University of Brasília

Institute of Exact Sciences
Department of Computer Science

Human Factors in the Design of Chatbot Interactions: Conversational Design Practices

Geovana Ramos Sousa Silva

Dissertation submitted in partial fullfilment of
the requirements for the Master's Degree in Informatics

Advisor
Prof.a Dr.a Edna Dias Canedo

Brasília
2023

University of Brasília

Institute of Exact Sciences
Department of Computer Science

Human Factors in the Design of Chatbot Interactions: Conversational Design Practices

Geovana Ramos Sousa Silva

Dissertation submitted in partial fullfilment of
the requirements for the Master's Degree in Informatics

Prof.a Dr.a Edna Dias Canedo (Advisor)
University of Brasília (UnB)

Prof.a Dr.a Mairieli Santos Wessel
Radboud University, The Netherlands

Prof.a Dr.a Ana Paula Chaves Steinmacher
Northern Arizona University, Flagstaff AZ

Prof. Dr. Ricardo Pezzuol Jacobi
Coordinator of the Postgraduate Program in Informatics

Brasília, January 31, 2023

Dedication

To all the women who were denied their right to get educated.

Acknowledgements

First and foremost, I thank God for giving me strength, wisdom, and the opportunity to conduct this work. “I can do all things through Christ who gives me strength.” (Philippians 4:13).

I would like to express my deepest love and gratitude to my parents, Luiz and Mariana, for their constant love and support and for encouraging me to undertake this wonderful journey. You were essential for me to finish it successfully.

I am deeply indebted to Prof. Edna Dias Canedo for being an excellent advisor and an incredible research partner. I greatly appreciate the trust you have placed in me and my work. I am grateful for the opportunities you have given me, and I hope to continue working with you for many years.

I would like to extend my sincere thanks to all volunteers participating in the survey and case study. This work would not be possible without your help.

Last but not least, I thank the colleagues and professors from my postgraduate program (PPGI), the Laboratory for Decision Making Technologies (LATITUDE), and the University of Brasília (UnB), who helped me and enabled me to grow as a professional.

The present work was carried out with the support of the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) through the access to the Journals Portal. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001.

Resumo

Contexto: Chatbots são agentes inteligentes que imitam o comportamento humano para conduzir conversas significativas. A natureza conversacional dos chatbots impõe desafios aos designers, uma vez que seu desenvolvimento é diferente de outros softwares e requer a investigação de novas práticas no contexto da interação humano-IA e seus impactos na experiência do usuário. Como o diálogo humano envolve diversas variáveis além da verbalização de palavras, é fundamental projetar diálogos bem pensados para que os chatbots proporcionem uma interação humanizada e otimizada. **Objetivo:** O principal objetivo deste trabalho é identificar práticas de design textual, visual ou interativo de interações de chatbots baseadas em texto e como elas podem potencializar ou enfraquecer algumas percepções e sentimentos dos usuários, como satisfação, engajamento e confiança, para a criação do guia Diretrizes para Design Conversacional de Chatbots (DDCC). **Método:** Utilizamos vários métodos de pesquisa para gerar e validar o guia. Primeiro, realizamos uma Revisão Sistemática da Literatura (RSL) para identificar as práticas de design conversacional e seus impactos. Essas práticas foram inseridas no guia DDCC por meio de análise qualitativa e codificação dos resultados RSL. Em seguida, o guia foi validado quantitativamente por meio de um *survey* e qualitativamente por meio de um estudo de caso. O *survey* teve como objetivo avaliar a clareza e a utilidade do guia baseado por meio da leitura do guia por parte dos participantes da pesquisa e nas suas respostas a um questionário adaptado do Modelo de Aceitação de Tecnologia. O estudo de caso teve como objetivo avaliar a utilidade do guia com base em sua aplicação prática pelos participantes em uma situação que simula um cenário real e em entrevistas de acompanhamento. **Resultados:** A pesquisa mostrou que desenvolvedores de software com diferentes níveis de experiência concordaram fortemente que o guia poderia induzir maior satisfação e engajamento no usuário. Além disso, eles também concordaram fortemente que o guia é claro, compreensível, flexível e fácil de usar. Embora os participantes tenham sugerido algumas melhorias, eles relataram que os principais pontos fortes do guia são a objetividade e a clareza. O estudo de caso confirmou os resultados da pesquisa, pois os participantes relataram sentimentos positivos em relação ao guia e uma intenção de usá-lo. Suas extensas percepções fornecidas por meio das entrevistas realizadas revelaram que

suas experiências anteriores com chatbots e em cargos específicos de desenvolvimento de software influenciaram seu design e adoção de práticas. **Conclusão:** O guia se mostrou útil para desenvolvedores com diferentes níveis de conhecimento, com potencial para se tornar um forte aliado dos desenvolvedores no processo de design conversacional.

Palavras-chave: Chatbot, design conversacional, interação humano-AI, fatores humanos

Resumo Expandido

Fatores Humanos no Design de Interações de Chatbot: Práticas de Design Conversacional

Introdução

Sistemas baseados em Inteligência Artificial (AI) estão ultrapassando a barreira acadêmica para serem cada vez mais utilizados, principalmente para agilizar a prestação de serviços ao usuário [1] e cada vez mais aceitos pelos usuários [2]. Um tipo de sistema baseado em AI que ganhou espaço em vários setores é o *chatbot*. Os chatbots são agentes inteligentes alimentados por algoritmos de aprendizado de máquina para imitar o comportamento humano [3], e os usuários tendem a recorrer para eles por causa da facilidade, velocidade e conveniência de conversar com um chatbot [4].

Como a tecnologia voltada para agentes de conversação está cada vez mais confiável e eficiente, os chatbots estão se tornando cada vez mais parte do dia a dia das pessoas comuns e há registros de seu uso para fins variados. Considerando a crescente presença de chatbots na vanguarda de diversos setores, é necessário não só investir em algoritmos de compreensão natural, mas também dedicar tempo e esforço para proporcionar interações agradáveis de chatbot para os usuários. Portanto, o design conversacional é uma parte essencial do processo de desenvolvimento do chatbot, que consiste em visualizar e especificar o fluxo do diálogo. Um design conversacional bem pensado fortalece a realização dos objetivos do chatbot em todos os contextos por meio de uma colaboração mútua entre chatbot e usuário. Essa colaboração é essencial para obter aos usuários o que eles esperam obter, pois os chatbots não podem prever o que os usuários desejam. Portanto, um design conversacional bem pensado tranquiliza e sensibiliza o usuário a colaborar com o diálogo, conduzindo-o ao seu objetivo e consequente satisfação.

Nesse contexto, é necessário investigar os requisitos conversacionais do chatbot que sejam independentes de tecnologia, centrados no usuário e focados em alcançar objetivos de negócios por meio de conversas. Assim, neste trabalho, realizamos um estudo das práticas de design conversacional de chatbots e como elas impactam os usuários. O principal objetivo deste trabalho é identificar práticas de design textual, visual ou interativo de interações de chatbots baseadas em texto e como elas podem potencializar ou enfraquecer

algumas percepções e sentimentos dos usuários, como satisfação, engajamento e confiança, para a criação do guia Diretrizes para Design Conversacional de Chatbots (DDCC).

Metodologia

Esta pesquisa está estruturada em 6 fases: 1) condução de uma Revisão Sistemática de Literatura (RSL) para identificação de práticas conversacionais e seus impactos nos usuários; 2) desenvolvimento do guia DDCC versão 1.0 aplicando Grounded Theory nos dados extraídos da RSL; 3) validação do guia DDCC versão 1.0 por meio de um survey baseado no questionário do Modelo de Aceitação de Tecnologia; 4) proposta do guia DDCC versão 2.0; 5) validação do guia DDCC versão 2.0 por meio de um estudo de caso, no qual os participantes irão usar o guia para melhorar conversas; 6) proposta do guia DDCC versão 3.0, a versão final.

Sendo assim, a construção do guia será feita com base nos resultados da RSL e sua validação e aplicação de melhorias serão feitas por meio de um *survey* e um estudo de caso. O *survey* visa avaliar quantitativamente a utilidade e facilidade de uso da primeira versão do guia em relação à sua leitura por potenciais designers de conversas de chatbots no contexto do design de conversas para chatbots. O estudo de caso visa avaliar qualitativamente a utilidade e facilidade de uso da segunda versão do guia em relação a seu uso por potenciais designers de chatbot de conversação no contexto de design de conversação para chatbots.

Resultados e Discussão

A SLR retornou um total de 1101 artigos. Após a aplicação do protocolo, selecionamos 40 estudos primários de diferentes contextos e com várias práticas sendo testadas com usuários para avaliar como eles se sentem sobre a presença ou ausência dessas práticas. Esses estudos revelaram um esforço significativo em tornar o chatbot mais humano com recursos antropomórficos e adequar esses recursos à melhor configuração, como testar alguns traços de personalidade. Além disso, houve tentativas de facilitar a comunicação com elementos interativos e tornar a conversa mais transparente com abertura e esclarecimento. As práticas conversacionais coletadas tiveram um impacto geral positivo nos usuários, mas muitas delas têm variáveis moderadoras que são difíceis de evitar por serem inerentes aos usuários.

A análise conjunta dos estudos primários selecionados revelou alguns padrões no design do chatbot que foram anexados ao propósito do chatbot. Para cada objetivo, os trabalhos geralmente se concentram em um conjunto de impactos e práticas testadas para potencializar os impactos positivos. Esses padrões foram adicionados ao nosso mapa conceitual, ponto de partida para a criação do guia. O mapa estabelece que essas relações podem ser reforçadas por meio de algumas práticas conversacionais, agrupadas em três objetivos: naturalidade, emocionalidade e transparência. Além disso, foram identificadas práticas

que devem ser evitadas.

O guia foi desenvolvido como uma página web que expõe, explica e exemplifica cada prática conversacional com linguagem e apresentação acessíveis. Todas as páginas têm uma linguagem direta para ser uma referência prática para profissionais de nenhum conhecimento avançado sobre desenvolvimento de chatbots. As práticas de naturalidade são: auto-apresentação, chamar o usuário pelo nome, conversa "fiada", ecoar respostas e linguagem casual. As práticas de emocionalidade são: respostas exclamatórias, mídia gráfica, mensagens empáticas e humor. As práticas de transparência são: apresentar capacidades, reconhecer limitações, fazer sugestões e pedir esclarecimentos. Por fim, as práticas que devem ser evitadas são: mensagens repetitivas, esconder a identidade real do chatbot, fontes de "máquinas" e forçar erros.

Na validação com *survey*, foram coletadas 66 respostas, quatro da versão em inglês e o restante da versão em português do survey. A amostra é bastante diversificada quanto à escolaridade e ocupação principal dos respondentes no momento da resposta, sendo que aproximadamente 20% dos participantes tem conhecimento no mínimo intermediário.

A partir do cálculo do Índice de Força Relativa para medir o grau de concordância para cada questão, concluimos que os participantes concordam fortemente que o guia: induziria maior satisfação do usuário; induziria maior engajamento do usuário; é fácil de usar; é claro e compreensível; é flexível para ser usado com chatbots de diferentes domínios; e que eles usariam o guia. Em relação as demais questões, os participantes concordaram moderadamente que o guia agilizaria o design; concordaram substancialmente que facilitaria o design; e discordaram moderadamente que o guia exige muito conhecimento sobre chatbots para ser compreensível.

Realizamos o Teste Exato de Fisher para verificar se há diferenças significativas entre as respostas de participantes com diferentes experiências no desenvolvimento de chatbots. Para todas as questões TAM do questionário, o valor de p está acima de 0,05, indicando que a experiência anterior dos participantes com chatbots não teve influência significativa em suas respostas.

Os pontos fortes do guia apontados nas perguntas abertas foram os exemplos para cada prática, a objetividade e clareza do guia. Por outro lado, a simplicidade foi vista como o principal ponto fraco no ponto de vista da maioria dos entrevistados. Em consonância com o que foi dito sobre os pontos fracos do guia, as sugestões de melhoria estão relacionadas principalmente à necessidade de exemplos mais aprofundados, aplicações do guia e referências de implementação para dar mais credibilidade ao guia.

Na validação com o estudo de caso, convidamos 10 profissionais que atuam com desenvolvimento de software para participar em um experimento que consistia em eles desenvolverem uma amostra de conversa para um chatbot de um aplicativo de meditação.

Primeiramente, eles desenvolveram a amostra de conversa sem terem conhecimento do guia. Após essa etapa, o guia lhes foi apresentado e eles tiveram a oportunidade de alterar a conversa caso identificassem que ela poderia ser melhorada. Por último, foram feitas entrevistas com os participantes para coletar suas opniões sobre a experiência.

Nenhum dos participantes tinha experiência anterior com o desenvolvimento de chatbots, o que indubitavelmente afetou a forma como desenvolviam suas conversas. Apesar de não descartarmos respostas de participantes com experiência, os sem experiência são os mais importantes, pois provavelmente buscariam um recurso para desenvolver um chatbot, como o guia proposto, do que desenvolvedores de chatbot experientes. Por outro lado, todos já interagiram com chatbots, e a maioria já interagiu várias vezes.

Foram feitas análises objetivas, narrativas e temáticas em cima dos transcritos da entrevista e das conversas entregues como resultado do experimento. Todos os participantes, exceto dois, optaram por alterar a conversa 1 após a leitura do guia e apresentaram a conversa 2 na segunda etapa. Desses, apenas um fez uma pequena alteração (uma linha), enquanto os demais fizeram mudanças significativas em suas conversas. Os dois participantes que optaram por não alterar suas conversas relataram que o motivo era que achavam que suas conversas já estavam de acordo com o guia.

Neste estudo de caso, não apenas avaliamos nossos participantes como potenciais desenvolvedores de chatbot, mas também buscamos entender suas experiências como usuários de chatbot. Embora suas experiências anteriores tenham claramente afetado a forma como eles projetaram as amostras de conversa, não foi possível encontrar uma correlação entre idade, sexo e nível de experiência dos participantes com suas escolhas de design. No entanto, a posição que os participantes ocupam apareceu como variável moderadora. Foi possível observar que os participantes que ocupam ou ocuparam cargos relacionados à experiência do usuário tiveram uma preocupação maior em tornar as conversas mais naturais e humanas desde a primeira conversa.

Muitos participantes estavam muito preocupados em como sua conversa seria implementada tecnicamente, o que afetou suas decisões de design. Embora os participantes não tenham experiência como desenvolvedores de chatbots, eles entendem que uma interação de texto livre requer algoritmos mais complexos do que uma abordagem baseada em menus. As práticas de uso também dependiam da consciência do designer sobre o domínio, contexto e público do chatbot. De acordo com esses fatores externos, o guia foi concebido como um cardápio de práticas que devem ser utilizadas com sabedoria. Conforme previsto, os participantes confiaram fortemente em suas experiências e conhecimentos anteriores para selecionar as melhores opções para este chatbot de meditação.

Por fim, todos os participantes concordaram que usariam o guia proposto para desenvolver um chatbot no futuro. Ainda assim, mais da metade apresentou sugestões de

melhorias, embora a maioria fossem pequenas alterações relacionadas ao layout do site. Das quatro sugestões de impacto significativo, três foram relacionadas a tornar o texto introdutório mais acolhedor e uma foi relacionada à apresentação do mapa conceitual.

Conclusões

Os resultados alcançados são promissores e mostram que o guia é útil para profissionais com diferentes níveis de experiência e é genérico e flexível o suficiente para uso em vários domínios. No entanto, é importante notar que o guia é limitado a chatbots baseados em texto, e a análise que forneceu a criação do guia é baseada nas tendências atuais na interação humano-chatbot, que é um campo em rápida evolução. Além disso, nossa validação abordou apenas a facilidade de uso e compreensão do guia, mas não a efetividade das práticas na interação com usuários reais, embora seja esperado que essa efetividade seja herdada dos estudos originários das práticas extraídas da RSL.

Palavras-chave: Chatbot, design conversacional, interação humano-AI, fatores humanos

Abstract

Context: Chatbots are intelligent agents that mimic human behavior to carry on meaningful conversations. The conversational nature of chatbots poses challenges to designers since their development is different from other software and requires investigating new practices in the context of human-AI interaction and their impact on user experience. Since human dialogue involves several variables beyond verbalizing words, it is vital to design well-thought dialogues for chatbots to provide a humanized and optimal interaction. **Objective:** The main objective of this work is to unveil textual, visual, or interactive design practices from text-based chatbot interactions and how they can potentiate or weaken some perceptions and feelings of users, such as satisfaction, engagement, and trust, for the creation of the Guidelines for Chatbot Conversational Design (GCCD) guide. **Method:** We used multiple research methods to generate and validate the guide. First, we conducted a Systematic Literature Review (SRL) to identify conversational design practices and their impacts. These practices were inserted into the GCCD guide through qualitative analysis and coding of SLR results. Then, the guide was validated quantitatively through a survey and qualitatively through a case study. The survey aimed to assess the guide's clarity and usefulness based on the reading of the guide by the participants and their responses to a questionnaire adapted from the Technology Acceptance Model. The case study aimed to assess the guide's usefulness based on its practical application by participants in a situation that simulates a real scenario and follow-up interviews. **Results:** The survey showed that software developers with different levels of experience strongly agreed that the guide could induce greater user satisfaction and engagement. Furthermore, they also strongly agreed that the guide is clear, understandable, flexible, and easy to use. Although participants suggested some improvements, they reported that the guide's main strengths are objectivity and clarity. The case study confirmed the survey findings, as participants reported positive feelings toward the guide and an intention to use it. Their extensive perceptions given through the conducted interviews unveiled that their previous experiences with chatbots and in specific software development positions influenced their design and adoption of practices. **Conclusion:** The guide proved to be useful for developers with different levels of knowledge, with the potential to become a strong ally

for developers in the conversational design process.

Keywords: Chatbot, conversational design, human-AI interaction, human factors

Contents

List of Figures

List of Tables

Abreviations and Acronyms List

AI Artificial Intelligence.

GCCD Guidelines for Chatbot Conversational Design.

SLR Sytematic Literature Review.

TAM Technology Acceptance Model.

Chapter 1

Introduction

This chapter presents the context of this work, the motivations for conducting such research, an overview of the proposed solution, and the methodology. Lastly, it describes how this document is organized.

1.1 Contextualization

Artificial Intelligence (AI)-based systems are crossing the academic barrier to be increasingly used due to two significant factors: the increasing availability of big data and hardware accelerators [6]. Accordingly, AI services' automation capability is used primarily to streamline user service provision [1] and is increasingly being better accepted by users [2]. One type of AI-based system that has gained ground in several sectors is a *chatbot*.

Chatbots are intelligent agents powered by machine learning algorithms to mimic human behavior [3], and users tend to recur to them because of the ease, speed, and convenience of chatting with a chatbot [4]. Although the concept seems futuristic, machines simulating human dialog have been around for a long time, resulting in different conversational systems. The first widely known chatbot in history was a conversational robot that simulated a virtual psychotherapist, Eliza, created in 1966 by Joseph Weizenbaum [7].

Since technology geared toward conversational agents is getting more trustworthy and efficient, chatbots are becoming more and more part of everyday life for ordinary people. There are records of its use for information retrieval, smart home control, services, navigation, entertainment, work, among others [8]. Governments have also explored the use of chatbots, especially to inform citizens and provide essential services [9, 10, 11]. Likewise, during the COVID-19 pandemic, chatbots were one of the front lines in educating citizens about the disease and the necessary measures as part of private, governmental, and non-governmental initiatives [12].

However, the service sector is the one that craves the most the use of chatbots. Companies recur to this technology because it improves customer service experience, reduces cost and resource requirements, and drives digitalization [13]. As a result, the global conversational-AI market is expected to grow at a compound annual growth rate of 21.8% per year and worth USD 18.4 billion by 2026, wherein the service segment is expected to account for the largest market size [14].

Considering the growing presence of chatbots at the forefront of various sectors, it is necessary not only to invest in natural understanding algorithms but also to dedicate time and effort to provide pleasant chatbot interactions for users. Therefore, chatbot teams see the conversations as an object of design to provide a better user experience [15] since bad conversational decisions from designers can negatively impact users' perception of the chatbot [16]. Hence, conversational design is an essential part of the chatbot development process. According to Google [17], conversational design is "a design language based on human conversation" and "a synthesis of several design disciplines, including voice user interface design, interaction design, visual design, motion design, audio design, and UX writing".

A well-thought conversational design has potential benefits that vary according to the chatbot domain and purpose, but comprehensively, it promotes brand likeability [18] and trust [19] in commercial contexts, ensures compliance with public administration principles in governmental contexts [20] and strengthens the achievement of chatbot objectives in all contexts. The latter is supported by the fact that all chatbots share the common purpose of helping users through mutual collaboration — the interaction itself. This user-chatbot collaboration is essential to get users what they expect to get since chatbots cannot predict what users want. Therefore, a well-thought conversational design reassures and sensitizes the user to collaborate with the dialogue, leading them to their goal and consequent satisfaction.

1.2 Research Problem

Even though chatbots' capacity to connect to their users has evolved, it is still a challenge to mimic human behavior, and user interaction is one of the biggest challenges developers face in chatbot development [21]. Human dialogues involve other variables besides verbalizing words. Andrew R. Freed [22, p. 76] justifies in *Conversational AI: Chatbots that work* why conversational design is more than just writing thoughtless chatbots responses:

> "In human-to-human conversations, we know that meaning is not just what you say, but also how you say it. Some people even formulize it. Albert Mehrabian's 7-38-55 Rule of Personal Communication is that spoken words are 7% of verbal

> communication, voice and tone are 38%, and body language is 55% [seen in [23]]. [...] Words matter, but so does tone. A difficult message can be softened with empathy. Just as you shouldn't blurt out the first thing that comes to mind, you shouldn't be thoughtless about what goes into your assistant's dialogue" [22, p. 76].

