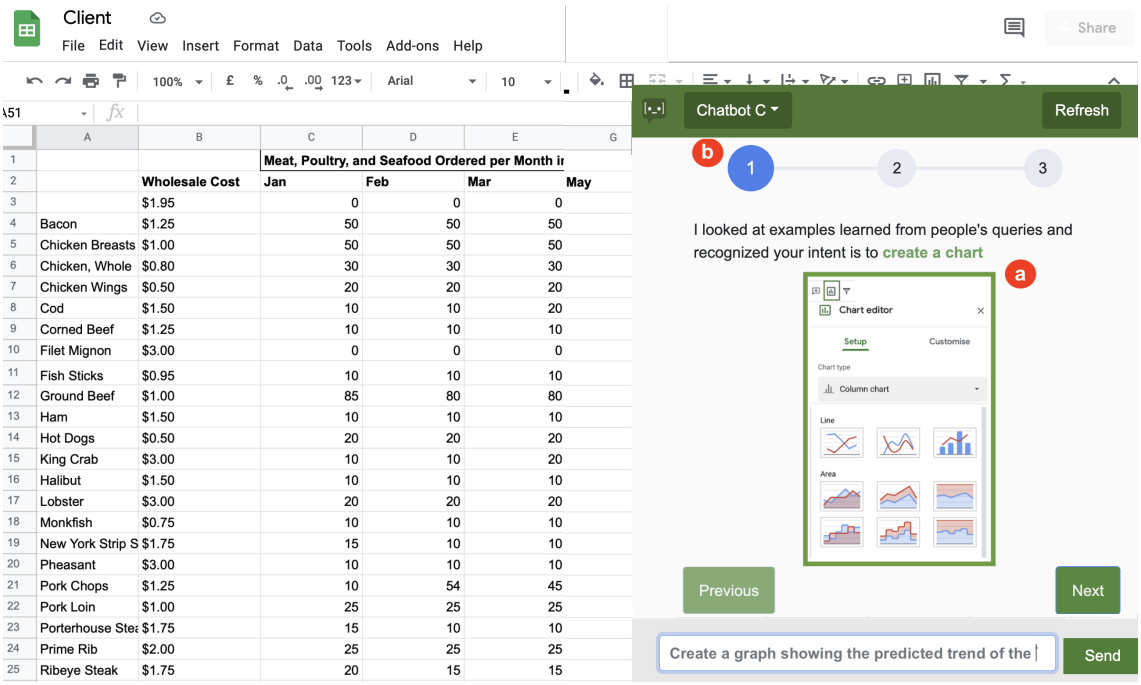


Figure 4.1: The common entry point for ChatrEx: (a) users submit a query, (b) error message is shown, (c) @explainbot feature can be invoked in response.

chatbot to initiate a conversation about automating spreadsheet tasks (Figure 4.1.a). As the entry point into ChatrEx, the introduction message enables users to become familiar with the interface, and guides them towards a helpful conversation. If a user sees an error message (Figure 4.1.b) after issuing the query, they can invoke the @explainbot feature (Figure 4.1.c) to see an explanation about what the chatbot understood and why the breakdown occurred.

4.1 ChatrEx-VINC: Visual in-context explanations

ChatrEx-VINC provides in-context visual example-based step-by-step explanations (DG4). Similar to the idea of example-based explanations based on the training set for a classifier [38, 11], ChatrEx-VINC shows examples from the training set of each keyword in the query (i.e., intent and entity) recognized by the chatbot. Fulfilling DG1, DG2 and DG3, ChatrEx-VINC distinctly explains the intent/entity that the chatbot comprehended successfully through training examples from the target class (normative explanations) (examples shown in Figure 4.2.a, 4.3.a, 4.4.a). Similarly, ChatrEx-VINC further explains the breakdown decision through the most similar or different examples from the alternative training classes (i.e., comparative explanations). In particular, when the breakdown occurs due to disambiguation, the explanation provides similar examples which matched the user's intent



	A	B	C	D	E
1			Meat, Poultry, and Seafood Ordered per Month i		
2		Wholesale Cost	Jan	Feb	Mar
3		$1.95	0	0	0
4	Bacon	$1.25	50	50	50
5	Chicken Breasts	$1.00	50	50	50
6	Chicken, Whole	$0.80	30	30	30
7	Chicken Wings	$0.50	20	20	20
8	Cod	$1.50	10	10	20
9	Corned Beef	$1.25	10	10	10
10	Filet Mignon	$3.00	0	0	0
11	Fish Sticks	$0.95	10	10	10
12	Ground Beef	$1.00	85	80	80
13	Ham	$1.50	10	10	10
14	Hot Dogs	$0.50	20	20	20
15	King Crab	$3.00	10	10	20
16	Halibut	$1.50	10	10	10
17	Lobster	$3.00	20	20	20
18	Monkfish	$0.75	10	10	10
19	New York Strip S	$1.75	15	10	10
20	Pheasant	$3.00	10	10	10
21	Pork Chops	$1.25	10	54	45
22	Pork Loin	$1.00	25	25	25
23	Porterhouse Ste	$1.75	15	10	10
24	Prime Rib	$2.00	25	25	25
25	Ribeye Steak	$1.75	20	15	15

Figure 4.2: An example of **ChatrEx-VINC** displaying normative visual training examples (a), highlighted in green, to convey chatbot's competencies .

or entity and were possibly misrecognized (Figure 4.3.f). In contrast, when the breakdown occurs due to a task being infeasible for the chatbot, the explanations provide feasible alternative examples which users can follow instead of the original intent or entity that the chatbot is not trained for (Figure 4.4.f).

To provide a better understanding of the chatbot during the breakdown, these examples (Figure 4.3.f, 4.4.f) are accompanied by corresponding match percentages (Figure 4.3.c, 4.4.c). This is analogous to confidence scores within an intent-based model [63] that represents the similarities between the user's intent and examples in the training set. Further, the explanations also highlight the competencies of the chatbot in green (Figure 4.2.a, 4.3.a, 4.4.a) and breakdowns in red (Figure 4.3.f, 4.4.f) along with a clear dialog message. To show the real-time system status more interactively (as suggested in [42]), we adopted a design similar to the "Status Tracker" UI [57] to show the step-by-step explanations in the form of latest status and updates, displayed in chronological order (Figure 4.2.b, 4.3.b, 4.4.b). When each of these steps are visited by the user, they are updated with GREEN check marks (Figure 4.3.d, 4.4.d) allowing users to follow the explanation steps intuitively. Addressing DG5, users can control the navigation of these explanations by using "next" and "previous" buttons.

For the demonstration, consider an example query for the disambiguation, "Create a graph showing predicted trend of the sales values for the year 2020". Once the user type-in

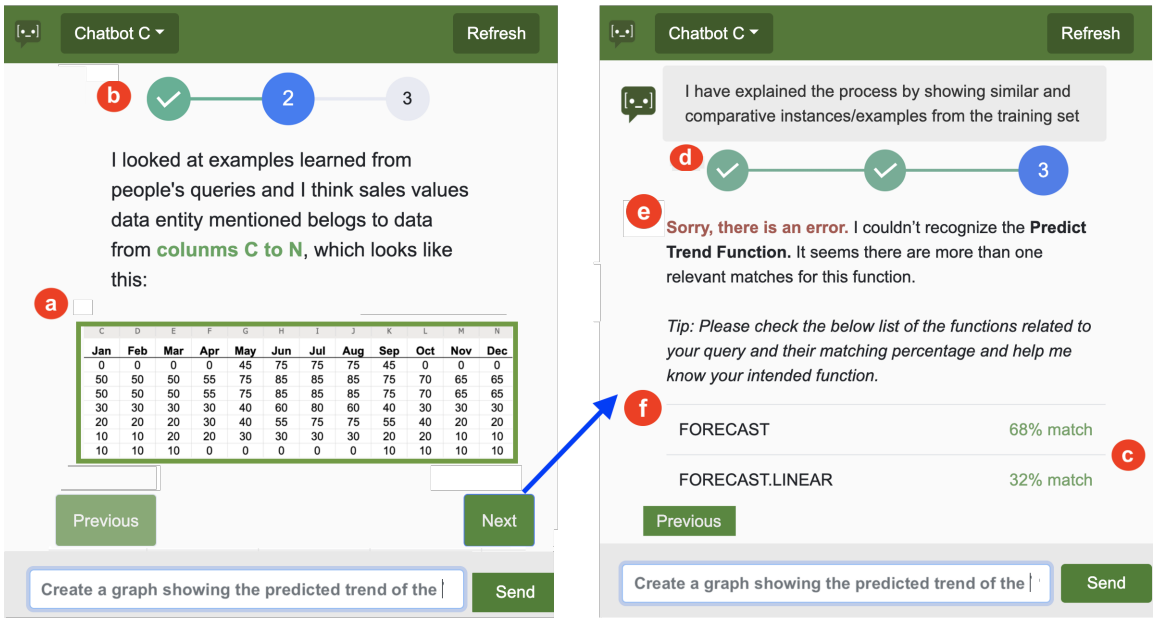


Figure 4.3: An example of **ChatrEx-VINC** step-by-step within-chat explanations illustrating the competencies *(Left)* and breakdown *(Right)* for disambiguation task. Comparative visual examples are shown for most similar visual training examples (e), the potential matches (f), and match percentages (c)

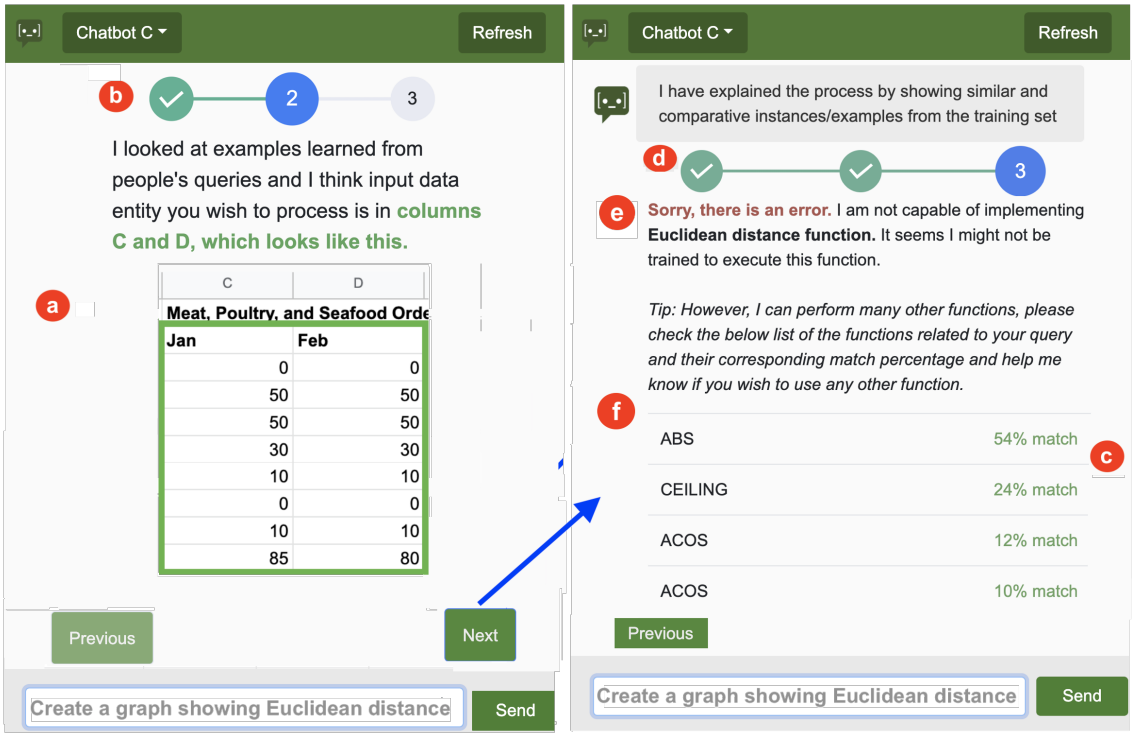


Figure 4.4: An example of **ChatrEx-VINC** step-by-step within-chat explanations illustrating the competencies *(Left)* and breakdown *(Right)* for infeasible task. Comparative visual examples are shown for most similar visual training examples (e), the potential matches (f), and match percentages(c)