Related works have already demonstrated that some conversational design practices can positively influence user perception [24, 25, 26]. However, it is not always feasible for chatbot developers to conduct user studies to define the correct conversational design practices. Therefore, it is necessary to investigate chatbot conversational design practices that are technology-independent, user-centric, and focused on achieving business goals through conversations. Furthermore, since user experience is defined as a person's perceptions and responses that result from the use of a system [27], by establishing global design practices, it is possible to provide a friendlier user experience that positively impacts users regardless of the technology behind the chatbot.

1.3 Aims and Objectives

In this work, we conduct an in-depth study of chatbot conversational design practices and how they impact users. Our main objective is to unveil textual, visual, or interactive design practices from text-based chatbot interactions and how they can potentiate or weaken some perceptions and feelings of users, such as satisfaction, engagement, and trust. The specific objectives are as follows:

- Propose user-centric guidelines for text-based chatbot conversational design based on a systematic literature review;
- Validate the proposed guidelines' usefulness and ease of use from the perspective of current or potential chatbot developers.

1.4 Expected Results

The main contribution of our work is a validated set of guidelines that can be used to enhance text-based human-chatbot interaction. These guidelines will help chatbot developers design chatbots that hold more natural and pleasant conversations with users, regardless of their experience level or chatbot framework.

By improving human-chatbot interaction, our guide has the potential to arouse feelings in users that can enhance the services offered by the entity that is represented by the chatbot. For example, in the customer service context, an optimal user experience with the chatbot can make users want to come back to consume the services the chatbot

is representing. Likewise, in learning contexts, improving the user experience of the interaction between the chatbot and the learner can enhance their learning ability.

1.5 Methodology

This research is structured into phases, shown in Figure 1.1. The first phase is the conduction of a Sytematic Literature Review (SLR), which is the source of knowledge for proposing the Guidelines for Chatbot Conversational Design (GCCD) guide in the next phase. Then, the following phases are iterations for validating and improving the guide.

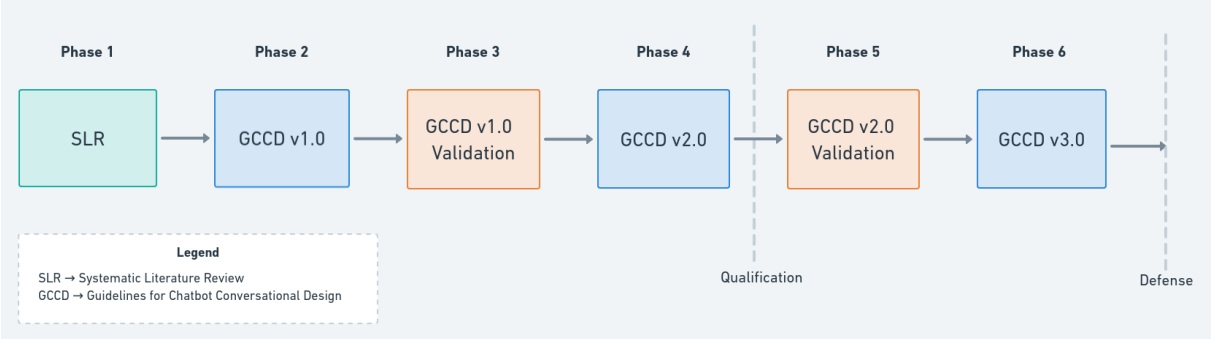


Figure 1.1: Steps to carry out this research.

Table 1.1 depicts the methods used in every phase of the methodology, besides the purpose of using such a method. The SLR is complemented by an analysis using Grounded Theory, which will help in the data extraction and transforming this data into guidelines that will compose the GCCD guide, whereas the other methods are part of the validation. We validate the guide both quantitatively — through a survey — and qualitatively — through a case study.

The proposed guide is aimed at any individual who is a chatbot designer or could potentially be. We consider as a "potential chatbot designers" individuals with knowledge in software design and modeling that could act directly with conversational design, more specifically, requirements analysts, UI/UX analysts, and programmers. In Phase 3 (the survey), we survey a less restricted audience regarding working experience and positions, being only necessary to be studying, working, or having worked with software development. On the other hand, in Phase 5 (the case study), we invite only participants with some professional experience in the positions mentioned above or strictly related.

Table 1.1: Research design method composition

Method	Phase	Purpose
SLR [28]	1	Unveil text-based conversational design practices and their impact on users.
Grounded Theory [29]	1, 2	Translate and transform SLR results into the GCCD guide.
Survey [30]	3, 4	Quantitatively evaluate the usefulness and ease of use of the first version of the guide regarding its reading by potential chatbot conversation designers in the context of conversation design for chatbots.
Case Study [31]	5, 6	Qualitatively evaluate the usefulness and ease of use of the second version of the guide regarding its usage by potential chatbot conversation designers in the context of conversation design for chatbots.

1.6 Publications

The references below point to a conference paper and a journal article that published some contents of this dissertation. The first was a broader investigation of chatbot development according to practitioners' perceptions, which helped explore the field and narrow the dissertation's theme and scope. The second one reports the SLR, the presentation of the GCCD guide, and its first validation, which was conducted as a survey.

Silva, G. R. S., & Canedo, E. D (2022). Requirements Engineering Challenges and Techniques in Building Chatbots. In Proceedings of the *14th International Conference on Agents and Artificial Intelligence - Volume 1: ICAART* (pp. 180-187), ISBN 978-989-758-547-0. https://doi.org/10.5220/0010801800003116

Silva, G. R. S., & Canedo, E. D (2022). Towards User-Centric Guidelines for Chatbot Conversational Design. *International Journal of Human–Computer Interaction*, 1-23. https://doi.org/10.1080/10447318.2022.2118244

1.7 Data Availability

The data that support the findings of this study are openly available on Zenodo at https://doi.org/10.5281/zenodo.7538681 [32].

1.8 Manuscript Organization

This manuscript is organized according to the phases presented in Figure 1.1. Therefore, in the list below, we present the chapters, what phases they document, and a summarized description.

- Chapter 2 — Background: presents essential concepts for the understanding of this work, such as chatbots and conversational design;
- Chapter 3 — Systematic Literature Review (Phase 1): presents the protocol and results of the SLR;
- Chapter 4 — Guidelines for Chatbot Conversational Design (Phase 2): presents how the SLR results were turned into the first version of the GCCD guide;
- Chapter 5 — Guide Validation — Survey (Phases 3 & 4): presents the survey settings and results GCCD's first version, besides the adjustments made to deliver the second version;
- Chapter 6 — Guide Validation — Case Study (Phases 5 & 6): presents the case study settings and results of GCCD's second version, besides the adjustments made to deliver the third version;
- Chapter 7 — Discussion: discusses the implications, contribution, and threats to validity regarding the results achieved by this work;
- Chapter 8 — Conclusion: summarizes the work carried out and the results achieved;

Chapter 2

Background

This chapter presents an overview of chatbots, introduces the concept of conversational design, and discusses the similarity and differences of this work with others that approach human-chatbot interaction.

2.1 Chatbots

The use of natural language, either by text or voice, has been widely used in various systems to facilitate and humanize user interaction with systems [33]. Nowadays, for example, it is possible to change Global Positioning System (GPS) routes without taking your hands off the wheel, book a flight or hotel room by chatting with a machine, search the internet just by talking to the phone or have a virtual agent notify you about important things just like a human would do. In this way, numerous conversational technologies have emerged, which can be categorized differently.

A chatbot is a type of conversational technology with the following defining features: understanding natural language input and the ability to interact and hold a conversation [34]. Still, they can differ in implementation by receiving only text input, voice input, or both. Moreover, their main functions range from performing tasks, troubleshooting, solving doubts, and providing personal assistance [35]. Additionally, they are strongly present in domains such as education, health, tourism, and general customer service [5].

Chatbot technology first appeared in 1966, as shown in Figure 2.1. That year marked the creation of the first chatbot in history, ELIZA [7]. Although it could mimic human behavior by acting as a psychotherapist, it was far from the current chatbots since it uses pattern matching, which significantly limits the coverage of subjects the chatbot can talk about. The following relevant chatbot creation, PARRY [36], managed to emulate emotions but other than that, it did not bring further advances to what was proposed by ELIZA.

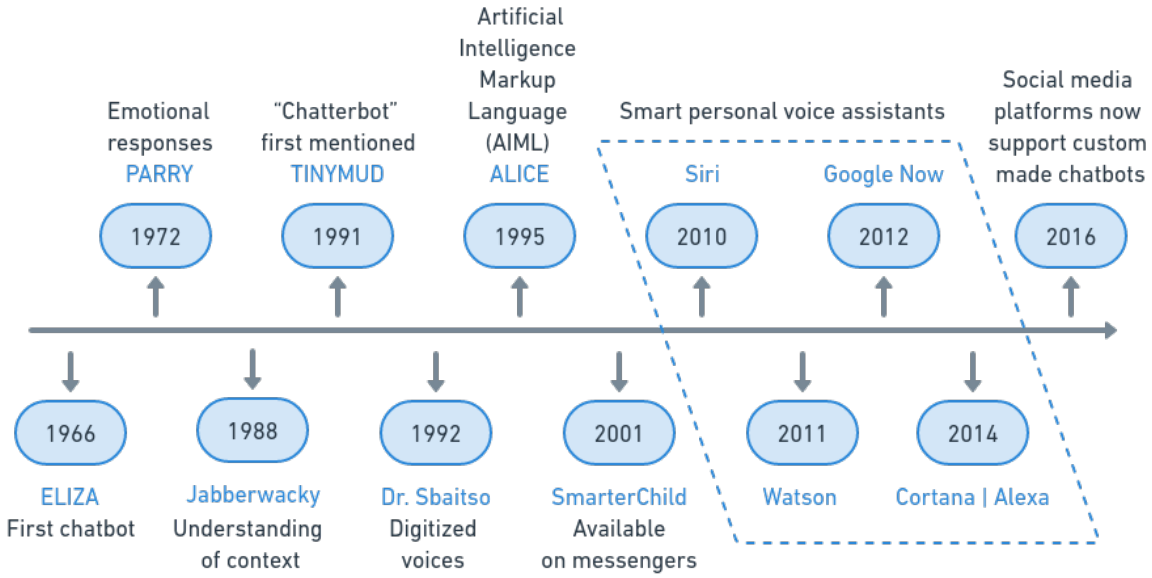


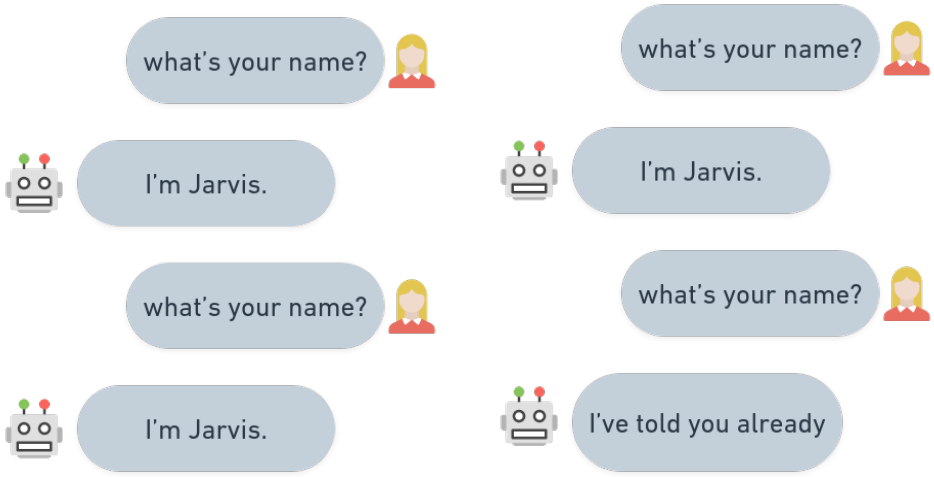
Figure 2.1: Timeline of the evolution of chatbot technologies (based on [5]).

However, in 1988, Artificial Intelligence (AI) was used for the first time in the development of the Jabberwacky [37] chatbot. Because of that, for the first time, a chatbot could deal with multi-turn conversations, using more than the last message to formulate a response. Furthermore, whereas ELIZA and PARRY were Frequently Asked Questions (FAQ)-based, Jabberwacky was context-based, although very limited. Figure 2.2 presents the difference between FAQ-based and context-based interactions.

Until then, chatbots were not yet referred to as *chatbots*. However, in 1991, the TINYMUD [38] virtual conversational agent was referred to as "chatterbot", a name later clipped to "chatbot". In the following year, a chatbot called Dr. Sbaitso [35] introduced digitized voices for chatbots, which is currently closely associated with a class of chatbots called *smart personal voice assistants*.

Even though the use of AI was already introduced in the development of chatbots, in 1995, there was a significant performance advance, as it was in that year that the ALICE [39] chatbot was launched, presenting a new language for developing AI, called Artificial Intelligence Markup Language (AIML). Still, although ALICE implied a much more robust language, it did not have many differences from ELIZA regarding functionalities.

In 2001, there was an advancement regarding the distribution of chatbots to the general public. SmarterChild [5] was available in popular messengers of that time, such as Microsoft Network (MSN) and America Online (AOL). However, it was only from 2010 onwards that we had an incredible leap in how chatbots are present in users' lives, perhaps motivated by the advancement of mobile technologies. From that year until 2014, there



(a) FAQ-based interaction (b) Context-based interaction

Figure 2.2: Difference between FAQ-based and context-based interactions.

was the release of highly popular personal assistants: Siri, Watson, Google Now, Cortana, and Alexa. Many of them are still part of people's lives to this day.

Although the currently most popular assistants started appearing in 2010, the interest in chatbots rapidly increased in 2016, according to Association for Computing Machinery (ACM), Google Trends, [40] and Scopus [35]. The first thing that boosted this increase that year was the opening for the inclusion of third-party chatbots on major social platforms, such as Slack, Telegram, and Facebook Messenger, that happened that same year [41]. This opening helped companies integrate their chatbots into their social networks since having chatbots in the front line of customer service is in the interest of any company as it can replace a team of human attendants and save resources [42].

In the following years, creating tools for chatbot development that encapsulate the complexity of developing chatbots further stimulated interest in the subject, such as Google's DialogFlow, Microsoft Bot Framework, and Rasa [43]. In the past, developers needed to generate complex machine-learning algorithms and build a chatbot from scratch. In contrast, current tools only require that developers feed the knowledge base and pass the training parameters.

2.2 Conversational Design

Chatbot practitioners face additional challenges in chatbot development since chatbots "have specific user experience (UX) requirements related to the way the information is presented, the mean of interaction, whether its text, buttons or speech" [44]. Another challenge of chatbots is designing for open-ended conversations [45] because conversational

flows must be broad enough to cover all conversation possibilities. Conversational design defines how it communicates with users, including its persona, formality of language, visual aids, and conversation shortcuts.

Having a well-defined conversational experience is essential because the customer's quality perceptions are increased if the brand's language is similar to the customer's expectation [46]. Therefore, in the development of chatbots, conversations must be seen as the object of design [47]. Moreover, knowing the target audience will dictate how the chatbot communicates since "younger users may be particularly sensitive to playful and emotionally engaging chatbots, whereas older users may be preoccupied with the efficiency and effectiveness of chatbots" [48].

Missing these specific aspects in the conversational design hurts the chatbot's inclusiveness capacity and causes unimaginable damage to brand images. One example is Microsoft's Tay Bot, released in 2016, which started to post offensive tweets when learning from users' tweets [49]. A research led by The Washington Post revealed that Amazon's and Google's smart speakers work best for white, highly educated, upper-middle-class Americans [50]. Since these failures are not uncommon, big companies have highlighted conversational design as an important part of chatbot development. Google released its own guide for conversation design [51], as well as IBM [52] and Amazon [53]. All of these guides put the user at the center of development. Google's guide has a specific section for gathering requirements that focuses entirely on thinking as the end-users and defining who they are, their needs, their context, and their journeys through conversations.

Since the technological barrier and complexity are becoming less of a problem in chatbot development, there has been an increasing interest in human-chatbot interaction to overcome users' resistance to accepting chatbots compared to human agents [54]. Furthermore, the naturalness of the interaction is critical to overcoming negative preconceptions users may have towards chatbots [55].

Designing chatbot conversations requires research and knowledge of this kind of interaction since users "adapt their language to communicate with intelligent agents" [56]. In light of these challenges, conversational design has recently been in the spotlight. It is an essential process for developing effective chatbots. According to Moore and Arar [57], it comprises the following activities: observe and engage with users, define user personas and goals, shape conversations, define an agent persona, presume user's and agent's messages, prototype and test.

From a more practical side, according to McTear [58], the conversational design ensures the promotion of engagement, retention, pleasant customer experience, and measuring the quality of the interaction. It is a multidisciplinary activity that requires the involvement of developers, designers, writers, and business strategists. Designers define the form of

interaction and the conversational flow, writers polish communication, developers ensure that the chat platforms support the interactive elements suggested by designers, and business strategists guarantee that the agent represents the brand accordingly.

2.2.1 Techniques

Conversational design is a process analogous to software engineering as it begins with requirements gathering and proceeds to prototyping, development, and maintenance. Its distinctive nature requires adequate techniques for designing, implementing, and maintaining natural language interactions. Below, we present some techniques that stand out in conversational design.

Wizard of Oz is a technique for validating an AI product without having a full implementation yet, where participants interact with an illusion of a working product. That is, it aims to simulate the behavior of a machine through a human operator, called a *Wizard*. It is helpful in the requirements gathering and prototyping stages.

The term Wizard of Oz was conceived by Jeff Kelley around the 1980s to describe the method of an experiment he created in his dissertation work at Johns Hopkins University. The term is a direct reference to the movie "Wizard of Oz", in which, at a certain point, the character Dorothy discovers that the Wizard of Oz is, in reality, a man who had controlling mechanisms behind a curtain to generate an image and deceive who wanted to talk to him.

In the case of chatbots, the *Wizard* and the user will talk through a machine, such as a computer, and the *Wizard* will impersonate the chatbot that is under evaluation [59]. This technique allows evaluating the conversational experience as a whole, including the conversation's personality and contents.

Model-Driven Development proposes that the developer does not need to manually interact with all the source code, concentrating on high-level models. It recommends that the application's initial development and future modifications be carried out only in the most abstract model.

A model is an accurate and formal representation of what must be developed but is technology-independent. This model allows for a comprehensive and compact view of the project at all architectural levels and facilitates stakeholder analysis and communication. In the case of chatbots, the model is composed of plain messages somewhat connected to represent the conversational flow, which will be later translated into a chatbot framework language [60].

Data-Driven Development is based on gathering, storing, analyzing, and interpreting data. Regarding its managerial aspect, it is about centralizing the data, coming from reliable sources, as a pillar of business decisions, avoiding guesswork and unfounded assumptions.

Regarding its technological aspect, it is instrumental in systems that use machine learning, as they learn based on a large volume of training data. This is the case of most current chatbots, which are fed by several phrases that will serve as a basis for understanding incoming messages. Thus, it helps build the conversational flow based on datasets from other information sources relevant to the chatbot conversation [61].

Crowd-Driven Development helps in the stage of development. It is based on the crowdsourcing process, a production model that uses collective wisdom and learning for resolution. It is used in general software development, such as for conducting tests or building datasets for machine learning applications. For chatbot development, crowd workers can provide concrete and practical feedback, suggest improvements on specific parts of the conversation [62], or contribute to increasing the dialogue model's training data.

Conversation-Driven Development is an exclusive technique for conversational design. It leverages the analysis of conversations between the chatbot and actual users to improve the conversational flow [42]. However, for that to happen, the chatbot needs to be inserted in an architecture capable of collecting conversations and presenting them for the evaluation of content curators [63].

Reading actual conversations reveals in great detail how the chatbot behaves as well as its users, allowing developers to infer points of improvement for the conversational flow. This tracking of user interaction is an advantage that conversational systems have by nature, as other software cannot build such comprehensive and automatic tracking methods for user interaction.

2.3 Related Work

In recent years, chatbot human interaction research has been at its high due to the advance in Artificial Intelligence and, consequently, the improvement in chatbot interaction capacity. Therefore, since achieving some complexity in conversational design is no longer hidden by the technology available, many studies focus solely on the user experience aspect of human-chatbot interaction. We have considered as related works the studies that review and present recommendations for designing human-chatbot interactions.

Chaves and Gerosa [24] conducted a literature review on disembodied, text-based chatbots to derive a conceptual map of social characteristics for chatbots. They analyzed 56 papers and highlighted how social characteristics can benefit human-chatbot interactions, the challenges and strategies to designing them, and how they may influence one another. This work primarily focuses on subjective social characteristics of human-chatbot interaction, while we primarily focus on the practical, functional, and interface aspects of conversational design.

Sugisaki and Bleiker [64] discussed how text-based conversational user interfaces differ from other forms of human-computer interaction and what challenges and opportunities arise from these differences. As a result, they extracted from high-level usability heuristics a set of 53 technology-agnostic checkpoints specifically for text-based conversational user interfaces. These checkpoints were examined by 15 practitioners and academics regarding content validity. This work uses usability heuristics as a source for deriving their guidelines, whereas our source is a literature review. Moreover, the guidelines are for quality assessment.