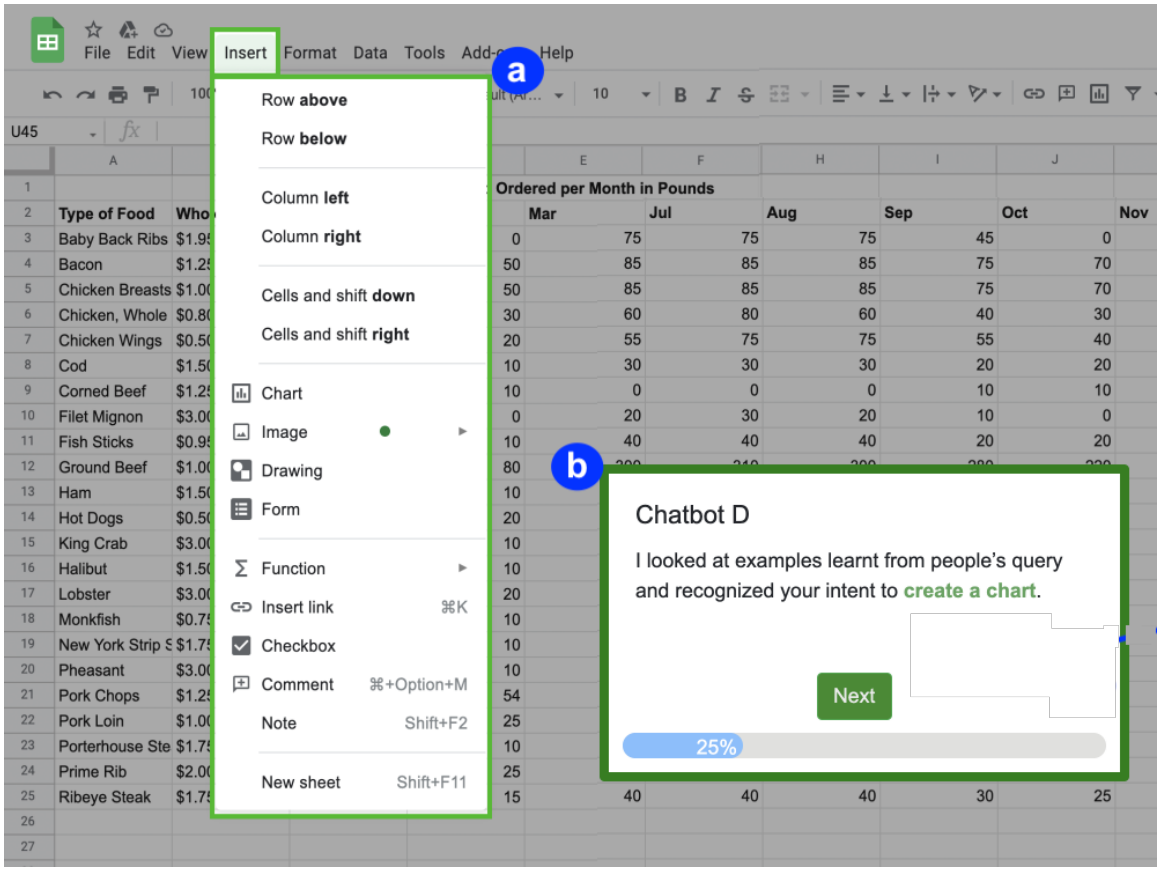


Figure 4.5: **ChatrEx-VST Competencies:** By clicking @explainbot, VST presents a visual tour overlaid on the application UI, highlighting normative visual training examples on the interface in green (a,b). *[Note: both ChatrEx designs support disambiguation and infeasible queries, here we showed infeasible query for ChatrEx-VST]*

the query, as shown in Figure 4.1, ChatrEx-VINC prompts a brief breakdown error message (Figure 4.1.b) and guides the user to seek explanations using the @explainbot feature. Upon the clicking of @explainbot, ChatrEx-VINC presents within-chat explanations as shown in the Figure 4.2. For this task, the chatbot step-by-step explains that it recognized the intent to create a graph (Figure 4.2.a), and the data entity from column C to N (Figure 4.3.a) and presented them with the corresponding normative visual examples. Further, the chatbot explains that the breakdown occurs highlighted in red due to the chatbot's misrecognition of the Predict Trend Function (Figure 4.3.e) and therefore the explanations provide comparative visual examples along with their match percentage which users can follow. The user can hover over each function name and seek the detailed description for each function as provided by Google sheets.

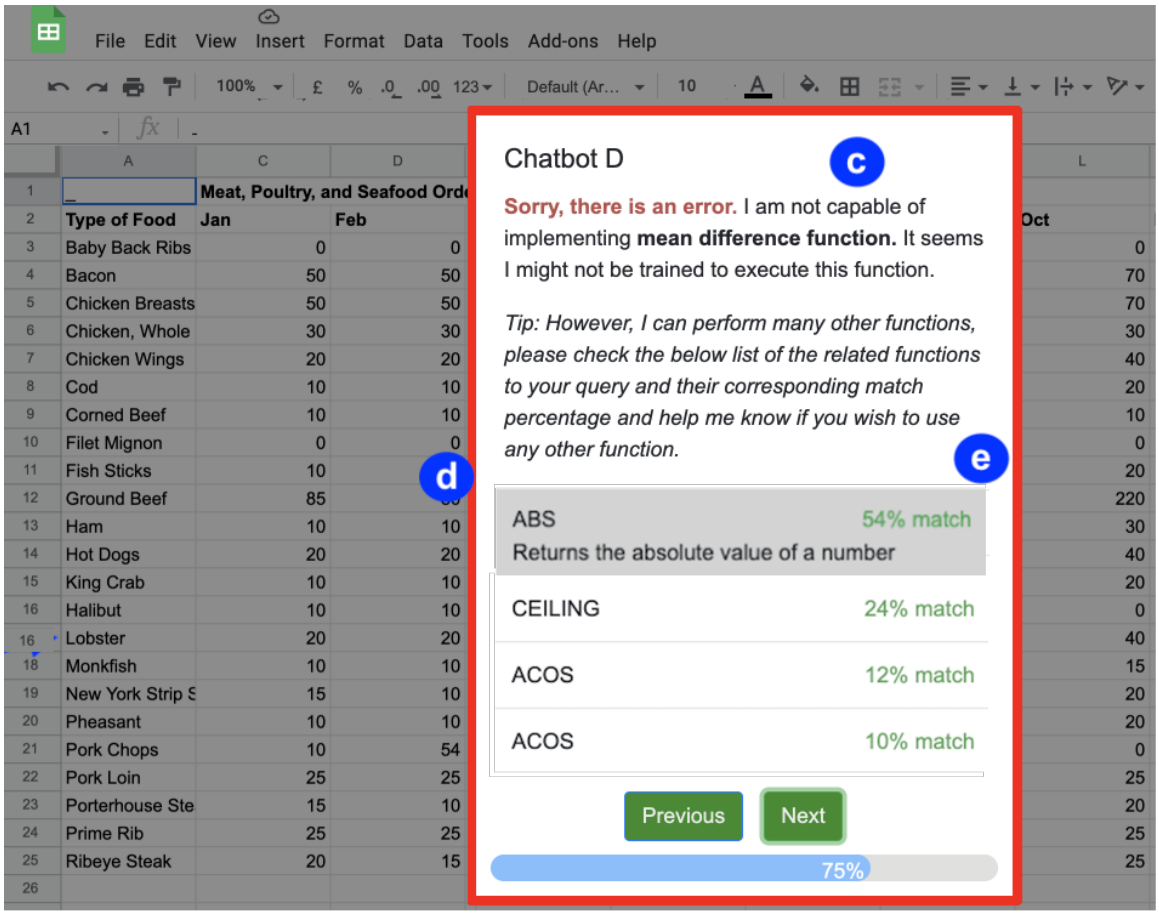


Figure 4.6: **ChatrEx-VST Breakdown decision:** provides comparative visual example-based explanations (c) through alternative visual training examples (for an infeasible query) along with match percentages (d,e). *[Note: both ChatrEx designs support disambiguation and infeasible queries, here we showed infeasible query for ChatrEx-VST]*

Similarly, for the infeasible task query, "Create a graph showing Euclidean distance between Column C and D", the chatbot recognizes the intent to create a graph, and the data entity Column C & D successfully by highlighting the corresponding visual examples from the application in green within the chat window (Figure 4.4.b). In contrast, for the breakdown reason, the chatbot explains that it was not trained to recognize and execute the Euclidean distance function, making the task infeasible (Figure 4.4.e). Further, the explanations provide alternative visual examples (Figure 4.4.f) along with their match percentage (Figure 4.4.c) which users can follow instead of the original intent or entity that chatbot is not trained to do.

4.2 ChatrEx-VST: Visual step-through explanations

In contrast to ChatrEx-VINC, ChatrEx-VST presents a step-by-step visual tour with examples overlaid (DG4) directly on the application user interface (Figure 4.5). We draw inspiration from in-application software walkthroughs or onboarding tours that explain features and functionality in a way that is relatable and engaging for users [47].

When a user invokes the @explainbot feature (Figure 4.1.c), ChatrEx-VST directs the user to seek more information about their query and explains what the chatbot understood. First, the chatbot asks for the user's permission to begin the tour and after confirmation, ChatrEx-VST minimizes the chat window and overlays a transparent background atop the UI. Next, it highlights (Figure 4.5.a, 4.5.d) the visual examples in the UI (e.g., menu items, data items, functions, etc.) corresponding to the intent or entity recognized from the user's query (DG1) along with a descriptive message. Similar to ChatrEx-VINC, ChatrEx-VST fulfills DG3 and DG4 by distinctly explaining the chatbot's competencies through normative visual explanations highlighted in green boxes (Figure 4.5.a, 4.5.b) and breakdown decision through comparative explanations highlighted in red boxes along with match percentages (Figure 4.5.d, 4.5.e). Addressing DG5, the user can easily navigate to the next or previous step on their own (Figure 4.5.f). Finally, the dialog box has a Finish button to end the overlaid tour and bring the users back to the chat window.

For example, consider an infeasible task query "Create a graph showing the mean difference of Column C with D". For this task query, the chatbot prompts the users to seek confirmation for beginning the visual tour. After confirmation, ChatrEx-VST overlays on the user's current page and recognizes the intent to create a chart (Figure 4.5.b) by highlighting similar visual examples from the target class directly on the application in green boxes (Figure 4.5.a). The user can then navigate to the next or previous dialogs using the button on the dialog box. The progress bar shows the user's progress. When the breakdown occurs, ChatrEx VST explains its infeasibility to execute the mean difference function due to lack of training (Figure 4.6.c). The chatbot, further, highlights the breakdown reason in a red box and displays the alternative visual examples (Figure 4.6.d) along with their match percentage (Figure 4.6.e) with the user's intent or entity. The users can hover over each function name and seek a detailed description. Similar to ChatrEx-VINC, ChatrEx-VST assists both disambiguation and infeasible task query, however, here we only showed the demonstration of ChatrEx-VST with respect to the infeasible task.

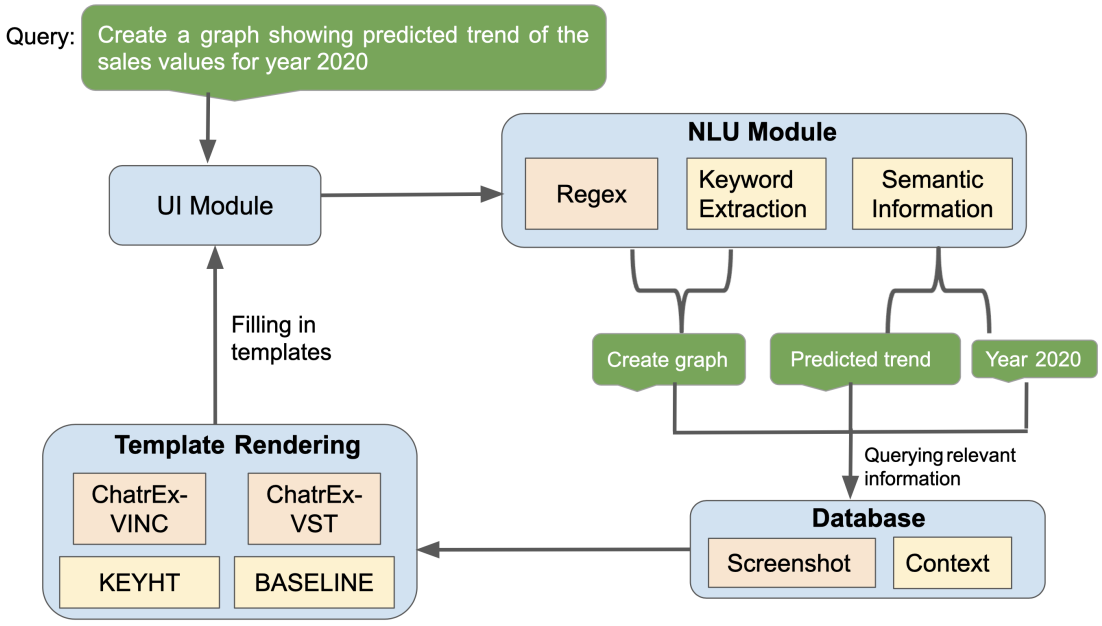


Figure 4.7: Implementation diagram for ChatrEx-VST and ChatrEx-VINC

4.3 Implementation details

Our main contribution in this thesis lies in exploring the interface design of visual explainable chatbots and our implementation focused on developing proof-of-concept prototypes, rather than contributing new algorithmic innovations in NLP or ML. We created interactive web-based prototypes to demonstrate the chatbot functionality and evaluate the different explanation designs with users for spreadsheet tasks related to statistical and visualization (e.g., creating graph) functions. To implement our ChatrEx designs, we took inspiration from chatbots that rely on intent-based models where multi-classifiers predict the intent in the user's query and calculate confidence scores with respect to all predefined intents. A breakdown occurs if all of these confidence scores are below a certain threshold. The highlighting of the breakdown in red is inspired by the typical ML approach used to identify keywords in the query having the highest weight on predicted intent. Aspects of the visual examples are inspired from recent work [11], where the training set included visual examples for predefined intent.

As shown in Figure 4.7, ChatrEx consists of two main modules: the UI module and the NLU (Natural Language Understanding module). The UI module lays out the various user interface components and receives the user's query. Next, this input query is sent to the NLU module, which uses regex and keyword extraction to detect a user's intentions and runs the query through another model to extract semantic information about the task. For

example, for the query “Create a graph showing predicted trend of the sales values for the year 2020”, the NLU model runs regex and keyword extraction to detect the user’s intent of creating a chart and further extract the semantic information or entities such as function name (e.g., predicted trend) and/or data items (e.g., year 2020). This intent and semantics extracted from the user queries are then sent to our pre-existing manually curated database which contains labeled queries and the corresponding series of screenshots and context for each predefined intent and/or entity. The identified intent (e.g., graph) and extracted semantics (e.g., predicted trend, year 2020) are mapped against our pre-existing database to retrieve the corresponding series of screenshots and context. The retrieved data are then used to fill our predefined templates for the different chatbot types (ChatrEx-VST and ChatrEx-VINC, and two other implementations used for comparison in the user study). To generate the explanation responses, these templates are then rendered using our UI module within each chatbot. The UI module is built using ReactJS and migrated to Chrome as an extension by adapting a boilerplate template [1].

4.4 Iterative design and prototyping process

We arrived at the novel designs for our in-application explainable chatbot interfaces by using a user-centered iterative design approach. We followed the guidelines advocated by the "research through design" [67] paradigm. In this section, we highlight how our ideation and prototypes evolved across the different stages of design and how we made use of low and medium-fidelity prototypes. We also describe how we used user feedback to help us iterate on design choices for our final prototypes.

4.4.1 Stages of low and medium fidelity prototypes

Based on the design goals (Section 3.2) informed from the recent literature, we began exploring the design space by conducting several brainstorming and sketching sessions to come up with various design ideas and their iterations. Throughout these brainstorming design sessions, our overarching goal was to investigate:

"What can be a useful way for a chatbot to explain itself if users wanted to know why it did or did not understand their query?"

In these sessions, we built low-fidelity prototypes in the form of paper mock-ups *(Stage 1: Exploratory design stage)* followed by image-based PowerPoint mock-ups *(Stage 2: Detailed*

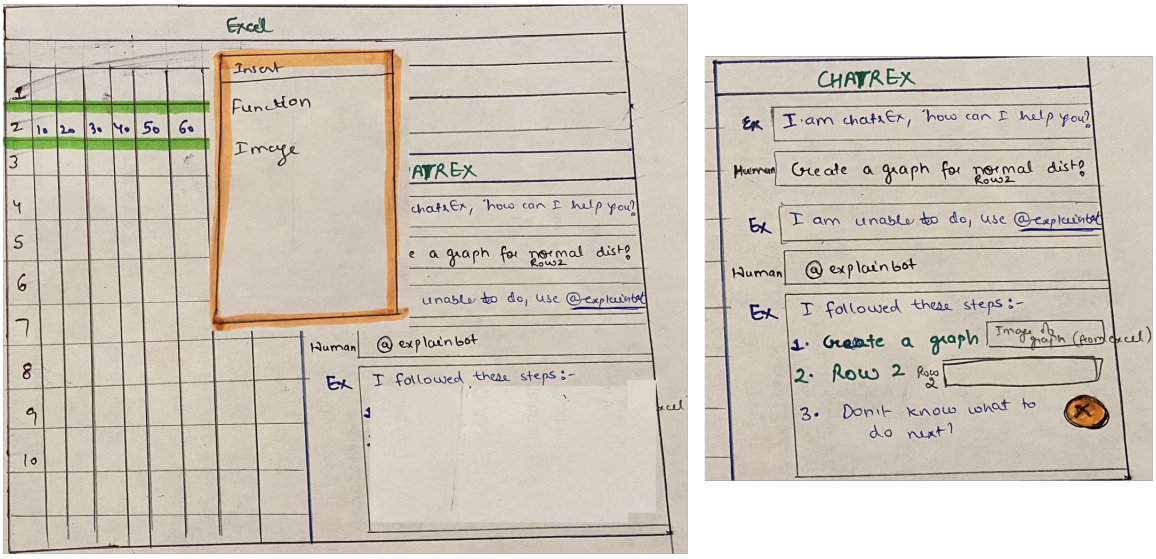


Figure 4.8: An example of paper mockups

prototypes). After carefully analyzing each of the design ideas and considering their pros and cons, we progress to medium-fidelity prototypes *(Stage 3: Partially interactive prototypes)* using the Axure prototyping tool (`https://www.axure.com`). These three stages of low and medium fidelity prototypes and the related ideation process is described below:

- **Stage 1: Exploratory design stage using Paper mock-ups**

  We began the exploratory design stage by sketching our different design ideas on paper. The paper sketches provided us an opportunity to explore a wide range of ideas and visualize high-level concepts. We proposed different solutions for designing the explanations based on our derived design goals. As these sketches were easier to generate and change, we could brainstorm and refine several ideas broadly for each design aspect at a time. An example of some of these sketches is shown in Figure 4.8.

- **Stage 2: Detailed prototypes stage using Image-based PowerPoint mock-ups**

  Next, we evaluated our different design sketches and decided to choose two ideas: within-chat and step-through explanations. We evolved these chosen design ideas and focused next to capture more detailed prototypes using image-based PowerPoint mock-ups. We created several iterations for within-chat and step-through based prototypes and assessed the positives and negatives of each iteration with the help of informal feedback from other team members.

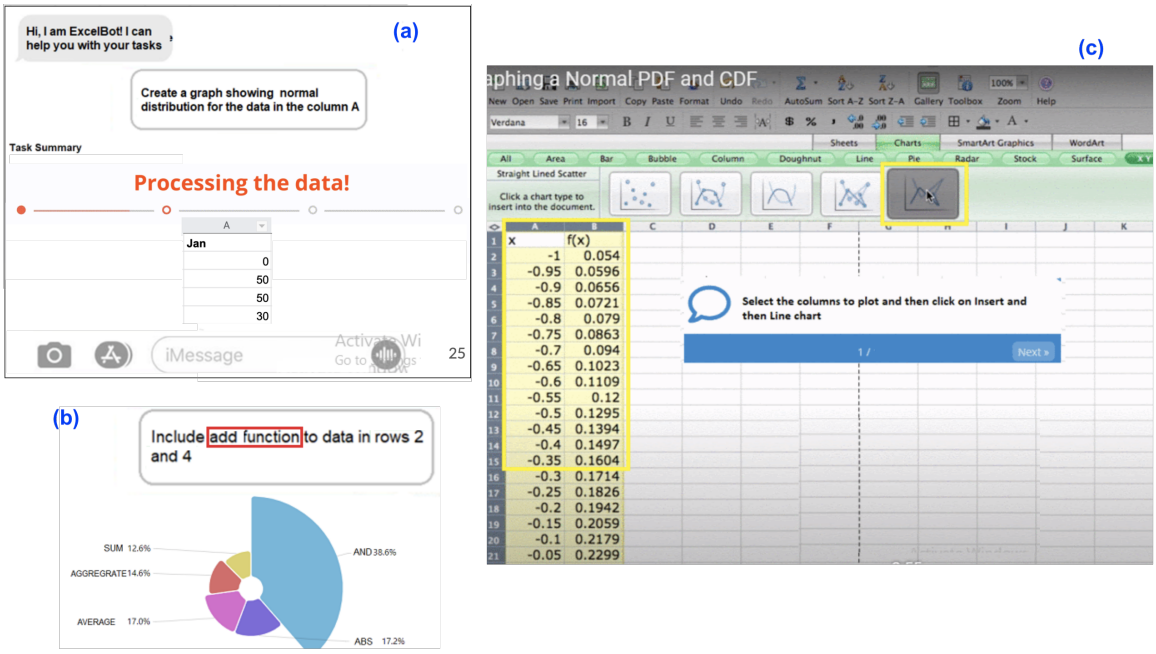


Figure 4.9: An example of Image-based PowerPoint mock-ups

For example: As shown in the Figure 4.9 , we presented two iterations (Figure 4.9.a, 4.9.b) for within-chat explanations where we explored different representations i.e. Graph and Status-Tracker UI. Similarly, we presented an iteration for step-through explanation (Figure 4.9.c). Further, these PowerPoint mock-ups allowed us to consider interaction choices and capture actual interactivity and dynamics among the various UI design elements and visuals.