Rapp et al. [25] carried out a systematic literature review of 83 papers that focus on how users interact with text-based chatbots in terms of satisfaction, engagement, and trust, whether and why they accept and use this technology, how they are emotionally involved, what kinds of downsides can be observed in human-chatbot conversations, and how the chatbot is perceived in terms of its humanness. This work is also more concentrated on subjective aspects of interactions, such as emotional experience and expression, whereas we are interested in how these aspects are practically implemented.

Amershi et al. [65] proposed generalist applicable design guidelines for human-AI interaction. The guidelines first originated from the literature and industry. Then, these guidelines were refined by a modified heuristic evaluation and tested by 49 design practitioners against 20 popular AI-infused products. Finally, after revisions, the guidelines were inspected by experts, which resulted in the final set of 18 guidelines. Although this work had the same focus as the present work, conversational design practices, the research refers to the full range of AI products, not only text-based chatbots.

Yang and Aurisicchio [26] derived ten guidelines from an interview study to explore how users' competence, autonomy, and relatedness needs could be supported or undermined in experiences with voice assistants. The guidelines recommend informing users about the system's capabilities, designing effective and socially appropriate conversations, and supporting increased system intelligence, customization, and data transparency. The interviews also unveiled determining factors for the success of the interaction, such as the users' knowledge of the conversational agent capabilities, conversation flexibility, and control over user data. Although this work is the closest to ours regarding their primary

objective, they define conversational practices through interviews, whereas we define them through a systematic literature review.

Feine et al. [66] conducted a systematic literature review to identify a set of social cues of conversational agents (CA) to develop a taxonomy that classifies the identified social cues into four major categories (i.e., verbal, visual, auditory, invisible) and ten subcategories. The taxonomy was used systematically to identify various social cues implemented in the text-based CA Poncho, the voice-based CA Alexa, and the embodied CA SARA. This work [66] only covers social cues and considers embodied and voice-based agents, whereas we cover the full range of textual, visual, and linguist design practices of text-based disembodied agents.

Guo et al. [67] reviewed the literature on building trust between users and chatbots in the financial domain to propose a set of design principles to make responses for designing more trustworthy conversational agents in the future. To validate the design principles, the authors conducted a Wizard of Oz study in which each participant was presented with agents that followed the design principles and agents that did not. Results indicated that users considered the agents that followed the principles more reliable and trustworthy. This work's scope is restricted to the financial domain and users' feelings of trust, whereas we aim at generalized guidelines considering a wider range of positive user feelings.

Mafra et al. [68] conducted a literature review to identify quality attributes for chatbots in academic and industry sources. The review ended up with six papers selected. The analysis of these papers resulted in 82 quality requirements for chatbots of three categories: usefulness, ease of use, and presence. This work did not conduct a systematic review and was based only on six studies, whereas we intend to conduct a systematic and broad review. Moreover, the guidelines are for quality assessment.

Komatani et al. [69] proposed design guidelines for developing dialogue systems. Systems developed with the aid of these guidelines took first place in two dialogue system competitions: the situation track of the second Dialogue System Live Competition and a pre-preliminary contest of the Dialogue Robot Competition. The three proposed guidelines are: make the system take the initiative, prevent dialogue flows from relying too much on user utterances, and include in utterances that the system understands what the user said. This work derived guidelines from chatbots that won competitions, whereas we are interested in deriving guidelines considering user feelings toward design practices reported in the literature.

Stanley et al. [70] conducted a review to integrate and find patterns across the literature and entities on accessibility guidelines for chatbots. The authors found seventeen different sources that were analyzed for proposing the accessibility guidelines, which consisted of 157 unique recommendations for chatbot developers, categorized into five

categories: content, user interface, integration, development process and training, and testing. This work is focused on finding accessibility guidelines, whereas we are interested in deriving guidelines considering general user feelings.

2.4 Chapter Summary

This chapter presented the conceptualization and timeline of chatbot technology. Although the first chatbot was delivered in 1966, it was very limited and did not compare to current chatbots. The big turn in chatbot technologies happened in 2010 when the most popular smart personal assistants were launched. Even so, chatbot development gained greater prominence from 2016 onwards, as that was when social networks opened their platforms to customized chatbots, which spurred the mass development of chatbots for companies. Still, going from chatbot requirements to meaningful conversations is very challenging. Therefore, conversations are becoming the object of design when developing chatbots, giving life to the process of conversational design. Therefore, many works have researched this subject, but ours differs because it focuses on pragmatic approaches to building a ready-to-use guide geared toward developers.

Chapter 3

Systematic Literature Review

We have conducted a Sytematic Literature Review (SLR) following the protocol of Kitchenham and Charters [28] to unveil chatbot conversational design practices and their impacts on users. The SLR is the process of identifying, evaluating, and interpreting relevant studies of an area or research question of interest [71], and it is composed of the following phases [28]:

1. *Planning*: identifying the need for a review, establishing objectives, and defining the review protocol, which consists of the following artifacts: research questions, search string, study selection criteria, list of data to be extracted, and quality assessment checklist;
2. *Conducting*: it consists of putting the review protocol into practice by utilizing the artifacts produced in the previous phase to filter studies and extract information from them;
3. *Reporting*: documenting the results of the review, in this case, as a research paper.

3.1 Research Questions

The motivation for conducting the SLR is to disclose state-of-art practices in text-based chatbot conversational design and how they impact users. It is important to notice that our focus is entirely on conversational design and practices directly seen by users. Therefore, we are not interested in the technical aspects of chatbot development. The research questions are shown in Table 3.1.

We consider as text-based chatbots the ones that respond primarily through text, although users can have tools that voice over the responses or parse voice input into text. The decision to consider only text-based chatbots respects the intrinsic differences between this type of interaction and voice-based interactions and the existence of exclusive

Table 3.1: Research Questions

ID	Research Question
RQ.1	What are the textual or visual approaches used in text-based chatbot conversational design?
RQ.2	What are the positive or negative impacts on users of the identified textual or visual approaches in text-based chatbot conversational design?
RQ.3	Are there moderating effects of other variables on the identified impacts of practices?

practices for one or the other, such as the use of images in text-based interactions and speech intonation in voice-based interactions. Moreover, although the two forms of interaction share some conversational practices, it would be necessary to investigate their impact separately because it is not guaranteed that these two methods share the same impact on users, which would significantly extend our scope.

3.2 Search String

The search string was created through the PICOC method (Population, Intervention, Comparison, Outcome, Context) [72]. The population refers to the object of study; the intervention is the means used or caused by the population to achieve some goal; the comparison is what is being compared with the intervention; the outcome is a result of the intervention; and the context is the focus of the study, its restrictions, and limitations. Table 3.2 shows the final definition of the PICOC terms to build the generic search string. Comparison is not applicable because we are not comparing the intervention with anything.

Table 3.2: PICOC terms

PICOC	Keywords	Related Words
Population	chatbot	chatterbot, conversational agent, conversational interface, conversational system, dialogue system,
Intervention	interaction	conversation, expectation, experience, impact, perception, usability, user journey
Comparison	*Not applicable*	*Not applicable*
Outcome	satisfaction	accept, content, effective, enjoy, happiness, preference, quality, trust
Context	text-based	not embodied, not speech, not spoken

We first defined our main keywords for the PICOC: *chatbot*, *interaction*, *satisfaction*, and *text-based*. Then, we did exploratory research with related words defined by us in order to see if the results were relevant and what were the missing words we did not think.

We have used VOSViewer[1] to help visualize the missing words. After some iterations of these steps, we came up with the final adjusted generic string:

> *(chatbot OR chatterbot OR "conversational agent" OR "conversational interface" OR "conversational system" OR "dialogue system")* ***AND*** *(interaction OR conversation OR expectation OR experience OR impact OR perception OR usability OR "user journey")* ***AND*** *(satisfaction OR accept OR content OR effective OR enjoy OR happiness OR preference OR quality OR trust)* ***AND*** *(NOT embodied AND NOT speech AND NOT spoken)*

The digital databases chosen to run the string were ACM Digital Library, IEEE Xplore, and Scopus. They were chosen for being extremely relevant to software engineering research [73], for indexing a great number of conferences and journals, and for being able to run our generic search string directly in its entirety. Table 3.3 shows the specific string for each source.

The initial idea was to run the strings for title, abstract, and keywords. However, the search engines did not share this specific option. In ACM, it was either the title or the abstract without repeating the string. Hence we chose the abstract. In IEEE, "All Metadata" refers to title abstract and keywords, as well as "TITLE-ABS-KEY" in Scopus.

We took advantage of the available filtering and string options for each source to apply some of our exclusion criteria and automate the exclusion of unwanted papers as suggested by Costal et al. [74]. For example, in ACM Library, we applied a manual filter for "Content Type" selecting only "Research Article" and selecting a date range from 2011 until the time of the search. In IEEE Xplore, we applied a manual filter for the date range. Finally, in Scopus, all filters were added to the string, which were: the "Document Type" as "Conference Paper" or "Article"; the "Publication Year" as older than 2010; "Language" as "English" or "Portuguese" or "Spanish"; and the "Source Type" as "Journal" or "Conference Proceedings".

3.3 Selection Criteria

As suggested by Kitchenham and Charters [28], we have defined both inclusion and exclusion criteria. The inclusion criteria refer to the main theme of the papers and which ones should be accepted in the context of chatbots. The inclusion criteria (IC) are shown below:

(IC) Tests one or more text-based chatbot conversational practices with users by:

[1] https://www.vosviewer.com/

Table 3.3: Search Strings per Source

Source	String
ACM Digital Library	[[Abstract: chatbot] OR [Abstract: chatterbot] OR [Abstract: "conversational agent"] OR [Abstract: "conversational interface"] OR [Abstract: "conversational system"] OR [Abstract: "dialogue system"]] AND [[Abstract: interaction] OR [Abstract: conversation] OR [Abstract: expectation] OR [Abstract: experience] OR [Abstract: impact] OR [Abstract: perception] OR [Abstract: usability] OR [Abstract: "user journey"]] AND [[Abstract: satisfaction] OR [Abstract: accept] OR [Abstract: content] OR [Abstract: effective] OR [Abstract: enjoy] OR [Abstract: happiness] OR [Abstract: preference] OR [Abstract: quality] OR [Abstract: trust]] AND NOT [Abstract: speech] AND NOT [Abstract: embodied] AND NOT [Abstract: spoken] AND [Publication Date: (01/01/2011 TO *)]
IEEE Xplore	(("All Metadata":chatbot OR "All Metadata":chatterbot OR "All Metadata":"conversational agent" OR "All Metadata":"conversational interface" OR "All Metadata":"conversational system" OR "All Metadata":"dialogue system") AND ("All Metadata":interaction OR "All Metadata":conversation OR "All Metadata":expectation OR "All Metadata":experience OR "All Metadata":impact OR "All Metadata":perception OR "All Metadata":usability OR "All Metadata":"user journey") AND ("All Metadata":satisfaction OR "All Metadata":accept OR "All Metadata":content OR "All Metadata":effective OR "All Metadata":enjoy OR "All Metadata":happiness OR "All Metadata":preference OR "All Metadata":quality OR "All Metadata":trust) AND NOT "All Metadata":voice AND NOT "All Metadata":speech AND NOT "All Metadata":embodied AND NOT "All Metadata":spoken)
Scopus	TITLE-ABS-KEY ((chatbot OR chatterbot OR "conversational agent" OR "conversational interface" OR "conversational system" OR "dialogue system") AND (interaction OR conversation OR expectation OR experience OR impact OR perception OR usability OR "user journey") AND (satisfaction OR accept OR content OR effective OR enjoy OR happiness OR preference OR quality OR trust) AND NOT voice AND NOT speech AND NOT embodied AND NOT spoken) AND PUBYEAR > 2010 AND (LIMIT-TO (DOCTYPE , "cp") OR LIMIT-TO (DOCTYPE , "ar")) AND (LIMIT-TO (LANGUAGE , "English") OR LIMIT-TO (LANGUAGE , "Spanish") OR LIMIT-TO (LANGUAGE , "Portuguese")) AND (LIMIT-TO (SRCTYPE , "p") OR LIMIT-TO (SRCTYPE , "j"))

(1) applying practices to real text-based chatbots;

(2) simulating practices through Wizard of Oz experiments or similar;

(3) by showing examples of interactions containing the practices and gathering users' impressions;

(4) by explaining the idea of the practices to users and gathering their impressions about it.

Our inclusion criteria do not require the implementation of a chatbot as long as it uses methods that simulate or present a real-like chatbot environment to the user. For instance, Wizard of Oz is a method that allows designers to assess users' reactions and impressions without needing to fully implement the system, with a human filling in the gaps in functionality [75].

The exclusion criteria aim to exclude papers that fit the inclusion criteria but do not present some methodology, focus, aspect, or approach [74]. The exclusion criteria (EC) are shown below:

(EC1) It does not present the impact of the conversational strategy or the individual practices;

(EC2) The object of analysis does not refer entirely to text-based one-on-one interactions between chatbots and humans (e.g., spoken interaction, embodied agent, machine-machine interaction, social media bots);

(EC3) The focus is not on conversational design and user interaction;

(EC4) It is written in a language other than the one understood by the authors (Portuguese, Spanish, and English);

(EC5) It is not a primary full research paper (e.g., book chapters, magazine articles, dissertation, thesis, literature reviews, work in progress, position paper, duplicated work);

(EC6) Published before 2011.

Regarding EC1, it guarantees that the select works correctly identified and actively tested or observed the impact of these practices with users, being either a human sentiment or a behavior pattern. The identified impact can result from using a single practice or a set of practices to comply with a broader conversational strategy.

It is crucial to set apart embodied agents from text-based agents with avatars, as verified by EC2. Embodied agents communicate intentions or messages via body expressions, while text-based agents only communicate via text, although they can be represented by

an image, either human, robotic, or zoomorphic. Simpler and directly, we accepted agents with an icon or figure as long as it is static. Moreover, we are not interested in chatbots that are only "bots", which we consider as agents that communicate but are incapable of engaging in a whole conversation with a human.

If the work is focused on the technicalities of the conversational practice and not user experience, it will be discarded by EC4. Technicalities can be easily detected by identifying the variables being measured. For example, when considering sentiment analysis and adaptive responses as conversational practice, if the user study is interested in the accuracy or performance of the algorithm, it will be discarded. On the other hand, it will be considered if the variables measured in the user study are user satisfaction or other user-centered variables.

The other ECs are related to quality standards to ensure that the selected papers are credible, have a fully developed research, and were adequately reviewed. Moreover, it is essential to define a short time range for accepted papers since chatbots are emergent and constantly evolving technology, making some older works lose relevance, especially in human-AI interaction. Therefore, we discarded studies before 2011 since it was in the early 2010s that chatbots started to get known by the general public [25] with the release of the mainstream assistants that are relevant up to nowadays, e.g., Siri in 2010, Watson in 2011, Google Now in 2012 [5].

3.4 Quality Assessment

Even though the exclusion criteria already give us relevant works, it is still necessary to run a quality checklist to ensure that the practices are valid and their impacts were properly and scientifically measured to be considered in our review. For that, we have used the following checklist to accept papers:

(QA1) Are conversational practices pragmatic and replicable?

(QA2) Is the methodology clear, adequate, and well-defined?

(QA3) Is the number of participants sufficient for wider inference? (≥ 40)

(QA4) Is there a comparison or control group testing the violation or absence of the conversational practices?

(QA5) Are results clear and relevant?

(QA6) Are the impacts statistically calculated? If not, the qualitative analysis is adequate?

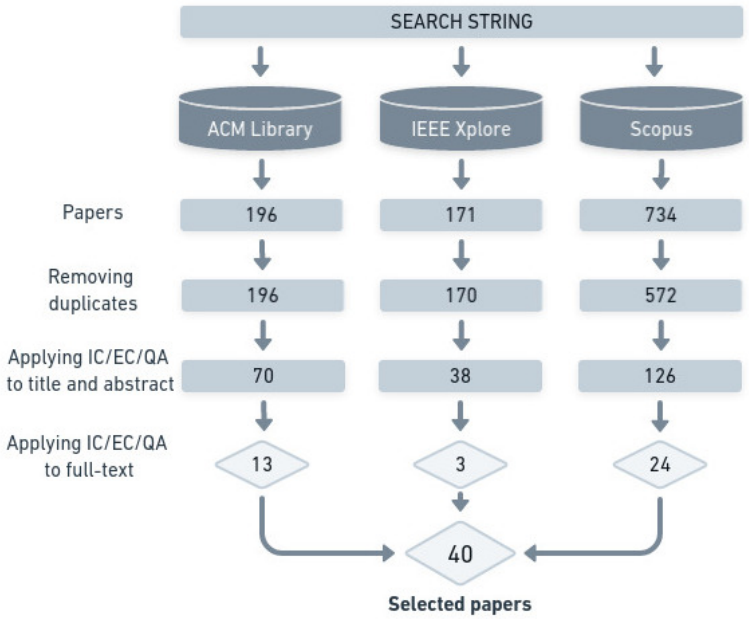
(QA7) Are the impacts of conversational practices properly presented and classified as positive, negative, or neutral?

(QA8) Are limitations and threats to validity presented?

A study was discarded if it did not meet all questions of this checklist. Since this work is concerned not only with practices but their impacts on users, this checklist established a rigorous cutting line to guarantee scientifically well-measured results, and QA2 to QA8 especially checks this. In QA3, we opted for excluding studies with less than 40 participants based on the recommendation of the Nielsen Norman Group [76] that relies on the estimation technique of Sauro and Lewis [77]. Moreover, QA1 aided the removal of works that approached subjective practices, for example, a work that proposes "proactive" chatbots but does not test proper applications of it, such as the chatbot "starting the conversation", which is what we are interested in.

3.5 Conducting

For conducting the review we have used Parsifal[2], which is a free open source web platform for supporting SLR. Parsifal was chosen because its features and workflow were based on the SLR process used in this work that was proposed by Kitchenham and Charters [28]. It streamlines review by providing easy and fast navigation through titles and abstracts during filtering and automatic detection of duplicated papers. Figure 3.1 shows the remaining papers after each step of the review.



IC=inclusion criteria EC=exclusion criteria QA=quality assessment

Figure 3.1: Remaining papers after each step of the SLR.

[2]https://parsif.al

We collected studies through the search string until February 2022, resulting in a total of 1101 papers (196 from ACM, 171 from IEEE, and 734 from Scopus) as seen in Figure 3.1. We had to remove 163 duplicated studies, leaving 938 papers (196 from ACM, 170 from IEEE, and 572 from Scopus) for applying the inclusion and exclusion criteria by reading the title and abstract. Although the papers analyzed by title and abstract were about chatbots, 704 were removed due to criteria violation. Lastly, we read the full text for deeper analysis and application of quality assessment of 234 papers (70 from ACM, 38 from IEEE, and 126 from Scopus), which resulted in the removal of 194 works and the final number of 40 selected papers (13 from ACM, 3 from IEEE, and 24 from Scopus). The detailed dataset is available on Zenodo [32].

3.6 Data Extraction

In systematic reviews, data extraction is vital for building quantitative and straightforward views of the studies. However, our review has the objective of serving as a basis for constructing the guide, therefore being quite restrictive as seen in IC/EC/QA, and it does not tend to have a broad view of the area. Accordingly, the categories of information to be extracted, are highly tied to the research questions. Apart from this reasoning, Table 3.4 presents each paper's year for tracking the evolution of works and the number of users that participated in the experiment. Naturally, the higher the sample, the higher the relevance of impacts.

Table 3.5 presents the data extracted from each selected paper of Table 3.4. Regarding RQ.1, we have extracted data for the column "Conversational practice(s)" as excerpts from selected papers. Therefore, several terms are used for the same practice (e.g., dynamically delayed responses and adaptive response speed) to preserve the integrity of the extracted information. Additionally, many papers used more than one practice to achieve a goal. Thus we extracted these goals to fill the column "Strategy". For this column, we did open coding to assign a strategy to each paper, which is the process of iteratively working on a set of concepts that will later be grouped and classified [29].

The column "Impact(s)" was extracted to support the discussion of RQ.2. Although this column was also filled with excerpts, it was adapted for better understanding. If the identified impact referred to a user's feeling or behavior, we filled it with the excerpt itself (e.g., enjoyment, satisfaction, and self-disclosure). On the other hand, if the impact the paper evaluates refers to how the user perceives a chatbot attribute, we added the expression "perception of" before it. The impacts refer to the combined use of the listed practices, except in the cases explained in the footer of Table 3.5.

Lastly, columns "Moderator(s)" and "Context" support the discussion of RQ.3. The moderators were extracted precisely as initially written in the paper and refer to variables that reduce, empower or change the impacts of practices. These variables were either measured statistically with a significant result or observed by authors in their experiments and presented with qualitative analysis. The chatbot's context is also an indirect moderator since, after the data extraction, it is possible to check for different impacts of the same practices in different domains. The context was also extracted through open coding.

3.7 SLR Results

This section describes the results of the SLR's data extraction and answers the research questions by interpreting these results. It also discusses the implications of implementing the conversational practices considering the aggregated result of the selected studies since some address the same practices.

RQ.1. What are the textual or visual approaches used in chatbot conversational design?

Many practices are used in the literature to cause good impressions on users, although some practices ended up having a negative or neutral impact in some works. These practices are mainly used to humanize the chatbot and use different visual, linguist, or interactive design practices.

Regarding visual design practices, avatars stand-out as a straightforward way of making the chatbot look literally human [PS10, PS14, PS19], however, defining a chatbot avatar is not as easy as it seems because gender and looks can cause impressions in users even before they interact with the chatbot [PS19]. It is vital to remember that we only considered static avatars, as explained in Section 3.3.

Other visual design practices are emojis or emoticons [PS9, PS24, PS3], which interplay with linguistic design practices since they accompany or substitute text messages to convey feelings and emotions. As well as emojis, GIFs, and memes [PS24, PS33] can be used to enhance the expressiveness of emotions.