- **Stage 3: Partially interactive prototypes using axure-based medium-fidelity prototypes**

  Finally, we explored the structure, layout, content, and functionality of the distilled image-based PowerPoint mockups. We created partially interactive web-based chatbot prototypes to simulate the functionality along with the 'look and feel'. These prototypes demonstrated enough of the chatbot's functionality so that we could get user's feedback and evaluate explanation designs. We created these web-based chatbot prototypes using the Axure Rp9 software (`https://www.axure.com/new-in-9`). We used various existing features of Axure, such as widgets, and assigned pre-defined actions such as OnClick, OnMouseOver, and OnMouseOut in response to actions for building the interactions within these widgets. For these medium-fidelity prototypes, we mimicked in-application chatbots by presenting the Axure prototypes overlaid atop a screenshot from Google Sheets. These prototypes were hard-coded and allowed the users to explore these explanations for a limited set of spreadsheet queries. Finally,

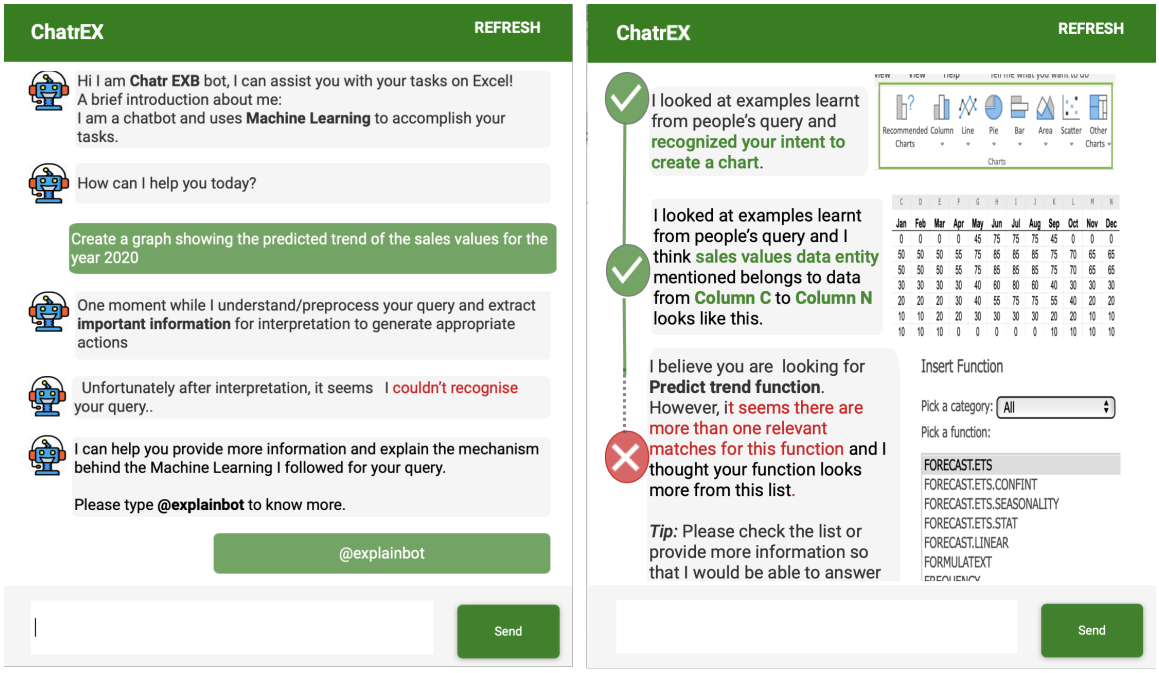


Figure 4.10: An example of Axure-based medium-fidelity prototypes (ChatEx-VINC)

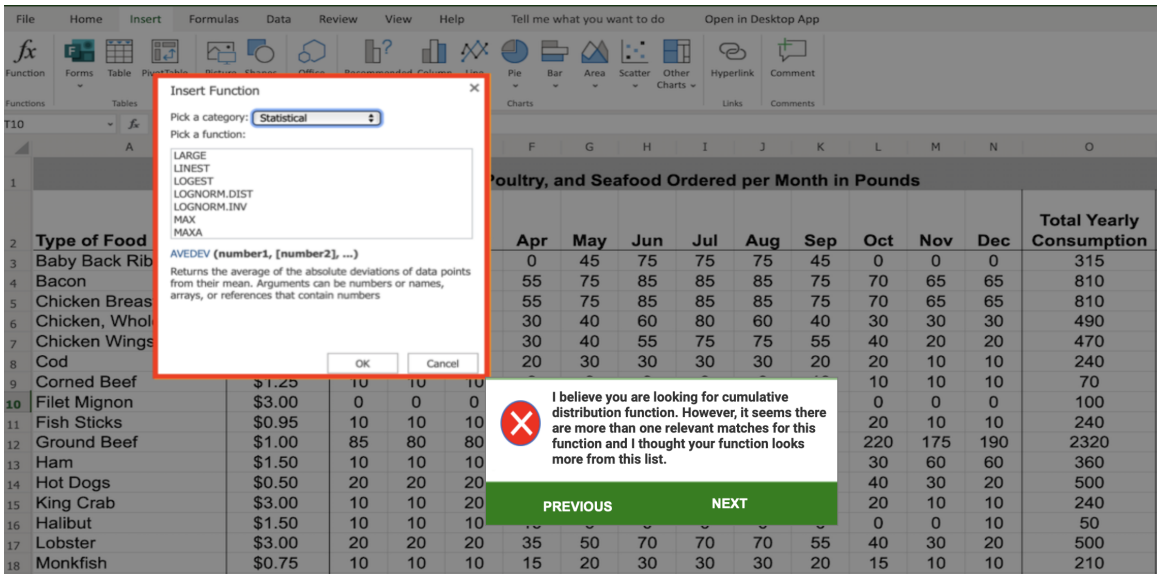


Figure 4.11: An example of Axure-based medium-fidelity prototypes (ChatEx-VST)

we published them to Axure Cloud (`https://www.axure.cloud`) to generate HTML websites that could be shared for preview and testing.

An example of the Axure prototypes for ChatrEx-VINC and ChatrEx-VST are shown in the Figure 4.10 and Figure 4.11 respectively.

4.4.2 User Feedback

Before implementing these Axure prototypes into a higher fidelity functioning web-based system, we conducted a brief formative study using the Axure-based prototypes and sought informal feedback from users. We took a qualitative approach and conducted semi-structured interviews with 11 participants from diverse backgrounds having varying experiences with using chatbots. This study was conducted online through an end-to-end encrypted video conferencing platform i.e. Zoom and the demographic questionnaire was hosted online on SurveyMonkey.

We presented the participants with each Axure-based prototype in random order and asked them to play around with the prototypes using pre-structured sample spreadsheet queries. We focused on eliciting their feedback and initial perceptions of the design of the chatbot explanations. Lastly, we asked them some questions to probe further into how well ChatrEx's explanations helped users to understand the "underlying chatbot working". Overall, the session lasted for approximately 45 minutes. Participants were encouraged to think aloud while interacting with the chatbot interfaces.

Key findings:

Among the design prototypes shown, we observed that almost all the participants found the general idea of seeking explanations from a chatbot to be useful and interesting. Notably, most participants found the visual step-by-step display of explanations to be fairly clear and comprehensible leading and suggested this could be a "less frustrating" way of interacting with chatbots. Further, participants indicated that the suggestive feedback and the user interface design of both ChatrEx VST and VINC were useful in acknowledging competencies and breakdown distinctly:

> *"It(ChatrEx) shows me in green what it understands and highlights the important aspects of the query, and also tells me what's wrong by highlighting parts of my query in red"-P03.*

However, a few of the participants expressed some concerns about certain user interface elements introduced in ChatrEx. For example, one of the participants mentioned that the initial conversation (before @explainbot as shown in Figure 4.2.a) was too overwhelming

which made it difficult for them to "grasp the information". Further, the green and red check marks seemed to overpower the entire interface of the chatbot and hence appeared irrelevant to the users. They pointed out that the highlighting as red and green boxes were sufficient enough to differentiate the competencies and breakdown. We addressed this feedback in our final implementation and removed the unnecessary conversation and check marks. As users were constrained in the Axure prototypes in terms of spreadsheet task queries, in our final implementation, we included a larger set of spreadsheet queries that users could explore dynamically.

4.5 Summary

In this chapter, we presented the design of ChatrEx, our novel web-based in-application explainable chatbot interfaces atop Google Sheets. In particular, we presented the system design and implementation of two variations, namely ChatrEx-VINC and ChatrEx-VST. Finally, we described our iterative prototyping process for exploring numerous design ideas and creating multiple low and medium fidelity prototypes. In the next chapter, we will design a user study to evaluate our ChatrEx system in terms of usefulness, transparency and trust.

Chapter 5

User evaluation of ChatrEx: User Study and Findings

In Chapter 4, we presented the design and implementation of web-based prototypes for two variations of ChatrEx: *ChatrEx-VINC* and *ChatrEx-VST*. To evaluate the extent to which the proposed ChatrEx-VINC and ChatrEx-VST designs help users understand the chatbot's explanation for a breakdown, we ran a usability study with 14 participants. In this chapter, we present the details of our study methodology and how we evaluated our web-based prototypes of ChatrEx using both quantitative and qualitative approaches. Further, we also discuss our key findings from this study.

5.1 User Study

To evaluate our ChatrEx designs, we compared them with two other types of chatbots (one baseline and one state-of-the-art chatbot closet to our work). We implemented the following two chatbot prototypes:

(1) **KEYHT**, which was adapted from recent work on verbal keyword highlighting and confirmation explanations [5]. As shown in Figure 5.1(Left), KEYHT highlights keywords (Figure 5.1) it understood in green and the keywords that it misunderstood in orange;

(2) **BASELINE**, which was our implementation of commonly used in-application chatbot designs that do not provide any explanations, but often recommend related search results during the system breakdown [42] as shown in Figure 5.1(Right).

The goal of this study was to assess the strengths and weaknesses of the different explanation designs, thus tackling the key research question:

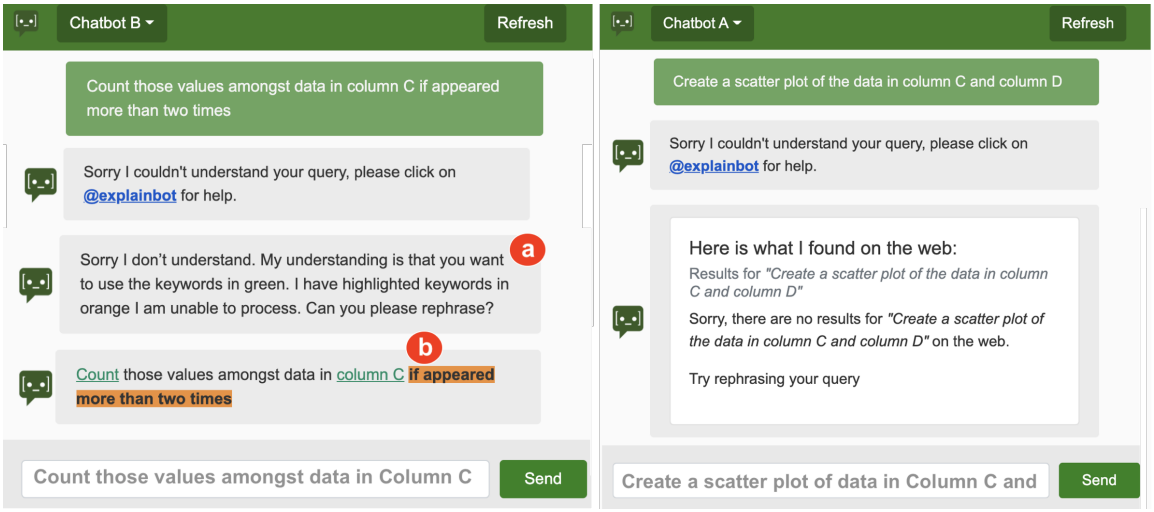


Figure 5.1: ***(Left)*** **KEYHT:** Verbal Keyword Highlight Explanation Design, displaying (b) the explanation by highlighting the competencies (Green) and breakdown decision (Orange) ***(Right)*** **BASELINE :** Traditional chatbots with No explanations for their decisions and resorting to web search

RQ1: *What are the strengths and weaknesses of chatbot interfaces with different explanation designs and how they are perceived by users in terms of (a) usefulness, (b) transparency and (c) trust?*

5.1.1 Participants

We recruited the participants mainly from our university's mailing lists and found additional participants through personal connections and snowball sampling by sending emails and messages. We ended up with a diverse pool of 14 participants (7F/7M) who came from different backgrounds (CS, Sciences, Arts) and self-reported having little to no experience with ML and only one participant had formal training in CS. Our participants were all between the ages of 18–34 and came from a diverse range of professions (client services, lab technicians, medical photographers, information designers, engineering students, researchers) and had different levels of education (1 Bachelor's, 1 Diploma, 8 Master's, 4 Ph.D.). For the purpose of having a wider overview of the perceptions and expectations of the users, we focused to recruit a broad range of participants who have varying experience in using chatbots to accomplish tasks in their regular lives. The participants were familiar with a range of chatbots, including Google Assistant(12/14), Alexa(11/14), Siri(13/14), but most participants (9/14) did not use these chatbots frequently (at most 3 times/week).