On the other hand, a wide variety of linguist design practices are used to enhance chatbot conversations, ranging from message content to message formatting, which can convey personality traits from the chatbot. For example, small talk is a practice of talking about casual and out-of-domain subjects [PS13, PS24], such as greetings, jokes, and the chatbot's fictional background. For that, self-disclosing the chatbot's non-human identity

Table 3.4: Identification of selected studies with the chatbot's context and number of participants for the user study.

Paper	Year	Context	Users
[PS1]	2017	Finance	199
[PS2]	2018	Shopping	175
[PS3]	2018	Health	58
[PS4]	2018	Customer Service	84
[PS5]	2019	Shopping, Banking and Travel	203
[PS6]	2019	Customer Service	112
[PS7]	2019	Interview	1280
[PS8]	2020	Open-domain	91
[PS9]	2020	Recommendation	96
[PS10]	2020	Delivery	193
[PS11]	2020	Customer Service	77
[PS12]	2020	Customer Service	159
[PS13]	2020	Mental Health	47
[PS14]	2020	Mental Health	212
[PS15]	2020	Recommendation	54
[PS16]	2020	Financial	410
[PS17]	2020	Booking	189
[PS18]	2020	Donations	790
[PS19]	2020	Shopping	240
[PS20]	2020	Interview	206
[PS21]	2021	Open-domain	263
[PS22]	2021	Customer Service	228
[PS23]	2021	Learning	58
[PS24]	2021	Delivery	171
[PS25]	2021	Customer Service	80
[PS26]	2021	Shopping	54
[PS27]	2021	Customer Service	257
[PS28]	2021	Customer Service	201
[PS29]	2021	Shopping	400
[PS30]	2021	Rental	150
[PS31]	2021	Shopping	426
[PS32]	2021	Recommendation	75
[PS33]	2021	Customer Service	155
[PS34]	2021	Recommendation	310
[PS35]	2021	Mental Health	210
[PS36]	2021	Open-domain	536
[PS37]	2022	Tourism	178
[PS38]	2022	Recommendation	289
[PS39]	2022	Open-domain	139
[PS40]	2022	Surveys	59

Table 3.5: Conversational practices extracted from each selected study.

Paper	Strategy	Conversational practice(s)	Impact(s) on users	Moderator(s)
[PS1]	Typeface	Machine-like typeface (OCR-A)	[-]perception of humanness	Familiarity with AI
[PS2]	Anthropomorphic cues	Human name, informal language	[N]perception of social presence, [+]mindless perception of anthropomorphism, [+]mindful perception of anthropomorphism	None
[PS3]	Emoji based dialogue	Emoji	[+]enjoyment, [+]attitude, [+]confidence	None
[PS4]	Social cue	Dynamically delayed responses	[+]perception of humanness, [+]perception of social presence, [+]satisfaction	None
[PS5]	Repair [breakdown]	Acknowledging misunderstanding and suggesting solutions	[+]preference	[User's] social orientation, experience with chatbots and technology.
[PS6]	Social cue	Sentiment-adaptive responses	[+]perception of empathy, [+]perception of humanness, [+]perception of social presence, [+]satisfaction	[User's] gender
[PS7]	Reserved and assertive personality	Reserved, calm, assertive, rational, careful, like a counselor	[+]willingness to confide, [+]willingness to listen	[User's] personality and context
[PS8]	Linguistic style	Code-mix	[+]perception of conversational ability, [+]perception of humanness	[User's] language proficiency
[PS9]	Nonverbal cues	Emojis	[+]social attractiveness, [+]perception of competence, [+]perception of credibility	None
[PS10]	Visual cues	Avatar	[N]perception of social presence	None
[PS11]	Preset answer options	Buttons	[-]perception of humanness, [-]perception of social presence, [N]satisfaction	None
[PS12]	Self-presentation	Introducing itself as a chatbot	[-]perception of social presence, [-]perception of humanness, [N]satisfaction	None
[PS13]	Self-disclosure	Small talk	[+]self-disclosure	Passage of time
[PS14]	Racial mirroring	Profile pictures and names that might have implied their racial identity	[+]interpersonal closeness, [-]disclosure comfort, [a][+]satisfaction	None
[PS15]	Interaction modes	Buttons	[N]satisfied, [+]understood	None
[PS16]	Socio-emotional features	Human name, respond empathetically, give encouraging statements, active listening skills, using the user's preferred name, turn-take and small talk	[N]trust, [N]privacy concerns, [-]data disclosure	Perception of social presence
[PS17]	Repair [miscommunication]	Clarification request	[+]perception of anthropomorphism, [+]adoption intent	None
[PS18]	Persuasive	Inquiry	[-]donation probability	Perceived identity of the chatbot
[PS19]	Gender cues	Female avatar	[+]forgive in the error condition, [+]satisfaction, [+]social disclosure	None
[PS20]	Active Listening	Paraphrasing, verbalizing emotions, summarizing and encouraging	[+]engagement, [+]interest, [+]chat experience	None
[PS21]	Extraverted	Many topics in a short amount of time, informal language, compliments and positive emotion words	[N]perception of humanness, [+]perception of social presence, [N]communication satisfaction	User personality
[PS22]	Typing errors	Dynamic error, temporal error and spatial error	[-]perception of humanness, [-]perception of social presence	None

[+]positive impact [-]negative impact [N]neutral impact

[a] higher "satisfaction" in comparison with "not mirroring" but not with baseline (robot avatar);

Table 3.5: (continued) Conversational practices extracted from each selected study.

ID	Strategy	Conversational practice(s)	Impact(s) on users	Moderator(s)
[PS23]	Humor	Jokes, conundrum riddles and funny stories	[+]motivation, [+]effort	Self-defeating humour
[PS24]	Social-oriented	Small talk, exclamatory feedback, GIFs and emoticons	[+]perception of social presence, [N]trust, [+]enjoyment, [N]intention to use	None
[PS25]	Warmth	Friendly initial message	[+]engagement	Brand affiliation
[PS26]	Mixed-modality interaction	Buttons, sliders and checkboxes	[+]enjoyability, [N]perception of supportiveness, [N]perception of efficiency, [N]perception of precision	None
[PS27]	Chatbot disclosure	Introduced himself as “Michael" and revealed himself as a chatbot	[-]trust	Acknowledge expertise or weakness
[PS28]	Chatbot disclosure	Introduced himself as “Leon". [...] At the end of the conversation, it was revealed [...] that the service agent [...] was in fact not a human person, but a chatbot	[-]trust, [-]perception of humanness	Service criticality and failure setting
[PS29]	Conversation initiation	System-initiated non-anthropomorphic assistant	[+]reactance	Anthropomorphic avatar, [user's] gender
[PS30]	Politeness	Polite greeting, polite goodbye, polite thanks and polite info on hours	[+]engagement	[User's] gender, age and personality
[PS31]	Anthropomorphism	Human name, informal language, typing cues, dynamic delay, jokes	[+]intention to buy, [+]offer sensitivity, [b][+]likeability	[Chatbot] disclosure
[PS32]	Linguistic	Lexical and structural alignment of responses	[+]user alignment	None
[PS33]	Social presence	Responses were designed to be informal, expressing emotions, and using numerous emojis and funny memes. Asking users their names to greet and address them by name.	[+]user engagement, [+]satisfaction, [+]brand likeability	None
[PS34]	Justification	Explaining why an item was recommended	[+]trust, [+]perception of transparency	[User's] age and experience with technology
[PS35]	Supportive messages	Attentional deployment, cognitive change, general emotional support, situation modification	[+]valence, [N]arousal	Participants who believed to be interacting with a human being
[PS36]	Social and Emotional Qualities	Politeness, small talk, sense of humor, emojis, short exclamations, express feelings, call me[user] by my name, ask questions.	[+]behavioral intentions	[User's] openness to technologies, empathy propensity and vulnerability
[PS37]	Register Compliance	Linguistic features	[+]perception of appropriateness, [+]perception of credibility	Domain
[PS38]	Rapport-building	Justify its recommendation	[+]trust, [+]satisfaction, [+]perception of usefulness, [+]perception of ease of use	
[PS39]	Authenticity signals	Female avatar	[+]perception of authenticity, [+]engagement, [+]satisfaction, [+]loyalty	[Avatar's] race congruence and professional dress
[PS40]	Humanization	Self-introduction, addressing respondents by their name, adaptive response speed and echoing respondents' answers	[+]perception of anthropomorphism, [+]perception of social presence, [+]satisfaction, [+]self-disclosure	None

[+]positive impact [-]negative impact [N]neutral impact

[b] “dynamic delayed response" has not improved likeability in individual experiments;

plays a vital role since the decision to reveal the chatbot's true identity as a virtual agent [PS27, PS28] will define its background.

Moreover, chatbots can be emphatic by adapting responses to what has been said by the user [PS6] such as demonstrating sadness after the user informed something went wrong or by echoing users' responses through reaffirming what the user has just said [PS40]. These practices can also be applied when the chatbot cannot solve a problem or does not understand a message to repair a breakdown [PS5]. Another way of conveying feeling through messages is by leveraging punctuation, such as exclamatory feedback [PS24]. Moreover, jokes and funny stories can also be used to pass on joy and excitement [PS23].

The initial message of the chatbot can be decisive for user retention. Therefore, chatbots can start by presenting themselves as a human or a machine [PS40, PS27, PS7], telling their name [PS14, PS31], welcoming the user with a friendly message [PS25] and presenting what its capabilities are [PS28]. This moment is also adequate for collecting the user name for later use when addressing the user during conversation [PS40, PS33]. Furthermore, when dealing with bilingual users, the chatbot can also code-mix, which consists of inserting foreign words in the middle of the message [PS8].

Regarding language choices, chatbots can be polite [PS30] but with some touches of well-dosed informality [PS36]. Language can also be personalized to match the domain by varying the use of verbs, pronouns, conjunctions, and other linguistic features [PS37]. Moreover, chatbots can mirror their users' language style [PS32]. The language can also convey some information indirectly by using distinctive typefaces, such as handwritten or machine-like [PS1]

Chatbots can help users understand them more by being honest and open with what happens behind the conversation. Acknowledging misunderstandings and suggesting solutions are ways of leading users out of a conversation breakdown [PS5]. Moreover, justifications and explanations are fundamental for users to understand why they receive some information, instruction, or recommendation from the chatbot [PS38, PS34].

Lastly, interaction design practices can humanize the agent, streamline conversation and avoid breakdowns. For example, typing cues and dynamic delayed responses increase the impression that there is a human being typing on the other side by masking the instantaneous response of the chatbot [PS31, PS4]. Beyond that, buttons and carousels are helpful for guiding users to quickly send the message the chatbot expects to function well [PS11].

RQ.2. What are the positive or negative impacts on users of the identified textual or visual approaches in chatbot conversational design?

The impacts investigated by the selected papers are tied up with the chatbot domain. Thus, works investigating the impacts of chatbots in the health context are more interested in sentiments that empower users' well-being, such as motivation and enjoyment [PS3, PS14]. For mental health, the user must develop feelings of closeness, friendship, and trust to self-disclose to the chatbot [PS13], which is an essential factor for the success of mental treatment. Conversely, commercial brands are more interested in user retention and satisfaction ratings, although the other feelings previously cited do not impede customer service success.

It is no secret that users prefer a human agent rather than a virtual agent [54]. Because of that, a chatbot that discloses itself as non-human can cause users to instantly lose trust [PS27], which is not surprising since it is natural that humans have higher confidence toward other humans rather than machines. However, this effect is not due to the displayed identity but to the perceived identity [PS18], that is, the identity that users believe is the true one.

Suppose the chatbot pretends it is human, and the user is suspicious about it. In that case, the negative impact may be much worse than disclosing the chatbot identity as non-human because the user will feel deceived [PS31], get angrier, and more frustrated [78]. Moreover, disclosing identity combined with showing what the chatbot is capable of and communicating its weaknesses can produce trust levels corresponding to that of undisclosed conversational partners [PS27]. Therefore, for this specific practice of disclosing identity, we can consider that the decrease in the perception of humanness after disclosure [PS12] is a natural effect that must be endured to avoid worse impacts of users finding out they are being deceived.

Moving forward, design practices used for humanization, social presence, or for conveying emotions have, in general, positive impacts on users. For example, in the experiment conducted by De Cicco et al. [PS24], a social-oriented chatbot increased users' perception of social presence and enjoyment by using small talk, exclamatory feedback, GIFs, and emoticons. Rhim et al. [PS40] used self-introduction, addressing users by name, adaptive response speed, and echoing respondents' answers as humanization techniques, which positively affected users' satisfaction and their willingness to spend more time with the humanized chatbot in comparison with the baseline chatbot.

Other works also used a large set of practices to increase the chatbot humanness and achieved positive impacts as well. For example, practices used towards active listening skills increased engagement and interest [PS20], the use of humor through jokes, conundrum riddles, and funny stories stimulated motivation and effort in an educational context

[PS23], anthropomorphic features enhanced users intention to buy and offer sensitivity in a customer service context [PS31] whereas social presence induced brand likeability in the same context [PS33].

Chatbot humanization not only has a positive impact, but it increases the users' perception of humanness or anthropomorphism [PS2, PS40], that is, these practices are successful in making the chatbot appear more human. Consequently, as humans tend to see other humans as bad, good, pleasant, unpleasant, and so on, the humanized chatbot must have a personality that pleases its target audience. Shi et al. [PS18] found that a chatbot designed to be persuasive by inquiring users to donate caused the opposite effect. In the work of Ahmad et al. [PS21], although the extraverted personality of their chatbot made users perceive it more as a human, it was not a determinant for satisfaction.

Zhou et al. [PS7] experimented with two interviewer chatbots, one was designed to be reserved and assertive and the other to be warm and cheerful. Results showed that users were more willing to confide and listen to the reserved and assertive chatbot in a high-stakes situation. Something that was not measured but may have impacted this experiment is that the reserved chatbot had a male avatar, and the other had a female avatar.

The moderate use of unnecessary personality traits in a context is not detrimental, as seen in the work of Diederich et al. [PS6] in which an empathetic chatbot enhanced positive feelings in users. However, missing essential personality traits may displease users, such as designing a therapist chatbot with a cold personality. In conclusion, the chatbot's personality depends on a study of the target audience to match their interaction style as well as the context it will be inserted in line with real situations. For example, we do not expect a job interviewer to comfort us as we do expect a therapist. When these studies are not feasible, the safer approach is to assign a pleasant yet subtle personality to the chatbot without overdoing it.

The impacts found by works researching chatbot avatars are conflicting, ranging from negative, neutral, and positive. For example, Tsai et al. [PS33] found that a human name in conjunction with a human avatar only boosts other humanization design practices, but they are insufficient to impact consumer response. In Pizzi et al. [PS29], the enforcement of anthropomorphism through avatars in automatically activated agents negatively impacted users, whereas the lack of avatar was positive. These works show that the impact of avatars is highly dependent on other variables.

Although humanization potentializes the willingness of users to self-disclose personally [PS40], when it comes to sensitive data, such as talking to a chatbot in a financial context, users have privacy concerns [PS16]. This may be because users trust more in machines to keep their information secure than humans. Therefore, humanized chatbots tend to

decrease their trust to disclose data. Still, the overall benefits of humanizing a chatbot compensate for this downside, although it is necessary to conduct more research on coping with these specific situations.

For customer service and recommendation chatbots, clarity and openness are essential attributes. Mozafari et al. [PS28] conducted a study to verify the impact of disclosing the chatbot's true identity and unveiled that, in general, users' trust is negatively affected when the chatbot reveals it is not human. However, disclosing it after a failure ameliorates users' perceptions of integrity and benevolence. Pecune et al. [PS38] and Wilkinson et al. [PS34] found that users place more trust in chatbots that justify their recommendations, in addition to being more satisfied with the suggestions they receive from these chatbots. Chatbot transparency can also be applied by acknowledging failures and helping the user to get on track by making suggestions, which has also proven to be a user preference for interaction [PS5].

Still, if avatars are used, gender and physical attributes can improve user perception. Toader et al. [PS19] found that female avatars, in contrast to male avatars, create stronger perceptions of warmth, generosity, and kindness besides being more forgiven when committing errors due to users' biased thinking based on social roles commonly assigned to women. Liao and He [PS14] conducted an experiment in which they presented to users different versions of a therapeutic chatbot that had varied racial identities expressed through the avatar's name and physical attributes. Results revealed that when users matched the racial identity of the avatar, it facilitated the interpersonal relationship. However, it made users more concerned about being judged.

The use of informality, such as casual language, GIFs, and jokes, are coupled with sentiments of joy, closeness, friendship, and motivation in health and learning contexts. However, they do not determine user satisfaction in commercial contexts [PS24, PS23]. Emojis were individually put under test in two studies, and both of them had very positive and expressive results regarding the use of these elements [PS3, PS9]. However, since politeness is also a positive design practice [PS30, PS36], the use of informality must be moderated so as not to exceed the limits of good manners and to cancel the professional image of the chatbot.

Typing errors were also investigated as a design practice of informality to make the chatbot more human. However, it did not have a positive impact since users thought it was "a lack of developer competence" [PS22]. Finally, interaction design practices, such as buttons, can streamline conversations. However, they have a neutral effect on satisfaction and decrease the perception of humanness due to the mechanical action of selecting a predefined response in contrast to the natural feel of freely building a response [PS11, PS15]. Therefore, such elements must be used with caution.

RQ.3. Are there moderating effects of other variables on the identified impacts of practices?

We have identified different moderating variables that were statistically measured and significantly changed the impacts of the practices. Chatbots with an anthropomorphic avatar that automatically initiates conversation leads to lower levels of satisfaction compared to the non-avatar agent [PS29]; using humor is generally positive, but the use of self-defeating jokes negatively impacts enjoyment [PS23]; the chatbot acknowledging its limitations smooths the negative impact of revealing the chatbot identity [PS27]; the chatbot revealing it is not human at the start of the conversation reduces users' trust, but right after a failure is positive [PS28]; and users finding out that they were talking to a chatbot when it was pretending to be human decreased positive effects of humanization design practices [PS31].

One practice that particularly suffers from the interplay with other measures is avatars. As we started discussing in RQ.2, human avatars potentiate other humanization techniques but increase the expectation of users. Therefore, users can feel deceived by the human image and be more frustrated with chatbot failures. Moreover, every detail of the human avatar, such as gender and physical attributes, moderate the impact on users. By a joint analysis of selected works, the safest approach seems to be a female human avatar, with racial mirroring in conjunction with self-disclosing the chatbot as non-human for transparency. Furthermore, although the perception of humanness may decrease, a robot avatar can also be used without negatively impacting satisfaction [PS14].

Other moderating variables that are beyond the control of designers must be taken into account when using the practices such as conversation duration [PS13], context of use [PS7, PS37], user's personality [PS7, PS30, PS21], language proficiency [PS8], age [PS34, PS30], gender [PS6, PS30] and experience with technology [PS34, PS5]. Besides these works and apart from the ones that did not find significant moderating effects from age and gender [PS17, PS6, PS11, PS4], others have not tested moderating effects of age and gender. Moreover, the participants' demographics generally exclude older adults and the elderly as well as adolescents.

In the work of Rana et al. [PS30], it was statistically found that women were more sensitive to the chatbot's polite triggers, being more positively impacted than men. In some cases, men even gave a lower rating to the polite chatbot. Considering that it is not always possible to run deep tests with target users, one strategy to mitigate these effects is to balance the use of practices to avoid exaggerations. For example, if a chatbot is extraordinarily informal and uses too many jokes, women may feel uncomfortable, whereas men feel joyful.

Ahmad et al. [PS21] found no correlation between specific user's personality traits

and chatbot preferences because the conjunction of many personality traits makes users unique, and they conclude there is a need for personalization during the conversation. Personalization is limited in conversational design but can be enhanced on the technical side by identifying users' traits with natural language processing, which is not in our scope. However, the previous recommendation of no exaggerating applies, preventing the chatbot from being inappropriate for certain groups.

Chaves et al. [PS37] found that the correct use of linguistic features (e.g., verbs, coordinate conjunctions, pronouns) positively impacts users, but the linguistic features to be used change from domain to domain. Similarly, Zhou et al. [PS7] presented evidence that users' preference for a chatbot personality depends if it is a high-stakes situation.

Designers should also check for users' perceptions of practices. As seen in the work of Ng et al. [PS16] and Shi et al. [PS18], regardless of the practice that has been used, what matters is what the user believes and perceives, which moderates impacts. For example, if the chatbot is designed to pretend it is human, but the user does not believe in this identity, the impact of humanness is annulled. This can be mitigated by running pre-tests to measure users' perceptions of practices.

Through qualitative analysis of participants' responses, Ashktorab et al. [PS5] and Rhim et al. [PS40] have found that the positive effect of textual practices is harmed by the lack of variability in messages causing users to see the chatbot as "an auto-machine". Therefore, designers should conceive different responses for conveying the same message when using any practice.

3.8 Chapter Summary

This chapter detailed the protocol used to conduct the SLR besides the results of this review. The summarized results are shown in Tables 3.4 and 3.5. The search string returned a total of 1101 papers, and after removing duplicates and filtering by title, abstract, and full text, there were 40 selected papers. These papers revealed a significant effort in making the chatbot more human with anthropomorphic features and tailoring these features to the best setting, such as testing some personality traits. Moreover, there were attempts to facilitate communication with interactive design practices and make the conversation more transparent with openness and clarification. The collected conversational practices had an overall positive impact on users, but many have moderating variables that are difficult to avoid because they are inherent to users.