#	Chatbot Type	Example Query (Infeasible)	Competencies (intent/ entity)	Breakdown Reason (intent /entity)
1	ChatrEx-VINC	Create a graph showing euclidean distance between data in column C from column D	**graph, column C and D data**	**euclidean distance (not trained)**
ChatrEx-VINC Response**: *Displays visual examples in-context for above-mentioned competencies and breakdown (intent/entity) indicating chatbot is not trained.*				
2	ChatrEx-VST	Create a graph showing the mean difference of Column C with D	**graph, column C and D data**	**mean difference function (not trained)**
ChatrEx-VST Response**: *Displays visual examples as tour for above-mentioned competencies and breakdown (intent/entity) indicating chatbot is not trained.*				
3	KEYHT	For data values from January to December in "Client" Workbook, create a heat map	**column C-N data, Client workbook**	**heat map**
KEYHT Response**: *Highlights the keywords chatbot understood in green the keywords that it misunderstood in orange, as shown above.*				
4	BASELINE	Create a histogram showing the frequency of data in column C	-	-

Table 5.1: Example query for **infeasible breakdown situation** covered in the user study and corresponding competencies as well as breakdown reason recognized by each chatbot. *[Note: As BASELINE is inspired from the traditional chatbots that doesn't provide any explanations, therefore "-" represents No explanations for competencies and breakdown reason]*

5.1.2 Study Instruments

We collected basic demographic information from the participants via a pre-test questionnaire e.g., age, gender, occupation, education. In addition to basic demographic information, the pre-test questionnaire asked participants about their previous experience in working with the traditional chatbots and Google Sheets along with the approximate usage frequency per week. The post-task usability questionnaire consisted of several 5-point Likert scale responses to rate the overall user experience (frustrating) and further assessing each chatbot's explanation designs in terms of usefulness, ability to recognize the breakdown reason, understanding how the chatbot works, ability to improve the query next time, and ability to trust the chatbot to finish their task. In order to assess how well ChatrEx's explanations aid transparency, the participants were asked to reflect upon the explanations and explain the "underlying chatbot working" with respect to query and reason of chatbot failure. Finally, the interview probed further into the strengths and weakness of each chatbot design interface, wherein, users were asked to rank the four prototypes they interacted

with, in terms of comprehensibility and trust. Additionally, the interview provided insights into users' interest in using these prototypes for their spreadsheet tasks.

#	Chatbot Type	Example Query (Disambiguation)	Competencies (intent/ entity)	Breakdown Reason (intent /entity)
1	ChatrEx-VINC	Create a graph showing the predicted trend of sales values for year 2020	**graph, column C-N data**	**predicted trend (match with FORECAST, FORECAST.LINEAR)**
ChatrEx-VINC Response**: *Displays visual examples in-context for above-mentioned competencies and breakdown (intent/entity) indicating chatbot misunderstood or disambiguate.*				
2	ChatrEx-VST	Create a graph showing normal distribution for data in column C	**graph, column C data**	**normal distribution (match with NORM DIST, NORM.DIST)**
ChatrEx-VST Response**: *Displays visual examples as tour for above-mentioned competencies and breakdown (intent/entity) indicating chatbot misunderstood or disambiguate.*				
3	KEYHT	Count those values amongst data in column C if appeared more than two times	COUNT function, column C data	**"if appeared more than two times"**
KEYHT Response**: *Highlights the keywords chatbot understood in green the keywords that it misunderstood in orange, as shown above.*				
4	BASELINE	Create a scatter plot of the data in column C and column D	-	-

Table 5.2: Example query for **disambiguation breakdown situation** covered in the user study and corresponding competencies as well as breakdown reason recognized by each chatbot. *[Note: As BASELINE is inspired from the traditional chatbots that doesn't provide any explanations, therefore "-" represents No explanations for competencies and breakdown reason]*

5.1.3 Study Design and Tasks

We used a within-subject design to minimize the impact of known high variation among participants. Each participant interacted with four web-based chatbot prototypes that represented one of the explanation designs (ChatrEx-VST, ChatrEx-VINC, KEYHT, BASELINE) in a random order (to eliminate order effects). For each chatbot, we asked users to try two distinct spreadsheet tasks each (8 in total) that represented two breakdown situations:

1) *Infeasible tasks (Table 5.1)*: these spreadsheet tasks resulted in a breakdown because the chatbots were not trained to recognize and perform them. For example, as shown in

Table 5.1.1, for the task "Create a graph showing Euclidean distance between Column C and D", our chatbots were not trained to recognize and execute the Euclidean distance function, making the task infeasible.

2) *Disambiguation tasks (Table 5.2)*: these tasks resulted in a breakdown because although they were feasible in the spreadsheet, they could be misunderstood by the chatbot as there could be more than one relevant matches for identified intent or entity. For example, as shown in Table 5.2.1, for the task "Create a graph showing predicted trend of the sales values for the year 2020", the chatbot would not be able to recognize the intent for "predicted trend" because there were multiple matches (e.g., FORECAST, FORECAST.LINEAR, etc) and it would need more specific information to process the query.

We explored a range of complex statistical functions as we considered different aspects of feasibility and disambiguation. We explained to the users that the goal of our study was not to complete the actual tasks in Google Sheets, but to assess the explanations that they saw during breakdowns in their interaction with different chatbot designs. We conducted pilot testing and iterated the queries several times to strike a good balance between challenge, allotted time, and comprehensibility.

5.1.4 Procedure

We conducted the study remotely through Zoom and participants were each given a $15 Amazon gift card in appreciation of their time. There were two parts to our study: 1) Usability test 2) Brief Follow-up Interview. Before starting the first part, participants were presented with a set of slides explaining the objective of the study, common scenario, and how to use ChatrEx's explanation feature. Participants were then asked to install our prototypes via a Chrome extension that would make our chatbot designs functional on Google Sheets (an example spreadsheet was provided). Next, participants filled out a pre-test questionnaire (via SurveyMonkey) that captured demographics and information about prior experiences with virtual assistants and spreadsheet applications.

We presented each of the 4 chatbots and spreadsheet tasks one by one in random order. The participants were given the following scenario: they were employees of a technology company that is working on designing chatbot interfaces that provide explanations for user's Google sheets queries. The company expected them to test out their newly designed explanations for various spreadsheet tasks and provide their initial feedback. For each of these tasks, we asked participants to phrase an appropriate query and use the @explainbot feature to seek an explanation as they would do if they were in the aforementioned scenario. The think-aloud protocol was followed and the participants were encouraged to think aloud.

When necessary, we also provided participants hints for constructing an appropriate query as the purpose of our study was not to test the user's understanding of spreadsheet features and their ability to come up with queries. After interacting with each of the 4 chatbots, users filled out post-task questionnaires (via SurveyMonkey) to assess their overall experience and ability to improve their query along with their perceptions of usefulness, transparency, and trust. To assess how well the explanations aid transparency, participants were asked to explain their understanding of each chatbot's underlying working and the reason for the breakdown in their own words.

For the second part, we carried out follow-up interviews to further probe into the strengths and weaknesses of each chatbot's explanation design. We asked users to rank the four prototypes they interacted with in terms of explainability and trust. Sessions were video and audio-recorded for transcription, and the participants were asked to share their screen through Zoom (only during the usability test). The usability test and follow-up interview took approximately one hour.

5.1.5 Data Analysis

We used a combination of statistical tests and a bottom-up inductive analysis approach [17] to explore our study data about users' perceptions of usefulness, transparency, and trust. We ran Pearson's Chi-square test for independence with nominal variable "Explanation type" (having four levels: ChatrEx-VST, ChatrEx-VINC, KEYHT, BASELINE) and ordinal variable (having three collapsed levels: Agree, Neutral and Disagree) to quantitatively determine the significance of the results. We also qualitatively observed and analyzed the participant's approach for breakdown recovery. We then created affinity diagrams using the gathered data from the task observations and interviews. Through discussion and use of affinity diagrams, we categorized our findings and identified key recurring themes.

5.2 Results

Overall, all of our participants ranked either ChatrEx-VINC (8/14) or the ChatrEx-VST (6/14) as the most explainable chatbot. As expected, participants found ChatrEx helpful in explaining the underlying chatbot's working and showing system status while processing the query. We next present users' perceptions of usefulness, transparency, and trust as they interacted with the different chatbots in our study.

5.2.1 Usefulness

As shown in Figure 5.2.a, users found the visual explanations by ChatrEx-VST (12/14) and ChatrEx-VINC (11/14) to be more useful than KEYHT (8/14) and BASELINE (0/14) and these differences in perceived usefulness were significant ($\chi^2(6, N=56) = 51.51$, $p<0.0001$). Participants' comments indicated that ChatrEx's clearly illustrated in-context visual representations were "more clear than words" and more "intuitive" in providing comprehensible information about the chatbot's understanding (competencies and breakdown) for their queries. For example, one participant commented:

> *"It [ChatrEx-VINC] understood what I [user] meant...showing the pictures of what it looks like in the spreadsheet, made really clear more than words, e.g. I said Column C and that's exactly Column C (P09)."*

Likewise, ChatrEx-VST's step-by-step tour highlighting the visual representations directly on the application was particularly useful for locating specific functions corresponding to the query:

> *"Highlighting the menubars and data columns in the worksheet itself makes the chatbot [ChatrEx-VST] looks more organic because that is also what a human would do, so I can relate to how it thinks (P10)."*

For ChatrEx-VINC, participants found it useful to have the instructions condensed within the chat window and felt that they had more freedom to go back-and-forth between the application UI and the chatbot UI. In contrast to ChatrEx-VST where participants said the overlay and visual tour "took over" the screen, ChatrEx-VINC offered more recognition than recall as the instructions could be used as a reference within the same screen:

> *"I liked [that] it[ChatrEx-VINC] was kept within the chat window...I could scroll back to the top to see what exactly I have said in case I needed to recall any information. The bot [ChatrEx-VINC] didn't expect me to remember it all. All the information just stayed there for me (P09)".*

> *"In ChatrEx-VINC, it feels like here is all the information and you can do what you will. The information is there but it's on me to take action" (P10)*

As expected, (9/14) participants were frustrated (Figure 5.2.b) to see the web links offered by BASELINE in response to a breakdown and did not find it useful: considered

them frustrating: *"It [Baseline] was a lot more frustrating because it didn't tell me anything. Definitely less useful because if I just wanted to Google, I would have done it myself (P04)."* Notably, the participants reported that they struggled in understanding the system status with respect to processing of the query by the chatbot. *"[With BASELINE] I really did not know what was happening at any point in the query" (P04).* The web links provided were considered irrelevant which annoyed users to spend their time and efforts in finding ways to approach the chatbot such that it can understand the query. *"This chatbot [BASELINE] was kinda condescending because it was like, 'I[BASELINE] do not know'. I[user] do not like the wall" (P10).*

Chatbot Type	Disagree	Neutral	Agree	Sparkline
ChatrEx-VST	7%	7%	86%	
ChatrEx-VINC	0%	21%	79%	
KEYHT	7%	36%	57%	
BASELINE	100%	0%	0%	

(a) The explanations/prompts I received from the chatbot were useful

Chatbot Type	Disagree	Neutral	Agree	Sparkline
ChatrEx-VST	57%	36%	7%	
ChatrEx-VINC	79%	14%	7%	
KEYHT	36%	36%	29%	
BASELINE	7%	29%	64%	

(b) My experience with using the chatbot was frustrating

Figure 5.2: Study results for "Usefulness" of explanations in each chatbot interface measured by prompts (a) and (b). Participants rated these prompts on a 5-point Likert scale ranging from Strongly Disagree (Rating 1) to Strongly Agree (Rating 5). In the above figures, Strongly Agree and Agree responses are added together and labelled as Agree. Similarly, Strongly Disagree and Disagree are clubbed and labelled as Disagree.