Chapter 4

Guidelines for Chatbot Conversational Design

Studying the papers selected by SLR enabled us to propose guidelines for conversational practices that can help designers in building user-centered chatbots. The first step was to build a conceptual map that synthesizes the selected studies' collective knowledge, which is further explored in Section 4.1. Then, based on the conceptual map, we built a simple and objective guide in web page format that explains conversational practices and in which situations they should be used, which is further explained in Section 4.2.

4.1 Conceptual Map

In the SLR, we kept the extracted data as close as possible to its origin in the paper, as explained in Section 3.6. Answering research questions by looking at the data in Table 3.5 paved the way to a deeper analysis done through open coding, axial coding, and selective coding. Open coding was already applied and explained in Section 3.6, whereas axial coding and selective coding are responsible for grouping codes into categories and finding interrelationships among them, respectively [29].

This analysis revealed some patterns of study focus depending on the chatbot's purpose. For example, papers that tested customer service chatbots were highly interested in satisfaction. In contrast, papers that tested chatbots in the context of health were more interested in feelings of well-being. These patterns are listed in the conceptual map shown in Figure 4.1.

This model starts by separating the purpose of chatbots and linking them to what should be the type of relationship that has to be built with the user. The listed relationships arose from a qualitative and joint analysis of selected papers and are linked to a group of impacts investigated by selected papers. From our analysis, it was possible to

identify the focus of each group of impacts, which were *transparency*, *naturalness*, and *emotionality*.

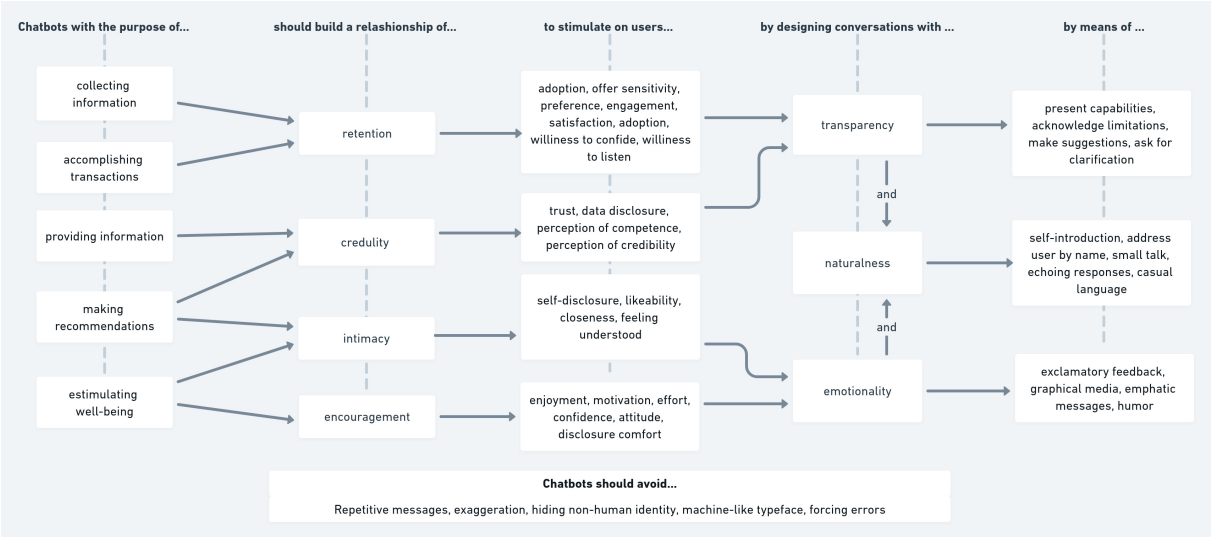


Figure 4.1: Conceptual map of chatbot conversational design according to purpose and user impact.

Chatbots that seek to collect information, such as those that conduct surveys and interviews, must keep users' attention, make them complete the questionnaire, and provide reliable information. Similarly, chatbots that accomplish transactions, such as booking and shopping, must keep their users interested to make them complete the purchase. Therefore, to *retain* users, the best approach is to acknowledge capabilities and limitations and set user expectations right away.

Credulity is a critical factor for users looking for information and recommendation in a chatbot since the reliability of responses impacts users' trust. Moreover, to make recommendations to users, chatbots must know more about them, which can be achieved by creating a relationship of *intimacy*. Lastly, to stimulate well-being, a chatbot must act as a companion and build a relationship of *intimacy* and *encouragement*.

We have identified three constructs for conversational design that can be used to build these relationships. The selected studies revealed that *naturalness* is essential in these use cases. *Transparency* is critical when the chatbot needs to be competent and effective, whereas *emotionality* is more critical when a more profound connection is necessary to accomplish the chatbot's purpose.

Emotionality and *transparency* are not mutually exclusive, but designers should switch focus on achieving the desired relationship. For example, users talking to a therapist chatbot are not concerned about being aware of everything in the chatbot. However, they want to be listened to, understood, and cheered. In this use case, focusing on being transparent more than emphatic is detrimental to the user experience. Therefore,

Emotionality's practices should be enforced, and *Transparency*'s practices, if used, should not be in the spotlight.

Regarding what *Chatbots should avoid*, *machine-like typeface* came from the findings of Candello et al. [PS1], and forcing errors from findings of Buhrke et al. [PS22]. *Repetitive messages* came from deeper discussions of participants' perceptions in the works of Ashktorab et al. [PS5] and Rhim et al. [PS40]. Moreover, *Exaggeration* should be avoided because users' personalities and demographics have been recurring moderators in selected works. Therefore, it is vital to stay in the middle ground to please the most significant number of users, especially when designers cannot afford to do extensive user studies before designing the conversations. Lastly, although disclosing the chatbot's true identity negatively impacts users' trust, as discussed in Section 3.7, *hiding non-human identity* has worse effects than disclosing it when the user finds out it is being deceived.

Although the open coding was responsible for reducing the list of practices seen in Table 3.5 to the one shown in the conceptual map, some of them do not appear in it intentionally because of conflicting results among papers or little positive impacts, such as buttons and avatars. Moreover, some practices such as code-mix, racial mirroring, and linguistic alignment are tied to specific contexts or depend on audience characteristics. Therefore, they were not included since the conceptual map intends to be context-independent.

4.2 Guide Structure

The proposed guide was developed as a web page that brings in a more accessible language the concepts that were defined in the conceptual map. It is called Guidelines for Chatbot Conversational Design (GCCD) and it is entirely avaiable on Zenodo [32] alongside other supplementary material. It is composed of the following pages: *Home*, *Conversational Design*, *Naturalness*, *Emotionality*, *Transparency* and *What to avoid*.

The *Home* page presents the justification of the guide and a summary of the guide's contents. The "Conversational Design" page follows the same structure as the home page, with a text approaching the importance of a well-executed design and the explanation of a shorter version of the conceptual map, intending to explain to users that the practices in the following pages are more relevant in specific contexts.

On the other hand, the pages *Naturalness*, *Emotionality*, *Transparency*, and *What to avoid* have a different structure, as shown in Figure 4.2. They start with a short paragraph approaching how to achieve the characteristic that entitles the page. Then, for each listed practice, there is a short explanation followed by a figure showing a generic example of a conversational practice that should be implemented or avoided.

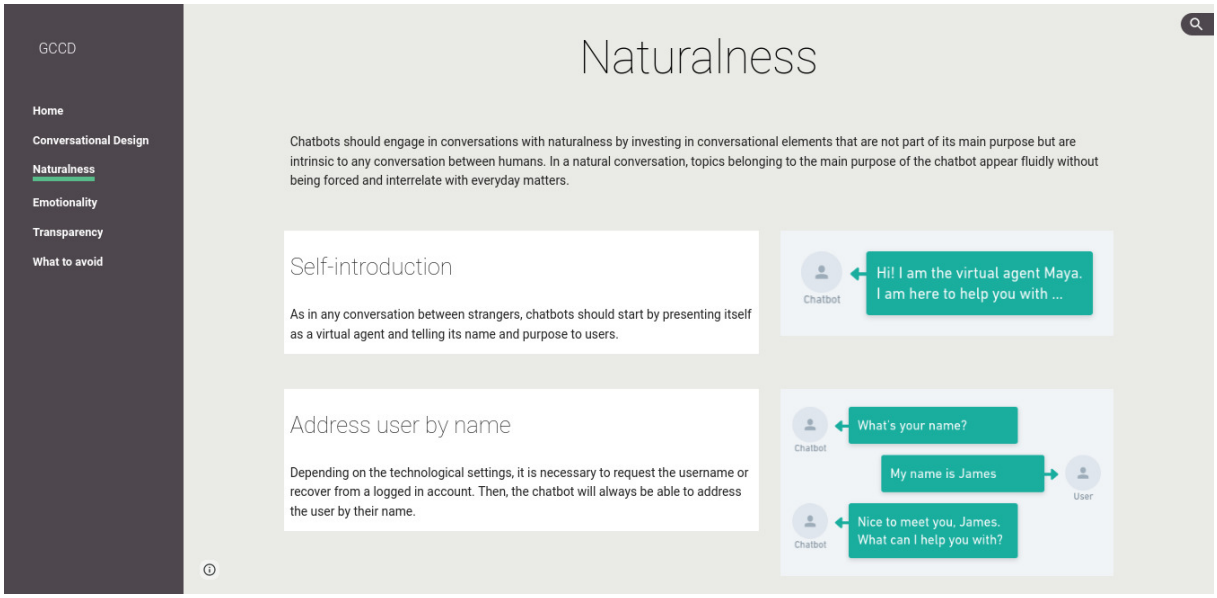


Figure 4.2: User's view of a guide's page that exemplifies practices for a specific focus, in this case, naturalness.

All pages have a straightforward language to be a practical reference for readers with none to advanced knowledge about chatbot development. The guide aims to be an accessible reference for designing effective chatbot conversations even though it is based on a joint analysis of scientific studies with strong theoretical foundations. It is important to note that it only approaches conversational design guidelines, not technological implementations or design processes, as it is out of our scope.

4.3 Proposed Conversational Design Practices

This section presents the practices that are part of the GCCD guide. Since the guide is intended for day-to-day use, its language distances itself from formalisms and complex terminologies, aiming for the guide's clarity and accessibility regardless of the reader's experience level. Moreover, each practice is accompanied by an image that exemplifies how this practice could be used in chatbot conversations.

Albeit the guide presents the conceptual map as a proposal of how the practices should be used, as previously said, the constructs and their corresponding practices are not mutually exclusive. Therefore, it is not intended to be followed blindly since it was possible to identify in the SLR that the chatbot's context and domain can demand different approaches. In this sense, our guide presents itself as a generic menu of validated practices with suggested use rather than a strict policy. Therefore, designers should reflect upon each practice to choose them wisely, considering aspects such as conversation flow length, particular target audiences, and stakeholders' necessities and requirements.

4.3.1 Naturalness

Chatbots should engage in conversations with naturalness by investing in conversational design practices that are not part of their primary purpose but are intrinsic to any conversation between humans. In a natural conversation, topics belonging to the primary purpose of the chatbot appear fluidly without being forced and interrelating with everyday matters. Practices are presented below and examples are shown in Figure 4.3.

Self-introduction As in any conversation between strangers, chatbots should start by presenting themselves as virtual agents and telling their names and purposes to users.

Address user by name Depending on the technological settings, requesting the username or recovering from a logged-in account is necessary. Then, the chatbot will always be able to address the user by name.

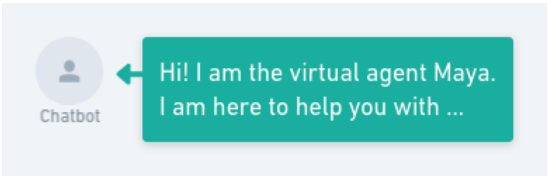
Small talk or chitchat Consists of pieces of dialogues with no particular goal or function, such as greetings, asking how someone has been doing, praise, and thanks. Be careful not to deviate too much from the chatbot's primary objective.

Echoing responses When a user sends a message, avoid simply responding "Ok". The chatbot should respond, including pieces of text from the user message whenever possible, in a way that makes it seem that the chatbot made that answer exclusively for that user message.

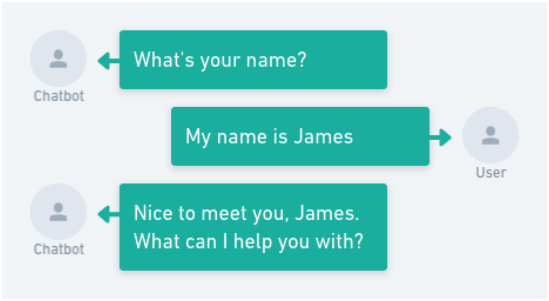
Casual language Personal statements in the chatbot's language help to make it less like a repository of information and more like an agent. Abuse of elements that make messages part of a conversation and not something that would simply be displayed on a web page. Balance the level of informality according to the chatbot's audience and purpose.

4.3.2 Emotionality

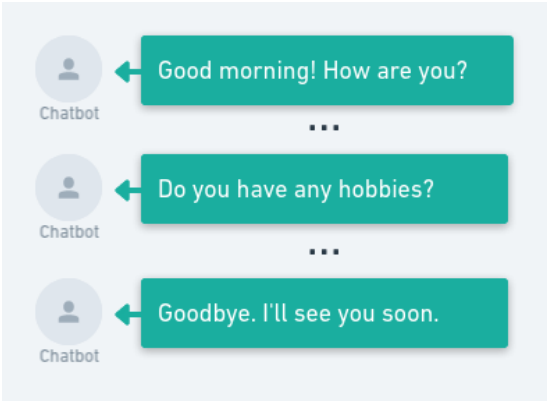
Here, emotionality refers to the chatbot's capacity to express feelings and show understanding of the user's feelings. This capacity is essential to build deeper connections with users and stimulate their self-disclosure. Practices are presented below and examples are shown in Figure 4.4.



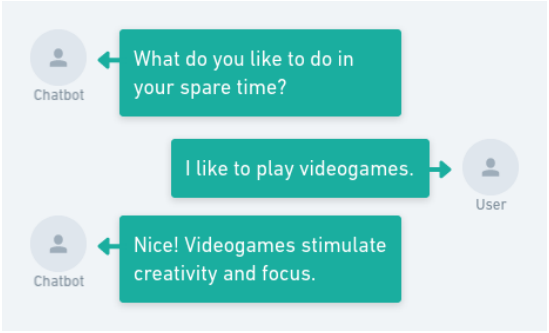
(a) Self-introduction



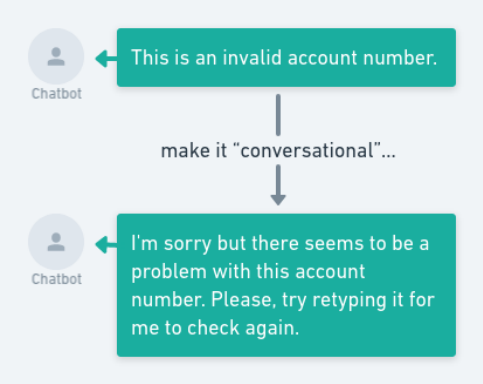
(b) Address user by name



(c) Small talk or chitchat



(d) Echoing responses



(e) Casual language

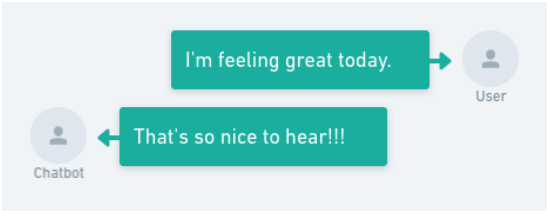
Figure 4.3: Examples of each practice that enforces *Naturalness*

Exclamatory feedback Leverage punctuation to convey emotions that a simple message could not convey. For example, you can use many exclamation marks to convey excitement and question marks to convey incredulity. However, be careful not to use question marks that make users feel contradicted.

Graphical media Images, GIFs, memes, and emojis can help make the conversation lighter and more fun and reinforce the expressiveness of messages. However, be careful not to use media that can have multiple meanings or that can be offensive.

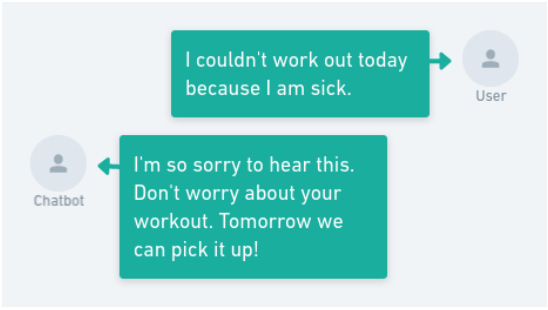
Emphatic messages Seek to adapt answers according to the user's feeling, such as being sorry when the user feels bad about something or expressing happiness when the user achieves some goal.

Humor No one likes to talk to someone low in spirits. Add a little bit of humor by using jokes or funny stories consciously and at the right times. Do not use derogatory jokes of any kind, not even about the chatbot itself.

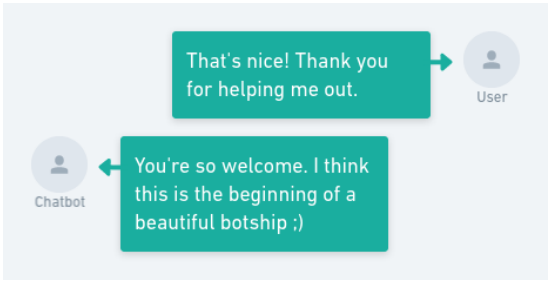


(a) Exclamatory feedback

(b) Graphical media



(c) Emphatic messages



(d) Humor

Figure 4.4: Examples of each practice that enforces *Emotionality*

4.3.3 Transparency

Making the chatbot honest and clear about itself and what is going on in the conversation helps to set the right expectations for users. High expectations can be detrimental to user

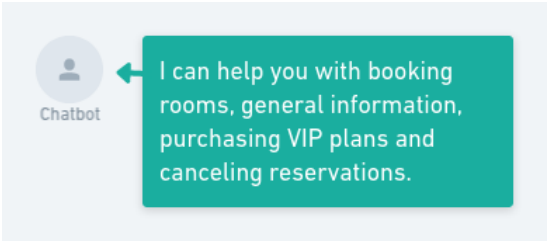
experience when they are not met due to chatbot limitations. Practices are presented below and examples are shown in Figure 4.5.

Present capabilities Start conversations by presenting the purpose and capabilities of the chatbot as a way of guiding users to ask the right questions and avoiding breakdowns in conversations.

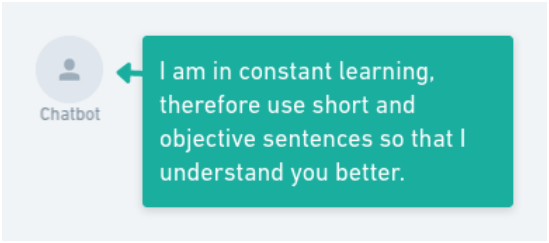
Acknowledge limitations Limitations can be presented at the start of conversations alongside capabilities or acknowledged after a failure. This avoids unwanted questions from users.

Make suggestions Chatbots are powered by knowledge bases that users are unaware of. Provide options of conversation topics to help users to stay on track with what the chatbot can do.

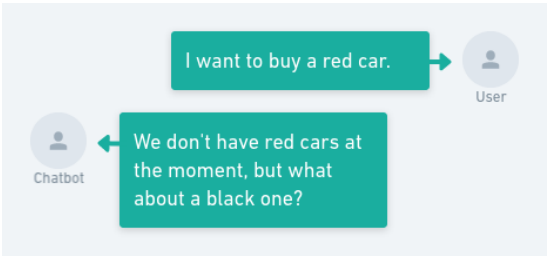
Ask for clarification Sometimes, chatbots fail to understand something present in the knowledge base. Therefore, it is always a good idea to first ask the user to rephrase something that has not been understood before acknowledging a total failure.



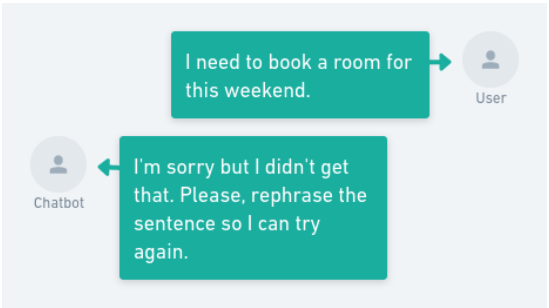
(a) Present capabilities



(b) Acknowledge limitations



(c) Make suggestions



(d) Ask for clarification

Figure 4.5: Examples of each practice that enforces *Transparency*

4.3.4 What to avoid

There are some practices that can invalidate the positive effects of the practices that we have listed on the previous pages. Therefore, here are some behaviors to avoid when designing conversations for chatbots. Practices to avoid are presented below and examples are shown in Figure 4.6.

Repetitive messages Even if you follow the practices in this guide, if the chatbot's responses are repetitive, the chatbot will look robotic, which is what we are trying to avoid here. Therefore, alternate the ways of saying the same thing.

Exaggeration Any conversational practice used with exaggeration may have opposite effects than the ones expected. Use common sense when using each practice, and be attentive to your audience's characteristics to adjust the use of practices.