Participants mentioned that KEYHT was somewhat useful in that the chatbot acknowledged *where* it went wrong, but it was not exactly clear *why* the breakdown occurred. KEYHT's compressed and verbose explanations lacking the suggestions were perceived as "vague", less indicative of how to solve the problem: *"KEYHT gives rough areas of the problem...didn't give me any suggestion showing [a] gap between me and the software's [process] (P12)." "The message talking about the green and orange text was extra and irrelevant (P06)"*

5.2.2 Transparency

To assess users' perceptions of transparency, we considered how well the users were able to (i) understand how the chatbot works (Figure 5.3.a), (ii) follow the reasons explained by the chatbot during a breakdown (Figure 5.3.b) and, (iii) whether or not the users knew how to take the next step to recover from a breakdown (Figure 5.3.c).

In terms of understanding how the chatbot works (as shown in Figure 5.3.a), all users ranked ChatrEx-VST (14/14) as their first choice, followed by ChatrEx-VINC(11/14), KEYHT (8/14), BASELINE (4/14). These differences between explanation type and users' perceptions of how the chatbot works were significant (χ^2(6, N=56) = 17.72, p=0.0070). Further reflecting upon ChatrEx's explanation, many participants felt they had a better understanding of how the chatbot processed their query in a step-by-step manner. *"Chatbot VST and VINC are very clear and don't have any room for interpretation (P09)."* All of the participants found ChatrEx-VST to be intuitive and indicated that the visual step-by-step tour showed them exactly how the chatbot processed their query in the application. The aspect of the explanation design informing the competencies supported by the visual feedback highlighted in green made ChatrEx-VST appear to be smart enough to figure out what participants were looking for. For example, participant's comment for ChatrEx-VST:

> *"It [VST] was able to recognize what section (e.g., file, edit, etc.) to go make a chart. It [VST] recognized columns, rows and all information on the actual spreadsheet (P03)."*

In terms of helping users recognize the reasons for a breakdown (as shown in Figure 5.3.b), users ranked ChatrEx-VINC (13/14) and ChatrEx-VST (12/14) as being more helpful than KEYHT (9/14) and BASELINE(4/14). These differences were significant (χ^2(6, N=56) = 20.76, p=0.0020). Participants comments' indicated that in both designs of ChatrEx, the red highlights and corresponding comparative explanations supported by visual examples helped them to know where and why the failure occurred for both disambiguation and infeasible tasks:

Chatbot Type	Disagree	Neutral	Agree	Sparkline
ChatrEx-VST	0%	0%	100%	
ChatrEx-VINC	14%	7%	79%	
KEYHT	36%	7%	57%	
BASELINE	57%	14%	29%	

(a) Based on chatbot's explanations, I understood how the chatbot works

Chatbot Type	Disagree	Neutral	Agree	Sparkline
ChatrEx-VST	7%	7%	86%	
ChatrEx-VINC	0%	7%	93%	
KEYHT	21%	14%	64%	
BASELINE	64%	7%	29%	

(b) I was able to recognize the chatbot's reason for the breakdown.

Chatbot Type	Disagree	Neutral	Agree	Sparkline
ChatrEx-VST	7%	14%	79%	
ChatrEx-VINC	7%	29%	64%	
KEYHT	36%	36%	29%	
BASELINE	43%	29%	29%	

(c) Based on chatbot's explanations, I would be able to improve my query next time.

Figure 5.3: Study results for "Transparency" of explanations in each chatbot interface measured by prompts (a), (b) and (c). Participants rated these prompts on a 5-point Likert scale ranging from Strongly Disagree (Rating 1) to Strongly Agree (Rating 5). In the above figures, Strongly Agree and Agree responses are added together and labelled as Agree. Similarly, Strongly Disagree and Disagree are clubbed and labelled as Disagree.

"It [VST] failed the first [disambiguation task] time because there were multiple past functions that were used for the same query. The second time [infeasible task], it failed because it wasn't capable of performing the mean difference (P02)."

As shown in Figure 5.3.c, users ranked ChatrEx-VST (11/14) and ChatrEx-VINC (9/14) higher than KEYHT (4/14) and BASELINE (4/14) in helping them to take the appropriate next step for breakdown recovery. These differences were significant ($\chi^2(6, N=56) = 13.08$, $p=0.0418$). Participants expressed that exactly pinpointing the problem in ChatrEx-VST and ChatrEx-VINC was a lot more usable. *"Instead of redoing the whole query, I just need to know what part I need to change. Because it[ChatrEx] clearly shows me what parts are understood and what parts are not" (P06).* Further, participants mentioned that the alternative or similar function list and match percentages (Figure 4.5.e) served as helpful cues to see the relevant functions and improve their query:

"Because it's 45% match and return values of normal distribution function...[it's] something I want to accomplish, so I will probably use NORM.DIST command as the function name (P11)".

ChatrEx's explanations also helped users to understand infeasible tasks that the chatbot was not programmed to perform due to lack of training and that they could explore alternatives:

"...the available functions list gave me a hint on what is/isn't available on Google Sheets and I realize that I asked it to execute or run a nonexistent function(P10)".

Although many participants (9/14) could get some indication of the breakdowns with KEYHT's highlights, they struggled to understand "why" the problem occurred: *"It [KEYHT] gives me a rough idea, but [it's] not clear enough...I [had to] guess on why it failed to understand (P09)."* Further, KEYHT did not provide a suggestive list of functions to help users specifically who are not familiar with the spreadsheet. Thus, it became guesswork for the users to keep on trying to feel confident. *"I could recognize the reason but I had to guess more to know exactly what I needed to put in there [rephrased query] (P09)."*

Likewise, 9/14 participants failed to recognize the chatbot's working and problems in the query leading to breakdown using BASELINE which did not give much feedback to them. *"I don't know why & where chatbot [BASELINE] can't understand me (P12)."* The participants considered BASELINE as a black-box, wherein they were unaware about the understanding of the chatbot as well as problems in the query leading to breakdown: *"BASELINE is just*

like a Black-box, you don't know what they are thinking, you do not know what is the problem and where the problem is" (P14). Moreover, BASELINE indulged them into approaching the brute force technique such as trial and error. *"I don't have any information about where the breakdown happens and I wouldn't be able to rephrase it all. I will be just doing the trial and error on the whole query (P11)."*

Since KEYHT and BASELINE overall did not provide any guidance on how to resolve the breakdown, participants felt that they would mostly resort to "trial and error" to improve their queries. In contrast, participants overall agreed that ChatrEx provided transparent explanations for the high-level underlying working of the chatbot. *"Definitely understood how it (ChatrEx) works and what I needed to do because of all these visuals that made it clear" (P10).* In fact, many participants showed their interest in using ChatrEx-VST and Chatr VINC only to learn about the underlying working of the chatbot with respect to their query (even when they don't experience a breakdown). From the ChatrEx explanations, the participants primarily showed their interest in understanding the part of the process which the chatbot completed (competencies) and the part with the problem (breakdown reason) for improving their query.

5.2.3 Trust

As shown in Figure 5.4, users ranked ChatrEx-VINC(7/14) and ChatrEx-VST(6/14) as more trustworthy than KEYHT(1/14) and BASELINE (0/14). This difference between explanation type and users' perceptions of trust was significant ($\chi^2(6, N=56) = 29.43$, $p<0.0001$). A recurring sentiment among participants was that the visual feedback and explanation from ChatrEx designs gave them more confidence about how the chatbot works and they could trust it more for their task:

> *"I trust the mechanism of VINC and VST...I would probably rely on that a bit better just because it at least explains and provides the suggestions I could use" (P06).*

> *"I trust this chatbot (ChatrEx) because it shows me in green what it understands and highlights the important aspects of the query, and also tells me what's wrong by highlighting parts of my query in red" (P03).*

Users also appreciated seeing visual confirmations directly within the application interface as they did not have to struggle to find an appropriate mapping on their own: *"I would trust CharEx-VST because...it was really highlighting the column right where it is on the*

Chatbot Type	Disagree	Neutral	Agree	Sparkline
ChatrEx-VST	21%	0%	79%	
ChatrEx-VINC	7%	7%	86%	
KEYHT	50%	14%	36%	
BASELINE	93%	7%	0%	

(a) Based on the chatbot's explanations, I would trust the chatbot to finish my task.

Figure 5.4: Study results for "Trust" for explanations in each chatbot interface measured by prompt (a). Participants rated these prompts on a 5-point Likert scale ranging from Strongly Disagree (Rating 1) to Strongly Agree (Rating 5). In the above figures, Strongly Agree and Agree responses are added together and labelled as Agree. Similarly, Strongly Disagree and Disagree are clubbed and labelled as Disagree.

worksheet...as a user I kind of recognized the location (P10)." This allowed participants to place their trust in the chatbot as at least it was trying to understand them and help them. In contrast, since the majority of participants failed to figure out the breakdown reason with KEYHT and BASELINE, they were hesitant to trust them.

Overall, we found that since participants could trust ChatrEx, they were more enthusiastic about using these chatbots for their future spreadsheet tasks. Even beyond spreadsheets, many participants expressed interest in seeking explanations using ChatrEx in other complex applications and some even said that they would enjoy the experience:

> *"It's kind [of] like a pair programming with the bot. It's nice to have something to bounce ideas back and gather information from within the [ChatrEx] bot instead [of] Google search (P09)."*

5.3 Suggestions for Improvement

Despite the overall positive findings for ChatrEx, participants identified some minor areas of improvement which could help these chatbot designs be adopted more widely. For example, with ChatrEx-VST, four participants mentioned that they wanted a more graceful exit from the tour mode so they could have better recall for the function to improve the query. *"the functions went away by the time it[ChatrEx-VST] ended the tour" (P04).* Two

participants who were beginners in using Google Sheets found a step-by-step explanation tour by ChatrEx-VST guided them through each step which might be less required for those who are Excel expert users. *"I am not super adept at Excel. I like that it[ChatrEx-VST] really guided me like holding my hand. If I was super adept, I would say stop holding my hand I know how to do this" (P10).* On the contrary, with ChatrEx-VINC, few participants reported that one of the difficulties they experienced was locating the functions on the software application similar to ChatrEx-VST. Despite these shortcomings, ChatrEx carries the potential to enhance users' mental model for AI systems such as chatbots in terms of usefulness, transparency, and trust.

Overall, participants showed their enthusiasm and interest in seeking explanations using ChatrEx and expressed their future utility among companies and real-world tasks. ChatrEx could be a step forward in providing the positive user experience and seems a bit close in improving the trust of the users for the chatbot:

> *"Chatbots [ChatrEx] in this study were felt more capable of performing more complicated tasks. Showed it's [ChatrEx's] capabilities in actually trying to figure out the breakdown and come up with a solution for every part of the problem" (P10).*

> *"I can trust it(ChatrEx) more because it tells me what I am doing. It doesn't just do it behind the scenes, it tells me what it is doing, it tells me what it understands for my sentence" (P03).*

5.4 Summary

In this chapter, we presented the evaluation and comparison of ChatrEx web-based prototypes with two other types of chatbots (one baseline and one state-of-the-art chatbot closet to our work) via an observational user study. The key findings obtained from this evaluation provided us with insights that the visual step-by-step explanations within ChatrEx-VST and ChatrEx-VINC were perceived as transparent and easier to understand. Most participants appreciated that ChatrEx offered better UX to distinguish competencies and breakdown. Overall, ChatrEX-VST and ChatrEx-VINC outranked KEYHT and BASELINE in terms of usefulness, transparency, and trust.

Chapter 6

Discussion

In this thesis, we have contributed the design and evaluation of two novel explainable chatbot interfaces (ChatrEx-VINC and ChatrEx-VST) that visually explain a chatbot's underlying functionality and decisions during a breakdown. Our findings indicate that users perceived both ChatrEx designs to be more useful, transparent, and trustworthy compared to explanations provided using verbal keyword highlighting [5] and traditional chatbots that provide no explicit explanations. More broadly, our work offers insights into the strengths and weakness of each explanation design in terms of usefulness, transparency, and trust and thus informs how to design user-centered explainable solutions for in-application task-oriented chatbots.

In this chapter, we now reflect on our key insights from this research, its limitations, and highlight opportunities for future research to design explainable chatbots from an HCI perspective.