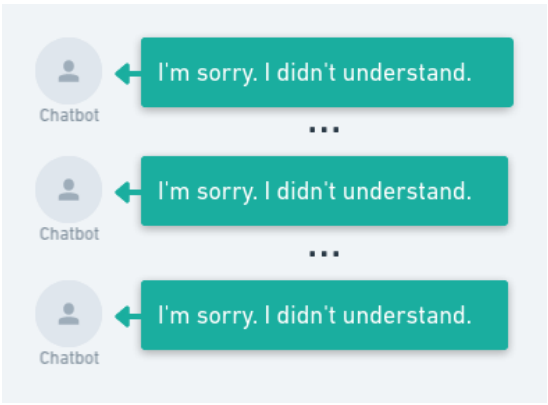
Hiding non-human identity Humans do not trust machines as they trust other humans. Therefore, there is a natural lack of trust in chatbots. However, it is very hard for a chatbot to mimic a human in every aspect and users likely suspect of a chatbot pretending to be human. Once they discover a chatbot is not a human agent, they will feel deceived and angry, and the impact is a lot worse than just disclosing the chatbot's identity as non-human right away.

Machine-like typeface When opting for typefaces for the chatbot's responses, avoid typefaces that can be associated with machines, such as OCR and typewriter font families.

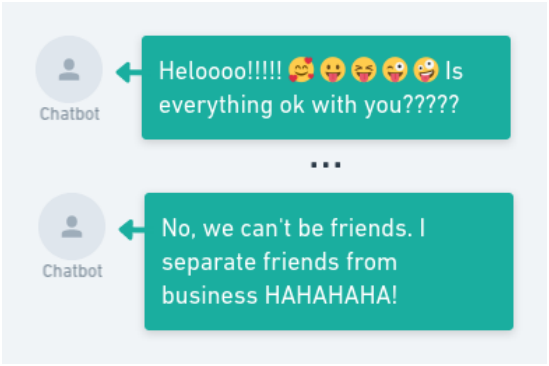
Forcing errors One may think that deliberately inserting errors in conversations, such as typos and slang, may increase the perception of humanness coming from the chatbot. However, users expect chatbots to be error-free, and faking errors cause a bad impression on them and makes the chatbot looks unprofessional.

4.4 Chapter Summary

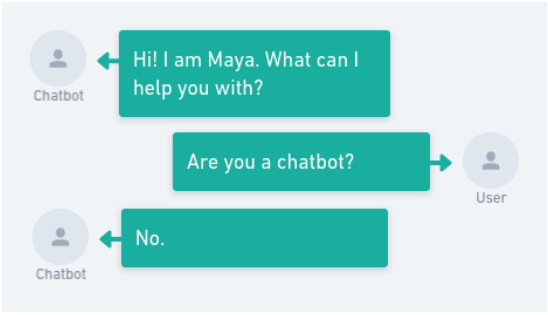
This chapter presented the conceptualization, structure, and composition of the GCCD guide. The SLR results served as the basis for creating a conceptual map that links chatbot purposes to the types of relationships they must build with their users. The map also establishes that these relationships can be enforced through some conversational practices, which were grouped into three objectives: naturalness, emotionality, and transparency. Moreover, it presents some practices that should be avoided. Finally, the conceptual



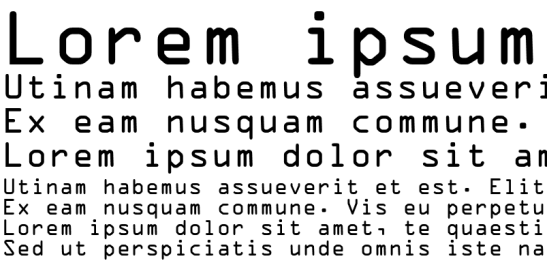
(a) Repetitive messages



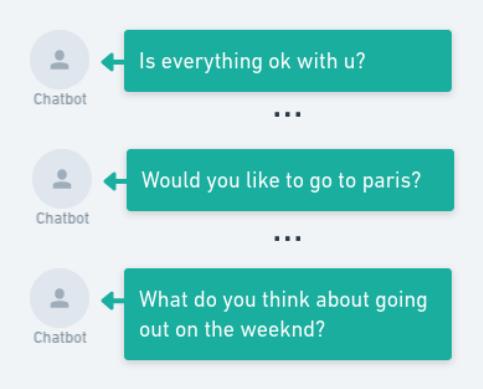
(b) Exaggeration



(c) Hiding non-human identity



(d) Machine-like typeface



(e) Forcing errors

Figure 4.6: Examples of each practice that should be avoided

map was the basis for constructing the guide, which was developed as a web application that exposes, explains, and exemplifies each conversational practice with an accessible language and presentation.

Chapter 5

Guide Validation — Survey

This chapter details the survey used to quantitatively validate the first version of the GCCD guide. It depicts the survey settings, participants, results, discussion and the improvements made to guide according to the results. This validation aims to assess the guide's ease of use and usefulness based on its reading by survey participants.

5.1 Survey Settings

The questionnaire was built and distributed through the Google Forms platform, as seen in our Zenodo repository [32], and required a time between 15 and 20 minutes to complete, considering the reading of the guide as a requirement for its completion. Participants were recruited primarily through personal contacts who do or have worked/researched/studied software development. Then, we shared the invitation on social networks and email lists targeting software development practitioners/researchers/students, emphasizing that the survey was "aimed at tech professionals who have already researched/worked with chatbots or may do so in the future".

Participants had to consent that participation was anonymous, voluntary, and with the exclusive purpose of contributing to the success of the research, in addition to the fact that the responses collected could be stored in perpetuity, which could be used anytime for journal publications, conferences, and blog posts. Moreover, they could leave the survey any time before clicking the send button without any discomfort since the process of responding was unsupervised. Finally, an email was provided in case participants had any problems or questions to the researchers.

The survey was initially conceived in Portuguese since it is the mother language of researchers and, consequently, many of the invited participants. However, it was also shared in English to reach a wider public. It was only necessary for the participant to select the language on the first page of the survey, and the following pages would appear

in the selected language. Besides agreeing to the terms of consent, participants had to confirm that they had read the guide completely before proceeding to questions about it. Therefore, we also developed versions of the guide in Portuguese and English, and the proper link would appear to participants according to their chosen language. Both versions are available in our Zenodo repository [32].

5.2 Survey Questions

The survey had questions approaching participants' general experience, experience with chatbots, and their perception of the guide's usefulness and ease of understanding. The original questionnaire from Google Forms is available in our Zenodo repository [32]. For practicality, the wording of the survey questions is transcribed below:

(Q1) What is your educational level?

(Q2) What is your current main occupation?

(Q3) Have you ever researched or worked with chatbots?

(Q4) Are you currently researching or working with chatbots?

(Q5) What is your level of experience or knowledge of chatbot development?

(Q6) Mark how much you agree with each statement [usefulness].

- (QU1) Using GCCD would enable me to design a chatbot more quickly.
- (QU2) Using GCCD would make it easier to design chatbots.
- (QU3) Using GCCD would make me design chatbots that induce greater user satisfaction.
- (QU4) Using GCCD would make me design chatbots that induce greater user engagement.
- (QU5) I would use GCCD for designing a chatbot.

(Q7) Mark how much you agree with each statement [ease of use].

- (QEU1) I find GCCD easy to use.
- (QEU2) I find GCCD clear and understandable.
- (QEU3) I find GCCD flexible to be used with chatbots from different domains.
- (QEU4) I consider that GCCD requires a lot of knowledge about chatbots to be understandable.

(Q8) In your opinion, what are the strengths of GCCD?

(Q9) In your opinion, what are the weaknesses of GCCD?

(Q10) Is there anything you would change in GCCD? If yes, please explain.

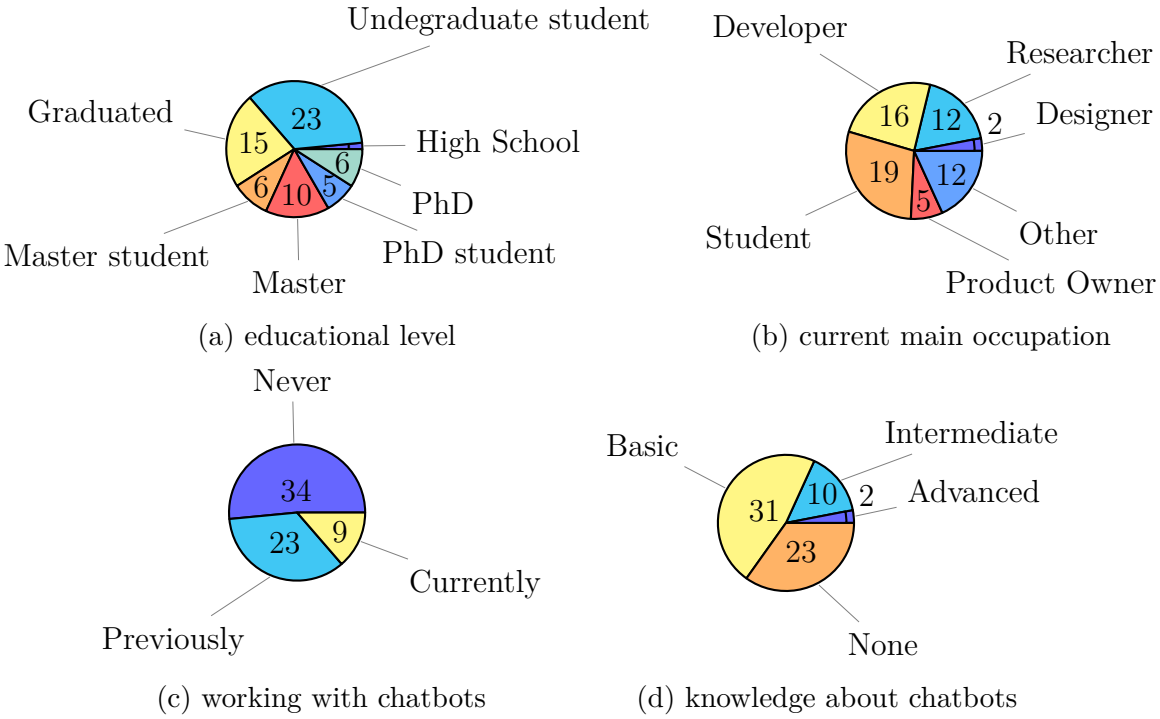
From Q1 to Q5, respondents were questioned about their profiles and had to select only one from pre-defined options. These questions were included to verify the diversity of the sample regarding participants' general experience and the roles they assumed or could assume in chatbot development. Finally, from Q8 to Q10, respondents had an open field at their disposal for complete answers, which can help to understand deviant values in closed questions, if necessary.

Q6 and Q7 were composed of statements in which users should opt for one number ranging from 1 to 5, representing a Likert scale of agreement (i.e., strongly disagree, disagree, neither agree nor disagree, agree and strongly agree). These statements were based on the Technology Acceptance Model (TAM) proposed by Davis [79], which is a model that measures the degree to which a person believes that using the guide will improve their performance (usefulness) and that it will not involve an unreasonable effort (ease of use). TAM is suitable because it has been widely used as a validation tool [80] and has been used in other similar studies to evaluate software design guidelines [81, 82, 83, 84, 85] and development guidelines [86, 87].

5.3 Results

The survey collected 66 responses, of which four came from the English version and the rest from the Portuguese version. All answers are available in our Zenodo repository [32]. Figure 5.1 depicts the profile of participants according to their responses to questions Q1 to Q5, in which Q4 and Q5 were condensed in the chart (d). It is possible to notice that the sample is very diverse regarding respondents' educational level and main occupation at the response time. Regarding their experience with chatbots, many have not experienced chatbot development and have only acquired basic knowledge. However, we have an appropriate amount of respondents with intermediate or advanced knowledge representing more experienced professionals.

Concerning respondents' perceptions of the guide's usefulness, Figure 5.2 shows that around 89% of respondents agreed on some level that the proposed guide would induce greater user satisfaction (QU3) and 83% that it would induce greater user engagement (QU4). Their perception is aligned with our SLR findings since we selected conversational design practices with positive impacts on users. Consequently, around 85% also agreed on some level that they would use the guide to design a chatbot (QU5).



(a) educational level

(b) current main occupation

(c) working with chatbots

(d) knowledge about chatbots

Figure 5.1: Profile of survey respondents.

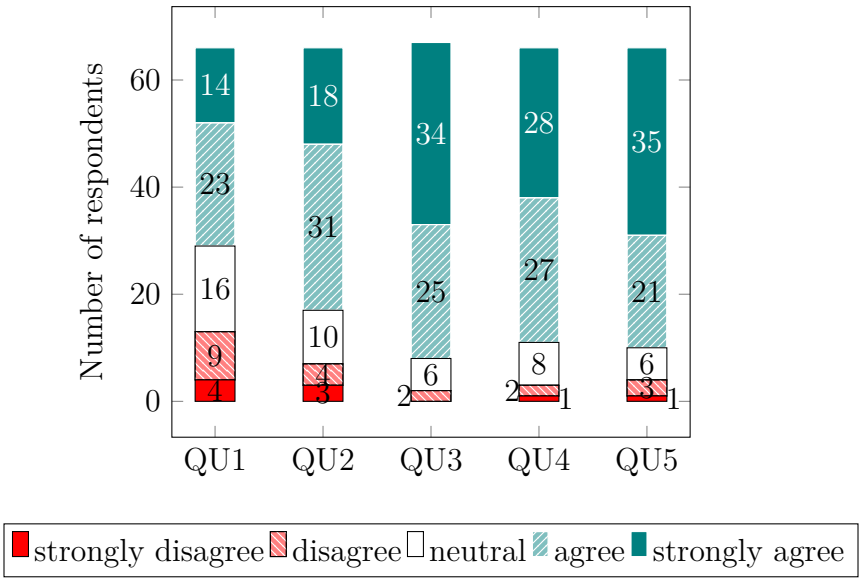


Figure 5.2: Respondents' level of agreement about GCCD'S usefulness to (QU1) quicken design; (QU2) facilitate design; (QU3) induce greater user satisfaction; (QU4) induce greater user engagement; and (QU5) if they would use it.

Although all aspects of usefulness had a majority of agreement, QU1 and QU2 were slightly less favorable than the other aspects. We analyzed responses from the open questions Q8 and Q9 to understand why. Some respondents are concerned with the technical difficulties of implementing these guidelines, which goes against quickening and facilitating design, which QU1 and QU2 respectively measure.

Figure 5.3 presents respondents' perceptions about the ease of use of the guidelines. Around 86% of respondents agreed on some level that the guidelines are clear and understandable (QEU2). Similarly, the vast majority also agreed on some level that the guidelines are flexible to be used with chatbots from different domains (QEU3), which confirms that the guidelines are broad enough.

In line with the percentage of agreement regarding ease of use, around 67% disagreed on some level that it requires much knowledge about chatbots to put the guide into practice (QEU4), meaning that respondents believe that starters would not have problems understanding the guidelines. This confirms that the guidelines are clear enough for all audiences.

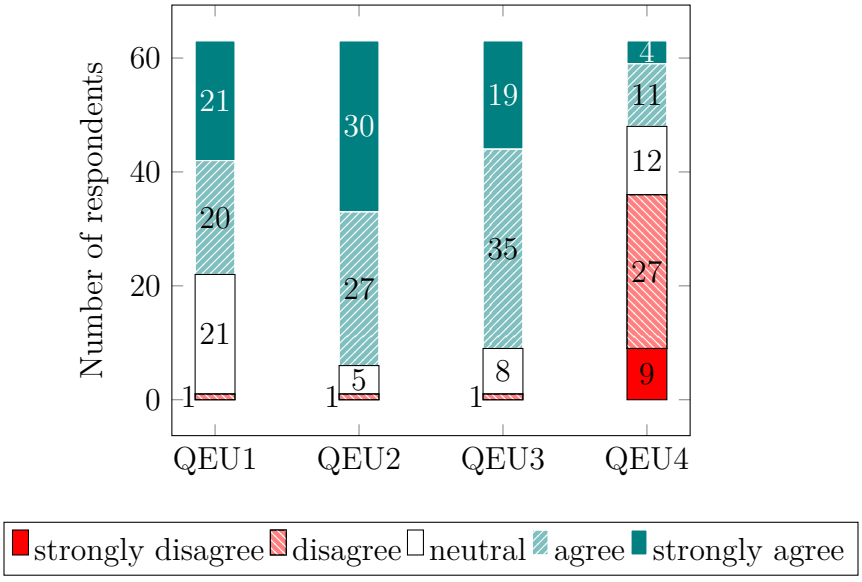


Figure 5.3: Respondents' level of agreement about GCCD'S ease of use regarding it being (QEU1) easy to use; (QEU2) clear and understandable; (QEU3) flexible to be used with chatbots from different domains; and (QEU4) it requiring a lot of knowledge about chatbots to be understandable.

Still, in Figure 5.3, it is possible to notice in QEU1 that the majority also agreed that the guidelines are easy to use. However, it has a more significant amount of neutral responses, which is also a reflection of respondents' concerns about the technical requirements to implement the guidelines, as discussed before, which implies a higher level of difficulty.

Responses from Q8 to Q10 are written comments about the guide that can help uncover possibilities for improvement. Starting with the guide's strengths, the examples for each practice were the high point for most respondents since they helped them visualize how it could be implemented. Other strengths mentioned were the objectivity and clearness of the guide, which reinforced the results shown by the TAM questions.

On the other hand, simplicity was seen as the main weakness from the point of view of most respondents. They missed a deeper approach to conversational design processes and technological implementation of guidelines. Other weaknesses mentioned by participants were the simplicity of the guide's web design and the theoretical nature, which could lead to technical difficulties in implementing it.

In line with what was said regarding the guide's weaknesses, suggestions for improvement are mostly related to the need for more in-depth examples, applications of the guide, and references of implementation to give more credibility. On the other hand, around 20 respondents did not have any suggestions for improvement because they considered that the guide fulfills its purpose or did not feel able to contribute. Table 5.1 presents a list of selected transcripts that support the main points we just presented. At least 3 participants mentioned each of these points.

Lastly, we calculated the Relative Strength Index to measure the degree of agreement for each TAM question. Moreover, we ran the Fisher's Exact Test to verify if there are significant differences among responses from participants with different experiences in chatbot development. These tests were chosen based on the work of Silveira et al. [81], which also used these calculations over a TAM questionnaire to evaluate usability design guidelines for monitoring interfaces.

As proposed by Wilder [88] and adapted by Silveira et al. [81], the Relative Strength Index (RSI) is shown in Equation 5.1, in which Ag refers to the frequency of responses of agreement (i.e. agree and strongly agree) and Dis to the frequency of responses of disagreement (i.e. disagree and strongly disagree). After calculating the RSI, the results can be labeled according to an interpretation of values [81], as seen in Table 5.2.

$$RSI = 100 - \frac{100}{\frac{Ag}{Dis} + 1} \tag{5.1}$$

For the Fisher's Exact Test, we divided the participants into two groups based on their answers to Q3 to check if there is a significant difference between responses from participants who have worked or researched with chatbots compared with those who did not. The calculation was done through a web tool [89]. Table 5.3 shows that, for all of the TAM questions, the p-value is above 0.05, indicating that participants' previous experience with chatbots did not have a significant influence on their responses.

Table 5.1: Transcripts of participants' responses supporting the guide's strengths and weaknesses.

	Construct	Transcripts
Strengths	Examples	"[...] the content is always presented with examples, which makes reading more dynamic and makes it much easier to understand." "The examples exposed in the guide help the reader to better understand the concepts." "The explanation of the concepts accompanied by examples of use."
	Clearness and Objectivity	"It presents in an organized and methodical way seemingly obvious questions about the expected behavior of a chatbot, but which can be easily ignored/forgotten when developing a chatbot, especially by inexperienced developers." "They created a clear and concise explanation of chatbot development and at the same time provide guidance for developers." "The clear and direct way of conveying information that is useful to build the chatbot."
	Presentation	"Friendly design, separation of information into topics" "Simple presentation and visual content, in my opinion, are items that allow for an efficient learning curve." "The fact that it is presented on different pages also makes reading more dynamic."
	Applicability	"The guidelines adopted are independent of the technology used in the construction of the chatbot." "A good summary, it is very useful to pass knowledge to beginners." "Provides a checklist of what to keep in mind when developing a chatbot"
Weaknesses and Improvements	Simplicity	"It is good to be summarized information, however, not all points are clear to those who read without knowledge. If you are going to fix this, it is interesting to pay attention not to miss the great positive point of having compact information." "GCCD touches more on the theoretical part of development, which sometimes ends up being a deterrent for fast development projects." "It is very generic and does not mention technical procedures, even if simplified, recommended for the adoption of practices."
	In-depth examples	"It could present an example of a chatbot that uses/used the DDCC." "[...] would make a single deeper example in some domain of chatbots application" "[...] it lacks more robust examples that can make it more understandable what problems/possible harm might occur if the proposed design guidelines are ignored."
	References	"Perhaps the inclusion of result references, such as research in the area of psychology, etc." "I felt a lack of bibliographic references to support claims about how the DDCC would make it easier to adapt a chatbot." "It is good as a general guide but needs to expand and be more research-based as it seems to assume things or not reference sources about what it promises"

There were some interesting suggestions from the survey's respondents that could be implemented without losing objectivity, which is one of the guide's strengths. One of them is adding references to the SLR papers or this paper to justify the practices and make them more credible. Moreover, many respondents missed a more extended example of the practices. In this sense, it is possible to add another page showing a fictional chatbot

Table 5.2: Data for the calculation of the degree of the agreement through the Relative Strength Index (RSI) for each TAM question.

ID	Ag	Dis	RSI	Interpretation
QU1	37	13	74.0	Moderate agreement
QU2	49	7	87.5	Substantial agreement
QU3	59	2	96.7	Very strong agreement
QU4	55	3	94.8	Very strong agreement
QU5	56	4	93.3	Very strong agreement
QEU1	41	1	97.6	Very strong agreement
QEU2	57	1	98.3	Very strong agreement
QEU3	54	1	98.2	Very strong agreement
QEU4	15	36	29.4	Moderate disagreement

Ag=frequency of responses of agreement
Dis=frequency of responses of disagreement

Table 5.3: Data for the calculation of p-value through Fisher's Exact Test comparing participants with or without previous research or working experience with chatbots.