6.1 Limitations

Our key focus was on developing a minimum viable implementation for explainable chatbot interfaces that simulates the breakdown and allows us to investigate their strengths and weaknesses towards enhancing transparency, trust and usefulness. Although our proof-of-concept interactive prototypes were useful for assessing users' initial perceptions and reactions when using explainable chatbots, additional work would be needed to fully understand how users would interact with ChatrEx's explanations with a more sophisticated implementation of the underlying NLP/ML-based algorithms and how users would use these chatbots for their own tasks in live deployments.

One limitation of the studies in this research is that we compared our explanation designs with limited repair strategies in the context of enhancing transparency and trust. Future work should investigate the relevance of our findings for other types of repair strategies. In addition, future studies should consider a larger number of participants, more diverse tasks, and perhaps some more varied successful and unsuccessful chatbox dialogue for the evaluation of these chatbot interfaces. In this study, we focused on only explaining the conversational breakdowns between a user and a chatbot. But, we realize that there are many other types of breakdowns that users may experience with feature-rich software (e.g., UI breakdowns, bugs) and it would be worth exploring how our explainable chatbot designs could be extended to support recovery from such software breakdowns. Lastly, we acknowledge that our exploration of one domain is a limitation. Although our implementation focused on supporting spreadsheet queries, the general design of our visual explanations can be adapted to any feature-rich application that allows clear and distinct one-to-one mappings between intents/entities and GUI interfaces and components. Nonetheless, our study is a starting point towards designing user-centered explainable solutions for in-application task-oriented chatbots and opens up several promising research directions for further enhancing users' perception of transparency and trust for these chatbots.

6.2 Future Work: Leveraging Explainable AI for breakdown recovery

Our research provides initial evidence that it can be useful for users to see *where* a breakdown occurred and *what caused* the breakdown when they are working with in-application chatbots. In particular, we demonstrated that by leveraging XAI approaches and designing explanations that provide visual guidance in-context of the UI, users can better understand the capabilities of the chatbot and how they could improve their interaction by rephrasing their queries. Even for tasks that were infeasible for the chatbot to perform, users still found it helpful to learn about the chatbot's limitations instead of wasting time and effort in using trial-and-error strategies. Interestingly, one participant expressed the desire to have an explanation option not only during the breakdown but also when the chatbot understands the user's query to provide confidence that the task was completed properly.

Given the promise and importance of XAI solutions explored in other contexts [11, 66], future chatbots should incorporate similar strategies to ChatrEx that allow people to learn how chatbots work and understand breakdown reasons. Our chatbot explanations can serve as a "teaching moment" for users to better understand where and why things went wrong. Instead of focusing on algorithmic-level explanations of the chatbot's functionality, it may be more important to focus on explaining the application UI-level functionality so that

users who do not know anything about ML can still find the chatbot to be transparent and trustworthy.

6.3 Future Work: Designing a hybrid of visual tour and non-tour mode

Both ChatrEx-VST and ChatrEx-VINC exhibit some unique strengths through their explanation designs. Although in this study we did not focus on the users having varying familiarity with spreadsheet GUIs, we observed that the more experienced users of spreadsheets considered ChatrEx-VINC's condensed within-chat explanations to be more useful. ChatrEx-VINC provided users with more control and freedom to access the information when required and allowed them to try to improve their query without leaving the screen. On the other hand, the tour mode of ChatrEx-VST that highlighted each step directly on the application UI was more intuitive for the less experienced users and helped them become aware of unfamiliar functions. One participant described the step-by-step feature of VST as if somebody was "holding their hand" in helping them work through a breakdown.

Feature-rich applications such as Google Sheets support many complicated tasks, so it is likely that even experienced users may be unfamiliar with several of the spreadsheet's other features and could benefit from designs such as VST. Future chatbots should support the strengths of both ChatrEx-VST and ChatrEx-VINC and allow users to toggle between the 'tour mode' and the 'non-tour mode' based on their requirements to leverage the benefits of each approach. We have so far in this study evaluated our proposed novel proof-of-concept designs with few state-of-the-art chatbots, however, there is an emerging class of repair models and dialog-based systems to automate these tasks. In the future, there is a rich opportunity in exploring these explanation designs with more sophisticated dialog-based chatbot implementations and, further, extending this work to explore how explanation designs can be incorporated in more chatbots with intent-based models that focus on offering alternatives and repair strategies for addressing conversational breakdowns.

Our main goal was to illustrate how a chatbot works (especially during breakdowns) because prior works [5, 42] have shown that most users usually lack a reasonable mental model of a chatbot's underlying functionality (it's a blackbox) and what it can or cannot do. Consequently, inspired by the "research through design" [67] paradigm, we contributed in developing a minimal viable implementation of a chatbot that allowed us to explore how visual explanations should be designed. Through our implementation we were able to augment the chatbot functionality with different visual in-context explanations that could support a range of infeasible and disambiguation tasks in the user study and assess them

with users, however, in the future, there is still more work that needs to be done at the intersection of ML and HCI to build upon recent work [37]. Nevertheless, our designs could be used to map user intents to specific portions of GUIs and interaction examples from other users and therefore could be adapted to other feature- rich applications besides spreadsheets that have similar UIs and menu structures.

6.4 Future Work: Empirically understanding human-chatbot interaction

With the rapid innovations in the field of AI, there is more need for HCI-oriented research that actually tries to understand human behavior and user perceptions of AI solutions, such as task-oriented chatbots [27, 32]. Our study provides various insights into how to make these chatbots more transparent (and increase user trust) with the use of explainable solutions. The key lies in leveraging visual explainable designs to enhance users' mental models of these chatbots, particularly targeting situations of breakdowns. Still, there are many opportunities for future research to investigate other automated ways to increase transparency (and make users more aware of why the system may be stuck) and make these black boxes AI systems more comprehensible.

The results from our study also showed that most of our participants who had little to no experience in ML were still able to understand the visual step-by-step explanations and found them to be useful. While much of the early focus of explainable AI solutions has been on algorithms, [66, 11], we decided to focus on designing more high-level visual example-based explanations. We believe that such explanations can be a starting point for further understanding and improving human-chatbot interaction, particularly for complex, application-specific chatbots. Lastly, we observed many interesting individual differences for the preference of explanation designs for novices and experts of the spreadsheet application. In the future, it would be a great idea to explore how our explanations would be perceived by a larger number of users in a field study in the context of even more diverse tasks and breakdowns (especially if they are chained together). Future studies could also investigate further the explanation needs for users having a different level of expertise with the underlying application.

6.5 Future Work: Exploring the potential of Explainable AI in enhancing ML learnability among end users

Previous user studies in the domain of explainable AI [34, 41] assume that users have advance experience in ML to understand the in-depth algorithm-specific explanations. Consequently, novice ML users struggle to understand the machine learning behind these AI systems [34, 59]. So far, in the chatbot domain, few studies have focused on designing Explainable AI solutions but it is yet not clear whether these explanations could be adopted by non-ML experts who can also be end users for many AI applications[5, 30]. Our study considered exploring the design solutions to target this sector of users. In fact, our findings showed that users who have little to no experience in ML found user-centered explainable solutions to be more comprehensible and usable to accomplish their tasks. Furthermore, our work shares the goal of enhancing users' mental models of AI systems in terms of transparency and trust by attempting to make these black boxes AI systems comprehensible. Future studies could investigate further the explanation needs for users having a different level of ML expertise.

Our study with ChatrEx-VST and ChatrEx-VINC revealed the empirical insights that continued to confirm the importance of looking at much more capable user-centered Explainable AI in the domain of application-specific chatbots. Through this research, we took an opportunity to innovate design solutions providing user-centered software help for improving the user interaction with task-oriented chatbots embedded within feature-rich applications such as Google Sheets. Our design solutions could improve the ML learnability for understanding the underlying working within the domain of task-oriented chatbots. With the growing interest in AI systems, the need for learning about ML concepts among the general population is increasing. However, learning resources such as online encyclopedias, textbooks, and articles are often rich in technical jargon and can be challenging to grasp. The positive user response from our visual explanation design solutions (ChatrEx-VST and ChatrEx-VINC) presents an interesting future direction to expand these solutions for educating the end users with ML concepts. Future studies could explore ways in which this approach could be used to innovate curated learning techniques for enhancing the ML learnability among end users. Specifically, they can benefit from a more guided and tutorial approach like ChatrEx-VST to learn the ML technical concepts using explainable chatbots.

Chapter 7

Conclusion

In this thesis, we have explored the design space of in-application explainable chatbot interfaces and contributed two novel designs of ChatrEx that provide visual example-based step-by-step explanations to illustrate the underlying working of a chatbot during a conversational breakdown. Our design goals derived from recent literature provide insights into the requirements for structuring and designing the explanations for in-application chatbot interfaces. The iterative design and prototyping approaches allowed us to design and evaluate low and medium fidelity prototypes for two variations of ChatrEx (ChatrEx-VINC and ChatrEx-VST) across multiple ideation phases. ChatrEx-VINC provides visual example-based step-by-step explanations in-context of the chat window whereas ChatrEx-VST provides explanations as a visual tour overlaid on the application interface. Our brief formative study that we ran informally using medium fidelity Axure-based prototypes elicited the initial feedback for the designs and led us to the final implementation of these two variations for complex spreadsheet tasks. Our comparative observational study using novel web-based chatbot interfaces of ChatrEx-VINC and ChatrEx-VST and two other implementations (i.e., KEYHT and BASELINE for comparison) shows that the explanations provided by both ChatrEx-VINC and ChatrEx-VST enhanced users' understanding of the reasons for a conversational breakdown and improved users' perceptions of usefulness, transparency, and trust. Users found these explanations of a chatbot's competencies and reasons for breakdown to be useful, transparent, and trustworthy. Our empirical findings have several implications and potential directions for future work, such as leveraging and adapting Explainable AI solutions to design in-application explainable chatbots and improve overall human-chatbot interaction.

Bibliography

[1] lxieyang/chrome-extension-boilerplate-react.

[2] Why Chatbots Fail: Limitations of Chatbots, 2019.

[3] 8 Epic Chatbot / Conversational Bot Failures [2020 update], June 2020.

[4] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. Guidelines for human-ai interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–13, New York, NY, USA, 2019. Association for Computing Machinery.

[5] Zahra Ashktorab, Mohit Jain, Q. Vera Liao, and Justin D. Weisz. Resilient chatbots: Repair strategy preferences for conversational breakdowns. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–12, New York, NY, USA, 2019. Association for Computing Machinery.

[6] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82 – 115, 2020.

[7] Nancy Baym, Limor Shifman, Christopher Persaud, and Kelly Wagman. Intelligent failures: Clippy memes and the limits of digital assistants. *AoIR Selected Papers of Internet Research*, 2019, Oct. 2019.

[8] Petter Bae Brandtzaeg and Asbjørn Følstad. Why people use chatbots. In Ioannis Kompatsiaris, Jonathan Cave, Anna Satsiou, Georg Carle, Antonella Passani, Efstratios Kontopoulos, Sotiris Diplaris, and Donald McMillan, editors, *Internet Science*, pages 377–392, Cham, 2017. Springer International Publishing.

[9] Petter Bae Brandtzaeg and Asbjørn Følstad. Chatbots: Changing user needs and motivations. *Interactions*, 25(5):38–43, August 2018.

[10] Andrea Bunt, Matthew Lount, and Catherine Lauzon. Are explanations always important? a study of deployed, low-cost intelligent interactive systems. In *Proceedings of the 2012 ACM International Conference on Intelligent User Interfaces*, IUI '12, page 169–178, New York, NY, USA, 2012. Association for Computing Machinery.

[11] Carrie J. Cai, Jonas Jongejan, and Jess Holbrook. The effects of example-based explanations in a machine learning interface. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, IUI '19, page 258–262, New York, NY, USA, 2019. Association for Computing Machinery.

[12] Hao-Fei Cheng, Ruotong Wang, Zheng Zhang, Fiona O'Connell, Terrance Gray, F. Maxwell Harper, and Haiyi Zhu. Explaining decision-making algorithms through ui: Strategies to help non-expert stakeholders. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–12, New York, NY, USA, 2019. Association for Computing Machinery.