ID	Experience	SD	D	N	A	SA	p-value
QU1	Yes	2	5	11	9	5	0.34
	No	2	4	5	14	9	
QU2	Yes	2	2	5	14	9	0.97
	No	1	2	5	17	9	
QU3	Yes	0	2	3	10	17	0.45
	No	0	0	3	15	16	
QU4	Yes	1	1	5	10	15	0.47
	No	0	1	3	17	13	
QU5	Yes	1	3	2	11	15	0.26
	No	0	0	4	10	20	
QE1	Yes	0	1	12	9	10	0.58
	No	0	0	9	13	12	
QE2	Yes	0	1	2	15	14	0.87
	No	0	0	3	14	17	
QE3	Yes	0	1	4	17	10	0.91
	No	0	0	4	20	10	
QE4	Yes	4	12	8	6	2	0.73
	No	6	15	4	6	3	

SD=strongly disagree D=disagree N=neither agree nor disagree A=agree SA=strongly agree

that was not designed with the guidelines and another version of it improved with the proposed guidelines.

Some comments about aspects beyond our scope indicated that many respondents did not understand the primary goal. Our proposed guide does not intend to teach conversational design from scratch or get into technical details but to present practices beneficial

to chatbot conversations. This misunderstanding can be mitigated by better explaining the guide's objective on the first pages of the guide and establishing its limitations.

Overall, all aspects regarding the usefulness and ease of use had positive results considering responses to the TAM questions. Considering the diversity of the respondent's level of education, we can infer that our proposed guide is helpful in both academic and industrial settings. Moreover, the lack of statistical influence of respondents' previous experiences with chatbots in their responses indicates that it can serve as a starting point for novices and help improve the chatbot development for experienced developers or designers.

5.4 Guide Improvements

According to participants' open responses to the survey, we made one minor change and two major additions to the guide. The minor one is a change in the home page to include the guide's purpose: "[...] it does not intend to teach conversational design from scratch or get into technical details but to present practices beneficial to chatbot conversations". This sentence was added in response to some participants missing out-of-scope contents in the guide. Furthermore, there were a few more minor textual corrections that did not change or add new meanings to the guide to accommodate the new content and keep textual correctness.

Regarding the major additions, the first one was a page of references listing the SLR selected papers as well as the paper published in the International Journal of Human-Computer Interaction as the source of information for the guide. This addition came to address participants' concerns about the guide's credibility. The second one was a page that exemplifies the use of the guide in a fictional airport chatbot, whose main objective is to present the flight status. This page addresses participants' claims for an in-depth and practical example of use. These new pages can be seen in our Zenodo repository with the complete second version of the guide [32].

5.5 Chapter Summary

This chapter detailed the methodology and results of GCCD's survey validation. We invited software development practitioners and academics to answer a survey based on the TAM questionnaire to validate if the guide is easy to use and understandable, besides collecting suggestions for improving these aspects, if necessary. The survey results revealed satisfactory scores for the guide's usability and understandability. In addition,

participants contributed with some suggestions that were incorporated, such as adding more robust examples and references that support the guide's practices.

Chapter 6

Guide Validation — Case Study

This chapter details the case study used to qualitatively validate the first version of the GCCD guide. It depicts the case study settings, participants, results, discussion, and the improvements made to guide according to the results. This validation aims to assess the guide's ease of use and usefulness based on its practical application in a situation that simulates a real scenario.

6.1 Case Study Settings

The case study involves participants designing a conversational flow for a fictional meditation chatbot. In short, the case study requires them to design first without and then with the guide, generating two conversations to be analyzed as well as the participants' perceptions about using the guide to produce this conversation. Unlike the survey, the case study has fewer participants as it intends to provide a more detailed and individualized analysis of the results of each participant. Therefore, more rigorous criteria are used in choosing the profile of each participant than in the survey.

Since the case study requires more time to be executed and more supervision than the survey, participants were gradually recruited to participate, with more individuals being invited as some would finish it. Their participation was voluntary, and they were assured that they would not be judged or subject to evaluation at any stage. Furthermore, they were asked to complete the stages according to their skills and knowledge, with attention, sincerity, and commitment, and they could withdraw from participating at any time. We provided them with a document containing these sayings (seen in Zenodo [32]) and step-by-step instructions to execute each stage. The three stages are detailed below.

6.1.1 Stage 1 — Elaboration of the Conversation Sample

The purpose of this stage was to create a baseline conversation for comparison with the conversation from the next stage. This stage was asynchronous, and there was no time limit for its completion. Participants were told to suppose that they were asked to design a conversational flow for a new chatbot that will be inserted into a meditation app. This chatbot has two functional requirements: (A) set times when the app will send reminders for the user to start meditation; (B) request meditation suggestions based on the feelings or current mood reported to the chatbot by the user.

The task consisted of participants presenting a sample of a conversation between a user and this chatbot, considering the following conditions for this conversation: i) it is the user's first interaction with this chatbot; ii) the conversation has at least one situation regarding functionality (A); iii) the conversation has at least one situation regarding functionality (B); iv) the conversation has between 20 and 30 interactions; v) the conversation has a beginning, middle and end, that is, without abrupt termination and the user has his requests attended successfully.

The conversation could be provided as a simple text, as the template presented to them. They were required to develop the best conversation for a potential user according to the proposed requirements and their current knowledge, without external consultation to any related material. In the end, they were required to send the resulting conversation sample to the researcher.

6.1.2 Stage 2 — Application of the GCCD Guide

The purpose of this stage was to verify if the proposed guide induced participants to improve the conversation from Stage 1. This stage was asynchronous, and there was no time limit for its completion. Participants were given the link to the GCCD guide (to see it for the first time, since having seen it before was an elimination criterion for participating) and were asked to read it carefully. After reading it, if they thought there was room for improvement, they could change the conversation from Stage 1. If they made changes, they were asked to send the new conversation to the researcher as they did for the first one. Otherwise, they should reply "Stage complete, no changes".

6.1.3 Stage 3 — Interview

The interview was synchronous, recorded, and conducted individually with each participant. Participants did not need to open their cameras since it was necessary to make them comfortable enough to participate and speak as sincerely as possible. Likewise, the researcher's camera was not opened to avoid influencing their responses because of

involuntary non-verbal perceptions. Participants were assured that the audio would not be disclosed, only its transcription, and any information that could identify them would be anonymized.

The interview was planned as semi-structured so that the researcher had previously thought questions available. However, depending on the participant's responses, there was room for additional in-depth questions if necessary. The pre-planned interview script is presented below.

1. Questions about profile:
   (a) Tell me a little about your professional experience as a developer/analyst.
      i. Academic/Educational background;
      ii. Working experiences;
      iii. Skills;
      iv. If you have already participated indirectly/directly in the development of chatbots;
   (b) Talk a little about the contact you had with chatbots as a user.
      i. How was it and what were your impressions?
      ii. What are your expectations when talking to a chatbot?
2. Questions about Stage 1:
   (a) Address the challenges and easiness you faced to design the conversation in Stage 1.
   (b) What previous knowledge and experiences helped you to perform this stage?
3. Questions about Stage 2:
   (a) Address the challenges and easiness you faced while reading and applying the guide in Stage 2.
4. Questions comparing conversations from Stages 1 and 2:
   (a) *(If they delivered changes in Stage 2)*
      i. What were the most or least useful practices for you, considering this topic of conversation?
      ii. In other use cases or themes, do you believe that the set of more or less useful practices of the guide would change when compared to the theme you were given?

iii. Do you believe that some practice would not be useful in any case or topic?

iv. In your opinion, what are the main differences between the conversations you created in Stage 1 and in Stage 2?

A. *(If you mentioned few/no difference(s))* what factors contribute to you not inserting large differences in the second elaborated conversation?

B. *(If you cited many differences)* What factors contribute to you making enough changes to generate this difference between conversations?

v. In your opinion, what would be the perceptions of potential users of the proposed chatbot regarding the conversation of Stage 1 and Stage 2?

(b) *(If they did NOT deliver changes in Stage 2)*

i. What factors led you to believe that the step flow did not need changes in Stage 2?

ii. Do you believe your conversation from Stage 1 fully complies with the guide?

A. *(If not)* Why did you not choose to make changes to comply with the guide?

iii. What were the most or least useful practices for you, considering this topic of conversation?

iv. In other use cases or themes, do you believe that the set of more or less useful practices and the level of difficulty of application of the guide would change when compared to the theme you were given?

v. Do you believe that some practice would not be useful in any case or topic?

5. Questions about the guide:

(a) If you were to develop a chatbot in the future, how and when would you use the guide in the development process, or what would be the determining factors for not using it?

(b) What do you think are the guide's strengths and weaknesses? And in the case of the weak ones, would you have any suggestions for improvement?

In total, 10 volunteers participated in this case study, all Brazilian. We invited individuals working as programmers, UI/UX analysts, or requirements analysts, as these are the positions most likely to use the proposed guide in a real scenario of chatbot development. In reality, we hypothesize that UI/UX and requirements analysts are the ideal users. However, teams do not always afford such positions, and programmers assume these duties. Table 6.1 presents an overview of participants' profiles. We assembled a very diverse sample regarding their backgrounds and age. Although we had only three women,

ID	Gender	Age	Position	Experience	Highest Education
P1	Male	30	Software Developer	10 years	Master in Informatics
P2	Female	24	Requirement/UX Analyst	2 years	Bachelor in Biotechnology
P3	Male	25	Software Developer	3 years	Bachelor in Computer Science
P4	Female	24	Fullstack Engineer	4 years	Bachelor in Software Engineering
P5	Male	54	IT Coordinator/Developer	30 years	Master in Applied Computing
P6	Male	39	IT Analyst	2 years	Master in Biotechnology
P7	Male	25	Frontend Engineer	4 years	Bachelor in Computer Science
P8	Male	41	Software Developer	19 years	Master in Applied Computing
P9	Male	24	Frontend Developer	3 years	Bachelor in Computation
P10	Female	24	Fullstack Engineer	5 years	Bachelor in Software Engineering

Table 6.1: Profile of the participants of the case study.

it is in accordance with the lack of female representativeness in software engineering and computer science [90, 91].

6.2 Methodology for Transcript Analysis

This case study generated artifacts that are available in our Zenodo repository [32]. For each participant, we had: a conversation from Stage 1, a conversation from Stage 2, and an interview transcript. The interviews were recorded in Microsoft Teams and were automatically transcribed by Sonix.ai[1], which also provided features for editing the transcripts for clarity.

The edition of the transcripts was limited to making the text comprehensible to the readers, given that the automatic transcription may fail at some points or the exact transcription of passages may not be comprehensible. Although Sonix's automated transcription was "Very confident" in at least 85% of each interview, it was necessary to review the transcripts from beginning to end in search of the passages that the tool could not capture correctly. Once these passages were found, the editor would listen to the audio and transcribe the passage manually.

In the last step of the edition, a round of reading was done without listening to the audio in pursuit of nonsense snippets. Although the previous step guaranteed the accuracy of the transcription, sometimes, the text would not make sense to a reader as it did in the recording. These cases were corrected solely when written comprehension was affected, using the following strategies: use of punctuation; removal of meaningless words or sentences, such as "eee" or "aaahnn" during thinking; removal of duplicated expressions during speech reasoning, such as in "a chatbot for... a chatbot for an airport"; removal of unfinished sentences due to dialog reasoning change; removal of out of context

[1]https://sonix.ai/

Question	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
Has developed a chatbot?	x	x	x	x	x	x	x	x	✓	✓
Has interacted with chatbots as a user?	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Made changes to conversation 1?	x	✓	✓	x	✓	✓	✓	✓	✓	✓
Thought that complied with the guide?	✓	-	-	✓	-	-	-	-	-	-
Would use the GCCD guide?	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Made suggestions for improvement?	x	✓	x	✓	x	✓	✓	✓	x	✓

Table 6.2: Summary of participants' answers to objective questions.

dialogue, such as problems with the call; addition of clarifications inside squared brackets; participant anonymization.

We ran an objective, narrative and thematic analysis on the transcripts. The objective analysis sought to give an overview of questions that could be answered with mere *yes* or *no*. In contrast, the other analysis sought to understand deeply participants' answers that approached many aspects. Narrative analysis is a qualitative method that consists of reading field texts taking into account the other aspects that permeate these texts, such as the context and collection environment, to retell the "story" from an analytical point of view [92, 93]. Thematic analysis is also a qualitative method, but it is more concerned with identifying patterns through coding and classifying field texts to make unified inferences from the varied sources [94]. The results of the three methods are presented in the following sections.

6.3 Objective Analysis

We have selected some questions that can be quickly answered with *yes* or *no*, based on the interview transcripts and the delivered conversations, to have an overview of the participants' performances in the case study. Table 6.2 presents these questions and their corresponding answers for each participant. Some questions were directly asked in the interview, and others could be easily answered by reading the transcript without depending on interpretation.

First of all, none of the participants had previous experience with chatbot development. Although we would not discard answers from participants with experience, the ones without experience are the most important since they would more likely seek a resource for developing a chatbot, such as the proposed guide, than experienced chatbot developers. On the other hand, all of them have interacted with chatbots, and most have interacted multiple times.

All participants except two chose to change their conversation 1 after reading the guide and delivered conversation 2 in the second stage. Of those, only one made a minor change (one line), whereas the others made significant changes in their conversations. The

two participants that chose not to change their conversations reported that the reason was that they thought their conversations already complied with the guide. Lastly, all participants agreed that they would use the proposed guide to develop a chatbot in the future. However, some had suggestions for improvement, which we will address in detail in another section.

6.4 Narrative Analysis

This analysis considered the following data: participants' profiles collected before (unrecorded) and during the interview, the transcripts, and conversations from Stages 1 and 2. This analysis aims to provide a summarized and descriptive vision of the interviews, based on the delivered conversations and participants' profiles, to serve as a basis for a generalized discussion. The adopted procedure after transcription was adapted from Nasheeda et al. [95]:

1. Chronologically plot: it consists of reading field texts several times in order to become familiar with the timing of events that occurred. In this case, the focused events were the time participants approached specific matters in transcripts and the changes made from conversations 1 to 2 that were delivered by them.

2. Developing the Story: as the title says, it consists of producing the narrative. In this case, after the plot, we thought of a cohesive narrative structure consistent with the interview's events: the participant's profile, changes made to conversation 2, the process of producing the conversations, and the participants' perceptions about the guide. The story is told from a third-person view, objective narrator, based on facts, not on perceptions.

6.4.1 Results

The proposed narrative structure was not rigidly followed but served as a guide to systematically produce the stories for each participant in such a way that their flow is similar. The narratives are presented below in the following sections.

Participant 1

This male participant has around 10 years of software development experience, with a focus on web development, although he also reported working with mobile applications. He graduated in Information Systems, has a master's and is pursuing a Ph.D. in Informatics while also being a programmer in software development projects. His closest

contact with chatbot development was when he included a chatbot provided by another company in an e-commerce he worked on. As a user, he considers that there is a lot of room for improvement in the chatbots he used, as they failed to be objective and solve his problems.

This participant opted not to change his conversation from Stage 1, as he considered that his conversation was already in compliance with the guide. He reported that he based his conversation on previous experiences with chatbots and focused on keeping his conversation objective. For that, in his chatbot instructions, he always gave the user instructions on how the answer should come, such as in the following transcript: "answer only with YES, otherwise, answer I AM REGISTERED".

His other responses showed that the main factor for not changing the conversation was really the existing compliance since he did not identify weaknesses in the guide and thought it was very explanatory. An addendum to this interpretation is that he indicated that he could apply more practices in a more extended conversation or a more complex use case.

Furthermore, he indicated that all practices are useful and can help avoid changes after implementation. For him, another strength of the guide is that it is brief, not too extensive, has examples, and is well-focused. When questioned if he would use the guide in chatbot development, he agreed and mentioned that he would use it to define requirements and make user stories. Lastly, he did not identify weaknesses in the guide.

Participant 2

This female participant has around 2 years of experience being a requirements/UX analyst, our only participant working in this position. She graduated in Biotechnology but is currently working in the position mentioned earlier on a software development project for public organizations. She is also pursuing a master's in Electrical Engineering. She has not participated in the development of any chatbot. As a user, she prefers talking to chatbots only for punctual questions.

This participant changed her conversation from Stage 1 and delivered a new one for Stage 2. Among her changes are: presenting the chatbot as a virtual agent; giving the chatbot a name and explaining to the user that it is there to help them meditate; asking the user for their name and referring to them using it; making the chatbot mourn because the user reported being sad; expanding many sentences to make them more expressive.

She reported one difficulty in producing the first conversation, which was her lack of knowledge about meditation apps, but she based herself on the knowledge of applications that have meditation but not as their main objective. In the second conversation, the challenge was to know whether or not to apply the practice in her conversation. Although

she found that all practices are useful and relevant, she understands that their use depends on analyzing the context, and she did ponder the impact of using each one in the given context.

Regarding the guide, she found the conceptual map very useful and ended up consciously following the map indications concerning what type of relationship she should build between the chatbot and the user. Based on her own opinion, after reading the guide and making the changes, she felt that her biggest gain was making the conversation more natural without losing objectivity. In the same line of thought, she mentioned that users from conversation 2 would feel much more connected with the chatbot, receiving the feelings that would be induced only by meditation.

As well as P1, she would use the guide in requirements elicitation as a reference but would also refer to the guide during development. Among the guide's strengths, she mentioned the organization, content length, presentation of information, presence of images, and bibliographic references. She also highlighted that the use case shows that making the conversation natural without losing objectivity is possible. Finally, as a suggestion for improvement, she mentioned that the guide's introduction does not have a language as accessible as the rest of the guide. Therefore, her suggestion was to reduce the formality of the introduction to make the guide's language more uniform.

Participant 3

This male participant has around 3 years of software development experience, including mobile and general systems development. He graduated in Computer Science and is currently working as a software developer for the private sector. He has participated in preparing for developing a chatbot at a project level, but it did not move forward. As a user, his experiences with chatbots were mostly negative, and he believes that the user experience can be improved by thinking more and better about the requirements.

This participant changed his conversation from Stage 1 and delivered a new one for Stage 2. Among his changes are: presenting the chatbot as a bot; giving the chatbot a name and explaining to the user that it is there to help them manage meditations; adding personality and expressiveness to some responses such as changing "Ok" to "Consider it done!"; and adding a friendly chatbot response to end of the conversation, as follows "Alright, I'll be here when you need me again :)".

His main concern and difficulty in producing conversation 1 were thinking creatively to produce a natural conversation for the user. He also used previous experiences with chatbots to produce the conversation. In the second conversation, he reported he had no significant difficulties besides creativity and tried to insert more humor, naturalness, and humanness.

Regarding the guide, this participant had similar perceptions as P2: all practices are useful but depend on the context and the practitioner's critical thinking. In his view, the main change from conversations 1 and 2 is that the chatbot gained an identity and started communicating more naturally, making users feel more connected to the chatbot. According to him, these changes were mainly driven by the guide's examples.

He affirmed that he would use the guide in chatbot development, more specifically in the planning phase and after implementation, for continuous improvements. From his point of view, the guide's strengths are the examples and intuitiveness. Although he did not mention weaknesses, he suggested keeping the guide up to date with new information, increasing the number of examples, and giving more guidance on starting the conversation from scratch.

Participant 4

This female participant has around 4 years of experience in mobile and web software development. She graduated in Software Engineering and currently works as a full-stack engineer in the private sector. Although she does not have previous experience with chatbot development, she is currently working with a feature of form filling via a chat interface. As a user, she had lots of experiences with chatbots, but most of the time, she was unsuccessful in solving her problem through them.

The main difficulty she reported having during Stage 1 was conciliating technical implementation with conversational requirements. She worried greatly about how the conversation would be really implemented as software which prevented her from keeping the conversation more natural. She also based herself on her previous experiences with chatbots to produce the conversation. She opted for not changing the conversation on Stage 2 based on her belief that the conversation already complied with the guide.

Differently from the previous participants, this participant did find some practices unhelpful, which were the ones of *Transparency*, due to the lack of objectivity based on her personal experiences. She also affirmed that the use of emotionality depends on the context. Moreover, she thought there could be difficulty in conciliating business requirements with the guide practices.

She confirmed that she would use the guide as it was practical and short. Other positive points, according to her, were the conciseness, examples, and images. Regarding suggestions for improvements, she mentioned a minor usability concern about the guide's request "read here", which was not very explanatory about what "here" really was.

Participant 5

This male participant has around 30 years of software development experience, which makes him our most experienced participant. He graduated outside IT but concluded a master's in Applied Computing and is pursuing a Ph.D. in Informatics. His development experience is concentrated in the public sector, with varied types of software and also occupying management positions. He never worked with chatbot development but has a few experiences as a user, in which he reported having difficulties solving his problems.

This participant significantly altered his conversation from Stage 1. Among his changes for Stage 2 are: presenting the chatbot as an artificial intelligence agent; giving the chatbot a name and explaining to the user that it is there to help them with the service of meditation; removing the request for only "1,2,3" responses from the user to use natural language; adding a situation of breakdown and asking the user to rephrase; making the chatbot explicitly refer to the information given by the user such as "I noticed that you like to wake up early."; using exclamation marks (from "Perfect" to "Perfect!!"); adding personality and expressiveness to some responses; saying goodbye to the user at the end of the conversation. For him, users from conversation 2 would feel more comfortable with the chatbot.