[13] Janghee Cho and Emilee Rader. The role of conversational grounding in supporting symbiosis between people and digital assistants. *Proc. ACM Hum.-Comput. Interact.*, 4(CSCW1), May 2020.

[14] William J. Clancey. Notes on "epistemology of a rule-based expert system". *Artificial Intelligence*, 59(1-2):191–204, 1993.

[15] Matteo Colombo, Leandra Bucher, and Jan Sprenger. Determinants of judgments of explanatory power: Credibility, generality, and statistical relevance. *Frontiers in Psychology*, 8:1430, 2017.

[16] Sven Coppers, Jan Van den Bergh, Kris Luyten, Karin Coninx, Iulianna van der Lek-Ciudin, Tom Vanallemeersch, and Vincent Vandeghinste. Intellingo: An intelligible translation environment. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, page 1–13, New York, NY, USA, 2018. Association for Computing Machinery.

[17] J. Corbin and A. Strauss. Grounded theory research: Procedures, canons, and evaluative criteria. *Qualitative Sociology*, 13:3–21, 1990.

[18] Justin Cranshaw, Emad Elwany, Todd Newman, Rafal Kocielnik, Bowen Yu, Sandeep Soni, Jaime Teevan, and Andrés Monroy-Hernández. Calendar.help: Designing a workflow-based scheduling agent with humans in the loop. *ACM CHI Conference on Human Factors in Computing Systems*, January 2017.

[19] Aritra Dasgupta, Joon-Yong Lee, Ryan Wilson, Robert Lafrance, Nick Cramer, Kristin Cook, and Samuel Payne. Familiarity vs trust: A comparative study of domain scientists' trust in visual analytics and conventional analysis methods. *IEEE Transactions on Visualization and Computer Graphics*, 23:1–1, 08 2016.

[20] Nicholas Diakopoulos. Algorithmic accountability. *Digital Journalism*, 3(3):398–415, 2015.

[21] Berkeley J. Dietvorst, J. Simmons, and C. Massey. Algorithm aversion: people erroneously avoid algorithms after seeing them err. *Journal of experimental psychology. General*, 144 1:114–26, 2015.

[22] Brian Feldman. Clippy Didn't Just Annoy You — He Changed the World, October 2016.

[23] Boris Galitsky. *Chatbot Components and Architectures*, pages 13–51. Springer International Publishing, Cham, 2019.

[24] Karen Gilchrist. Chatbots expected to cut business costs by $8 billion by 2022, May 2017.

[25] Alyssa Glass, Deborah L. McGuinness, and Michael Wolverton. Toward establishing trust in adaptive agents. In *Proceedings of the 13th International Conference on Intelligent User Interfaces*, IUI '08, page 227–236, New York, NY, USA, 2008. Association for Computing Machinery.

[26] Jonathan Grudin and Richard Jacques. Chatbots, humbots, and the quest for artificial general intelligence. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–11, New York, NY, USA, 2019. Association for Computing Machinery.

[27] Jonathan Grudin and Richard Jacques. Chatbots, humbots, and the quest for artificial general intelligence. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–11, New York, NY, USA, 2019. Association for Computing Machinery.

[28] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):93, 2018.

[29] David Gunning, Mark Stefik, Jaesik Choi, Timothy Miller, Simone Stumpf, and Guang-Zhong Yang. Xai—explainable artificial intelligence. *Science Robotics*, 4(37), 2019.

[30] Mohit Jain, Ramachandra Kota, Pratyush Kumar, and Shwetak N. Patel. 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, page 1–6, New York, NY, USA, 2018. Association for Computing Machinery.

[31] Mohit Jain, Pratyush Kumar, Ramachandra Kota, and Shwetak N. Patel. Evaluating and informing the design of chatbots. In *Proceedings of the 2018 Designing Interactive Systems Conference*, DIS '18, page 895–906, New York, NY, USA, 2018. Association for Computing Machinery.

[32] Danica Jovic. The Future is Now – 37 Fascinating Chatbot Statistics, August 2020.

[33] René F. Kizilcec. How much information? effects of transparency on trust in an algorithmic interface. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, page 2390–2395, New York, NY, USA, 2016. Association for Computing Machinery.

[34] Josua Krause, Adam Perer, and Enrico Bertini. A user study on the effect of aggregating explanations for interpreting machine learning models. *In Proceedings of KDD 2018 Workshop on Interactive Data Exploration and Analytics (IDEA'18)*, Aug 2018.

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

[36] T. Li, I. Labutov, X. Li, X. Zhang, W. Shi, W. Ding, T. M. Mitchell, and B. A. Myers. Appinite: A multi-modal interface for specifying data descriptions in programming by demonstration using natural language instructions. In *2018 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*, pages 105–114, Los Alamitos, CA, USA, 2018. IEEE Computer Society.

[37] Toby Jia-Jun Li, Jingya Chen, Haijun Xia, Tom M. Mitchell, and Brad A. Myers. Multi-modal repairs of conversational breakdowns in task-oriented dialogs. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*, UIST '20, page 1094–1107, New York, NY, USA, 2020. Association for Computing Machinery.

[38] Q. Vera Liao, Daniel Gruen, and Sarah Miller. Questioning the ai: Informing design practices for explainable ai user experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–15, New York, NY, USA, 2020. Association for Computing Machinery.

[39] Brian Y. Lim, Anind K. Dey, and Daniel Avrahami. Why and why not explanations improve the intelligibility of context-aware intelligent systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, page 2119–2128, New York, NY, USA, 2009. Association for Computing Machinery.

[40] Zachary C. Lipton. The Mythos of Model Interpretability. *arXiv e-prints*, page arXiv:1606.03490, June 2016.

[41] S. Liu, J. Xiao, J. Liu, X. Wang, J. Wu, and J. Zhu. Visual diagnosis of tree boosting methods. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):163–173, 2018.

[42] 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, page 5286–5297, New York, NY, USA, 2016. Association for Computing Machinery.

[43] A. Maedche, S. Morana, Silvia Schacht, D. Werth, and J. Krumeich. Advanced user assistance systems. *Business & Information Systems Engineering*, 58:367–370, 2016.

[44] Edoardo Maggio. Apple says that 500 million customers use Siri, January 2018.

[45] Robinson Meyer. Even early focus groups hated clippy, 2015.

[46] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *CoRR*, abs/1706.07269, 2017.

[47] Andy Mura. Why, How, And When To Use Walkthroughs To Enhance UX.

[48] Don Norman. The Design of Everyday Things. In *The Design of Everyday Things.*, page p.39. 2013.

[49] 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*, CHI '18, page 1–12, New York, NY, USA, 2018. Association for Computing Machinery.

[50] Foster Provost and David Martens. Explaining Data-Driven Document Classifications. June 2013.

[51] Filip Radlinski and Nick Craswell. A theoretical framework for conversational search. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*, CHIIR '17, page 117–126, New York, NY, USA, 2017. Association for Computing Machinery.

[52] S. Read and A. Marcus-Newhall. Explanatory coherence in social explanations: A parallel distributed processing account. 1993.

[53] A. Renkl, T. S. Hilbert, and S. Schworm. Example-based learning in heuristic domains: A cognitive load theory account. *Educational Psychology Review*, 21:67–78, 2009.

[54] Alexander Renkl. Toward an instructionally oriented theory of example-based learning. *Cognitive science*, 38, 09 2013.

[55] 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, page 1135–1144, New York, NY, USA, 2016. Association for Computing Machinery.

[56] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-precision model-agnostic explanations. In *AAAI*, volume 18, pages 1527–1535, 2018.

[57] Maria Rosala. Status Trackers and Progress Updates: 16 Design Guidelines, February 2019.

[58] Heleen Rutjes, Martijn Willemsen, and Wijnand IJsselsteijn. Considerations on explainable ai and users' mental models. In *Where is the Human? Bridging the Gap Between AI and HCI*, United States, May 2019. Association for Computing Machinery, Inc. CHI 2019 Workshop : Where is the Human? Bridging the Gap Between AI and HCI ; Conference date: 04-05-2019 Through 04-05-2019.

[59] S. Stumpf, Simonas Skrebe, Graeme Aymer, and Julie Hobson. Explaining smart heating systems to discourage fiddling with optimized behavior. In *IUI Workshops*, 2018.

[60] Jack Taylor. Is Microsoft Excel Hard To Learn?, 2019.

[61] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr, 2017.

[62] Joseph Weizenbaum. Eliza—a computer program for the study of natural language communication between man and machine. *Commun. ACM*, 9(1):36–45, January 1966.

[63] Jason Williams, Nobal B. Niraula, Pradeep Dasigi, Aparna Lakshmiratan, Carlos Garcia Jurado Suarez, Mouni Reddy, and Geoffrey Zweig. Rapidly scaling dialog systems with interactive learning. January 2015.

[64] Jun Xiao, John Stasko, Richard Catrambone, et al. An empirical study of the effect of agent competence on user performance and perception. In *AAMAS*, volume 4, pages 178–185. Citeseer, 2004.

[65] Wei Xu. Toward human-centered ai: A perspective from human-computer interaction. *Interactions*, 26(4):42–46, June 2019.

[66] Fumeng Yang, Zhuanyi Huang, Jean Scholtz, and Dustin L. Arendt. How do visual explanations foster end users' appropriate trust in machine learning? In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, IUI '20, page 189–201, New York, NY, USA, 2020. Association for Computing Machinery.

[67] John Zimmerman, Jodi Forlizzi, and Shelley Evenson. Research through design as a method for interaction design research in hci. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '07, page 493–502, New York, NY, USA, 2007. Association for Computing Machinery.

Appendix A

Study instruments information

In the appendix, we attached the questionnaire we used during the usability study and follow-up interview. Figure A.1 presents the post-task questionnaire used during the usability test for evaluating each chatbot interface.

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
1. My experience with using the **chatbot was frustrating.**	○	○	○	○	○
2. The explanations/prompts I received from the chatbot were **useful**.	○	○	○	○	○
3. I was able to **recognize the chatbot's reason** for the breakdown / conversational dead-end	○	○	○	○	○
4. Based on the chatbot's explanations/prompts, I understood **how the chatbot works**	○	○	○	○	○
5. Based on the chatbot's explanations/prompts I would be able to **improve my query next time**.	○	○	○	○	○
6. Based on the chatbot's explanations/prompts, I would **trust the chatbot** to finish my task.	○	○	○	○	○

Figure A.1: Post-task questionnaire for evaluating chatbot interfaces (ChatrEx-VINC, ChatrEx-VST, KEYHT, BASELINE)

Follow-Up Interview

1. Have you used chatbots before such as Siri, Cortana, Alexa etc? If yes, How do you trust these chatbots-your perception of trusting the chatbot?
2. How does the experience with chatbot prototypes you used in this study compare to previous experience with respect to chatbots you have used before?
3. Now you have interacted with all the four chatbots. Suppose, Chatbot E includes all these four chatbots that you have interacted with. You want to use any one of these four chatbots to see how you can create a chart showing comparison of data from January to February. You have four chatbot explanation options (A, B, C, D), choose one chatbot's explanation you want to use for this task query.

   For this task, please mention what chatbot would you use for this task and why?

4. Now that you have seen and interacted with 4 different types of chatbots' explanations, how would rank them in terms of their **explanations(EXPLAINABILITY)**: -such that Rank 1 indicates the chatbot which was strongly useful in recovering from breakdown and Rank 4 indicates not at all useful? **WHY**?
   a. Chatbot A Chatbot B Chatbot C Chatbot D
   b. Based on the different chatbot's explanations, which of these designs would you trust the most to finish your task and Why?
5. Among the designs that you saw, which types of explanations could help you in understanding the underlying AI/ Machine Learning working of the chatbot with respect to your query? Why?
6. (In general), Do you think these chatbots help you learn ML and how interested you would be in using them to learn Machine Learning? On a scale of 1 to 5 (5 is very likely). Why or why not?
7. To what extent would you be interested in using these prototypes/solutions to help with excel tasks in the future? On a scale of 1 to 5 (5 is very likely). Why or why not?
8. What would you like to see improved? (e.g., what would need to be improved before you would consider using the chatbot on a regular basis)?