Although he made many improvements in his conversations and noticed that the most significant gain was in fluidity, he reported difficulties in dealing with the chatbot's emotionality, which he attributed to his own personality. Still, his strategy in the first conversation was to use his real-life experience, which made him set a primary goal to make the chatbot as solicitous as possible. As well as P4, he initially struggled with how this conversation would actually be implemented as software, affecting his conversational decisions. However, he reported no difficulties in absorbing the guide's content and applying it, which made him identify in the guide many points that could be improved in his conversation, which made him worried about overdoing the changes.

Regarding the guide's practices, he pointed out a practice as not useful in any situation, which was *Small talk or Chitchat*, based on his own experiences, because he appreciates objectivity. On the other hand, he was most sensitized by the practices from *naturalness*, which he found the most important ones. The factors that most helped him to change the conversation were the guide's practicality, objectivity, concreteness, and examples.

Similarly to the previous participants, he would also use the guide during requirements elicitation. Furthermore, as an interesting contribution, he pointed out that he would even use the guide as a reference for the customer to direct their requests better. Moreover, he thought the guide was an excellent resource for beginners like him.

Participant 6

This male participant has around 2 years of software development experience. He graduated in Information Systems, specialized in Software Engineering, has a master's in Biotechnology, and is pursuing a Ph.D. in Informatics. He works in the public sector as an IT analyst and is also a college professor. His current work involves development and systems integration. He did not have experience with chatbot development but had plenty of interactions with them as a user, with most being positive experiences, unlike previous participants.

After reading the guide, he opted to improve conversation 1 and delivered a new one in Stage 2. His changes were not punctual, as the structure of the conversation and flow of questions were changed almost entirely. The notable changes were related to removing commands that stifled the user's response, allowing the conversation to run with natural language requests from the user. Although he made it more flexible for the user to communicate, he kept the chatbot's expressiveness and language pattern from the first conversation. According to him, *transparency's* practices were the ones that most induced him to improve the conversation since he identified this was the main deficiency of his first conversation.

Since he works closely with the messaging team in his job (they produce system messages sent to the user), he used this experience to make conversation 1. Although his messages were individually expressive and natural, he had trouble making them part of a fluid conversation. In conversation 2, he struggled a little bit to understand the conceptual map at first, but after reading the whole guide, he could understand it. In line with the map, he agrees that the use of *transparency* and *emotionality* will depend on the context.

For him, the main facilitator for improving the conversation was the use case. Although he was initially a little bit confused with the conceptual map, he found that the way the content was presented was really helpful and one of the guide's strengths. After improving the conversation, he thought users would feel like talking to someone, not something like filling out a form. Apart from the conceptual map, he did not mention other weaknesses or points that should be improved.

Participant 7

This male participant has around 4 years of software development experience and he reported having worked closely with UI/UX aspects. He graduated in Computer Science and is pursuing a master's in Informatics and working in a startup as a senior developer.

His experience with chatbots was only as a user, and he had many reservations about their use.

There was just one change in his conversation from Stage 1 to Stage 2. After the user asked for a meditation while also informing that the chatbot was tired, in his second conversation, he added a message from the chatbot that acknowledged the user's feeling as follows: "It's a pity, [username] :(, sometimes routine demands too much from us.".

He followed his life experiences and his own way of talking to produce the first conversation, and no major difficulties were reported. However, he considered that he could not build a close relationship with the user through his first conversation, but the guide helped him improve this aspect. According to him, practices from *What to avoid* are really important, and the most useful ones for his conversation were the ones from *naturalness* and *emotionality*.

As for the guide, he acknowledged that it is an important reference for developers. However, he did not dwell on the guide's strengths, although he mentioned that the guide is intuitive and lean. Lastly, as P1, he suggested that the guide's first page could be less academic and more inviting.

Participant 8

This male participant has around 19 years of software development experience. He graduated in Computer Science, concluded a master's in Applied Computing and is pursuing a Ph.D. in Informatics. He has worked on multiple collaboration projects with public bodies as a software developer. He had some interactions with chatbots as a user, and it was generally a positive experience.

He made a couple of changes to his first conversation: presenting the chatbot as a virtual helper for a given app; addressing the user by its name; echoing user intentions; expressing feelings according to the user's responses. Differently from other developers, he focused only on the conversation and not on how it would be implemented, admitting that perhaps his chatbot was not technically viable.

The only barrier of difficulty he faced was his lack of mastery over the subject of meditation. His source of inspiration for creating the conversation was user support via phone with recorded messages, in which the message asks the user to press a number to continue and change subjects. According to him, the user would feel more comfortable in conversation 2.

He thought that all practices were useful, but in this situation, he used naturalness the most, although he maintained the structure of a guided interaction through menu-based options. The main factor that induced him to change his conversation was the guide's examples. He would also use the guide during requirements elicitation and suggested

improving the page navigation and explaining the types of chatbots for the reader to understand what type this guide is for.

Participant 9

This male participant has around 3 years of software development experience and graduated in Computer Science. He has worked mainly with frontend development and has experience developing a chatbot for WhatsApp, but it was not deployed. He also had multiple interactions with chatbots as a user and he has a positive perception of them.

This participant made only one change to his chatbot's responses, which was adding the purpose of the chatbot. All of his other changes were related to how the user communicated — instead of using requesting with commands using a backslash, such as " addreminder", he changed for natural language requests, such as "Add a reminder". According to him, the biggest gain was in fluidity.

He did not report difficulties in reading and applying the guide. He acknowledged that in his first conversation, he favored ease of development over a more natural conversation, which was incompatible with what the guide was proposing. He considered all practices relevant, and emotionality caught his attention in this specific theme.

He did not identify weaknesses in the guide. Since he had previously developed a chatbot with the intention of deploying it to real users, he was questioned if he would have used the guide if it had been presented to him before. He stated that he would follow the guide because it is a referenced resource for chatbot development as opposed to his personal design decisions. Therefore, the guide would be helpful to him.

Participant 10

This female participant has around 5 years of experience in software development. She graduated in Software Engineering and currently works as a full-stack engineer in the private sector. She developed a chatbot but only for self-learning. As a user, she had many interactions with chatbots but only likes to use them in some situations due to their limitations.

She made three changes to her conversation. The first was making the chatbot acknowledge the user's bad feelings and mourn about them. The second was explaining the purpose of recommending a given meditation, which was to make the user feel better. Lastly, she added a finalization in the service in which the chatbot says goodbye with "See you next time". She acknowledged that the biggest gain in conversation 2 was conveying more emotions to the user.

Her strategy in conversation 1 was to put herself in the user's shoes to understand their needs in the conversation. She also used strategies that she liked in interactions with other

chatbots. However, she had difficulty complying with the two mandatory requirements in the same conversation sample. Regarding stage 2, she had no difficulties understanding the guide's contents, but she thought applying the practices correctly in her conversation was challenging.

She thought that practices from naturalness were the most important for her. Although she used emotionality more than transparency, she reported that she could have used more, which would have been better for the conversation. According to her, all practices are useful, but their selection depends on the domain and public.

For her, the use case was really helpful in showing how to make her conversation more natural and induced her to add more emotions to her chatbot. She said she would use the guide if she developed a chatbot. According to her, the guide's strengths are the very intuitive content presentation, although she found the conceptual map a bit confusing at first and only understood it after reading the guide thoroughly.

6.5 Thematic Analysis

This analysis considered only the transcripts of participants' responses that were self-contained, meaning that single response confirmations were not considered (the interviewer presented some point of view and the participant only confirmed). It aims to reach conclusions about the proposed guide and related aspects based on the codification and association of transcripts excerpts. The adopted procedure after transcription was adapted from Kiger and Varpio [94]:

1. Familiarizing yourself with the data: it consists of reading repeatedly the interview transcripts to get familiarized with the contents. This was actually initiated in the narrative analysis.

2. Generating initial codes: it consists of annotating transcripts excerpts with codes that represent a most basic segment of the raw data that can be assessed in a meaningful way regarding the phenomenon. This was done by assembling participants' responses on a sheet. The coding process started with P1, and the codes from it were reused in the analysis of the following transcripts. If there was an excerpt that did not fit previously created codes, a new one was created. In the end, the list of unique codes was generated, and they were refined in an iterative process if there was ambiguity between them. This sheet is available in our Zenodo repository [32].

3. Searching, defining, and naming themes: consists of examining the codes to find relationships and correlations that can form a theme. The themes derived not only

from the codes themselves, but from the inner findings of transcripts classified as a given code, and were consolidated as thematic maps.

6.5.1 Results

At the end of the coding process using Thematic Analysis [94], there were 32 codes used to classify transcripts, as seen in Table 6.3. Some codes were discarded due to not being relevant to the current analysis. The remaining codes were examined for correlation and theme formation, which was not possible for all codes. The correlations were represented as thematic maps, presented in Figures 6.1-6.5. These maps start on top with a theme represented inside a dark green container, which was generated from the codes below them represented in a light green container. In the blue boxes are the conclusions extracted from excerpts classified with the corresponding code, which helps understand the motivation of the theme.

ID	Code	#	TM1	TM2	TM3	TM4	TM5
C1	concerns in developing conversations	11	✓				
C2	decision to not make changes	3				✓	
C3	difficulties in making conversation 1	14		✓			
C4	difficulties in making conversation 2	11					
C5	experience with chatbot development	7					
C6	factors that induced changes	10					
C7	how they used the conceptual map	2			✓		
C8	improvement suggestion	12					
C9	intention to use the guide	11					
C10	lack of negatives about the guide	2					✓
C11	negative perception of the guide	7					
C12	participant profile	14					
C13	perception about compliance	6				✓	
C14	perception of conversations as a user	8				✓	
C15	perception of difference between conversations	19				✓	
C16	perception of practices as a user	19	✓			✓	
C17	perception of the conceptual map	7			✓		
C18	perception of the impact of practices on development	7					✓
C19	perception of the quality of existing chatbots	14	✓		✓		
C20	perception of the usefulness of practices according to theme	11					
C21	perception of the usefulness of practices in the development	20	✓		✓		
C22	perception of when to use the guide	9					✓
C23	perception on how to use practices	19			✓		
C24	perception on the theme of the chatbot	1					
C25	positive perception of the guide	29					✓
C26	reasons for not having negative perceptions	2					
C27	reasons to not use the guide	4					✓
C28	strategy in developing conversation 1	16	✓	✓	✓		
C29	strategy in developing conversation 2	14	✓	✓			
C30	use of chatbots as a user	9					
C31	when do they prefer to use chatbots	4					
C32	which practices they used	18					

Table 6.3: Codes generated during the thematic analysis, number of transcripts that received each code and thematic maps in which they were used.

Figure 6.1 depicts *Thematic Map 1 (TM1) — Participants sought objectivity when developing chatbot interactions since they also value it as chatbot users.* A total of six codes were used to compose this theme. The interviews revealed that chatbots are not well

received by participants since their experiences with them failed mostly in two aspects: going straight to the point and solving their problems. Therefore, these negative experiences made them prioritize the problem solution based on their own experience, in which they developed a conversation that would be pleasant for themselves as users. As part of this strategy, the practice of "small talk" had some of them uncertain about its usefulness and was avoided by them. However, they very much agreed with the importance of having natural conversations as per the guide.

Figure 6.2 depicts *Thematic Map 2 (TM2) — Those who focused a lot on the technical aspect of implementation had difficulties in applying naturalness on conversation 1 since they prioritized strategies that would be easier for developers to implement.* Particularly for developers, we have seen that they had a concern of making the conversation easy to implement as if they were to develop this chatbot. Although they did not have previous experiences with chatbot developers, they kept the idea that menu-based interactions would be easier to implement since, as developers, they understand that free text input requires a more complex natural language understanding mechanism. Still, they understood that this strategy would hurt naturalness, which was a very well-received construct as part of the guide.

Figure 6.3 depicts *Thematic Map 3 (TM3) — Participants differed significantly in how they used the practices as they relied heavily on their personal experiences.* When questioned about the usefulness of practices and their constructs, the answers very much agreed on the importance and usefulness of naturalness. However, there were some divergences in the use of transparency and emotionality as constructs and some practices from naturalness as well. The interviews revealed that the choice of practices depended on how they envisioned their first conversation, their vision about the theme, their personal experiences with chatbots, and their personality in some cases. The conceptual map was hardly mentioned, and although the mentions were positive, they struggled to use it in this limited conversation of the experiment. Unconsciously or not, the conceptual map was mostly correctly followed as it comprehends the context dependency and it is flexible regarding how much of the constructs should be used.

Figure 6.4 depicts *Thematic Map 4 (TM4) — Participants recognized the importance of the guide for better user-chatbot interaction and ensured that their conversations complied with the guide.* Only two participants decided not to change their conversation based on the belief that it was already in conformance with the guide's practices and principles. The ones who did change their conversations, when questioned about the differences between the conversations, reported that potential users would feel more comfortable, understood, positive feelings, and a closer relationship with the chatbot. As for the main difference from one conversation to another, most reported that the biggest gain was on the fluidity

and naturalness of the conversation.

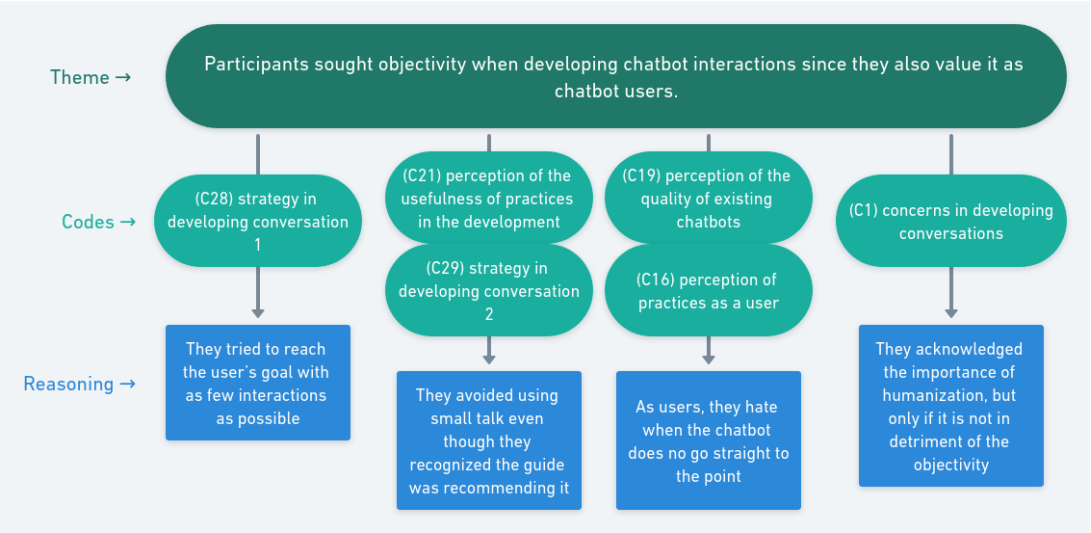
Figure 6.5 depicts *Thematic Map 5 (TM5) — Participants saw the guide as a great reference for all stakeholders to determine conversational requirements, especially for beginners.* Besides demonstrating positive feelings toward the guide, some developers pointed out how the guide would contribute to development teams, such as serving as a training resource and reference for stakeholders since they all envisioned the guide being used during the requirements elicitation and specification phase. The guide's main strengths pointed out by participants were the content presentation since they found it objective and easy to absorb. Although they did not point out weaknesses per si, they suggested some changes, mostly minor ones.

6.6 Discussion

In this case study, we not only evaluated our participants as potential chatbot developers but also sought to understand their experiences as chatbot users. The majority of our participants had bad perceptions and experiences with chatbots, mostly due to the chatbot not being objective, solving their problems, or offering follow-up with a human agent. Some also mentioned they preferred chatbots only for simple tasks. Our findings were very similar to the ones of [96], in which users preferred chatbots for simple and straightforward inquiries, and when their problems were not solved, it was not detrimental to user experience as long as they offered an easy path for following up with a human agent. Still in this line of thinking, other works concluded that the intention to continue using chatbots is affected by their perceived usefulness [97], in other words, users noticing concrete functionality and reliability [98].

Although their previous experiences clearly affected how they designed the conversation samples, it was not possible to find a correlation between participants' age and gender with their design choices. For example, one female participant was overly focused on user experience (P2) and the other was overly focused on the real implementation of her conversation (P4). As for age differences, the older participants (P5, P6, P8) did not present more difficulties in using the guide and developing conversations than the younger ones, and also had the same concerns as his fellow younger developers — implementation. Similarly, experience level was also not a decisive factor in design decisions. Still, since the sample size is small, this does not imply that a correlation does not exist, but it was not evident in this experiment.

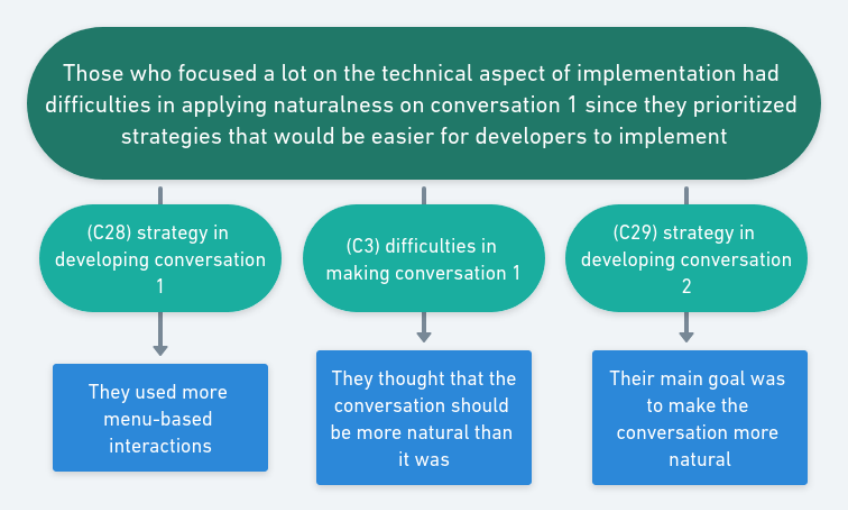
However, the position participants occupy did appear as a moderator variable. It was possible to observe that participants that hold or held positions related to user experience (P2 and P7) had a greater concern about making the conversations more natural and



Transcripts excerpts supporting theme (not extensive)

P1: [00:03:27] I think that some services even approach the question of being charismatic, very receptive, but they take a lot of time in some points to solve what the user really wants. So it turns out that sometimes you do a roundabout of things that are unnecessary to finally get what you want or get close and end up not solving it. This is a negative point I see.
P1: [00:04:25] I tried to be a little more straight to the point, not trying to evade too much of what was the actual objective.
P1: [00:06:13] I tried not to be so mechanical and look like it was a person answering, but in a way that got straight to the point of what the user wanted to do at the moment.
P2: [00:14:36] For example, one thing that worries me a lot, like, with UX, is that, in the name of making the conversation nicer, I need to make the user take a bunch more actions that he didn't need to do before. So, for example, the small talk part is something that worries me sometimes, the chatbot, for example, wanting to talk to you a lot and not doing what you want it to do. [...] It's just one more action and then the rest I don't think changes and that's important to me. So I managed to have a gain in the conversation without having to make the user take many more actions.
P2: [00:22:56] I think that, as a UX, it showed exactly what I said about objectivity, right? That for example, in the first image of the use case it is super robotic and you have a lot more user interaction. And then in the second, which is like the improved one, you have less user interaction and it is more friendly. So this is an important thing for me [...]
P3: [00:03:50] If they had the option I wanted and I couldn't find it, they would force the user to continue talking to the chatbot and make it very difficult to find out how that user could be able to talk to the attendant, who is the one that could really solve my problem. [...] So, in my opinion, I think it has a lot of potential, but the implementations that I used were quite frustrating.
P4: [00:03:38] I think it's more because of attempts to try to solve it and not being able to do it, you know? Using it[chatbots]. Now, if I had the idea, like, I'll talk to him, I'll be able to solve it. Then I think that would be another story.
P4: [00:12:34] Yeah, which is transparency. I think sometimes it can feel like it's actually padding things out or whatever. It's my feeling when I talk to a chatbot.
P5: [00:12:18] I avoid small talk a lot. I like to be... By the way, I don't talk much, I don't talk much, you know? So I don't like the situation very much. So, for me, this first part here of naturalness, I think the only thing I really might not use, these things, would be this small talk.
P7: [00:02:02] So, personally, I tend to look for a more direct use of things.

Figure 6.1: Thematic Map 1 (TM1): Participants sought objectivity when developing chatbot interactions since they also value it as chatbot users.



Transcripts excerpts supporting theme (not extensive)

P4: [00:04:19] I think the biggest challenge is because I ended up confusing the way I interacted with him a bit, thinking like this: "but how would that be done from behind". But that's because I'm a developer. I thought, for example, in some sentences, I think: "it would be better to repeat it as if it were a menu, as if it were going back in the menu". But looking at the guide, we see that it is not good for us to keep repeating it as if it were a menu, because we want a more natural interaction, natural language, you know, as if two people were really talking. P4: [00:05:13] I was thinking like this, if I had been developing this, here I would go back to a previous step.

Q5: [00:05:09] Do I need to install? Does the guy need to have an account, be registered, have some record, something like that? I kept thinking about these things, then I said: "Oh no, I'm going to focus on the conversation itself" because otherwise I would have stayed there, I would have to do about 50,000 iterations there until I reached the point where the person could actually talk to a chatbot calmly. Registering a user, a lot of things, would be a bit complicated.

P5: [00:06:49] After I extrapolated that first part of technicality that I saw that I would not be able to advance without thinking a lot about technical requirements, I went more into this matter of personal experience, that is, what as a user, I would like to receive information from a chatbot.

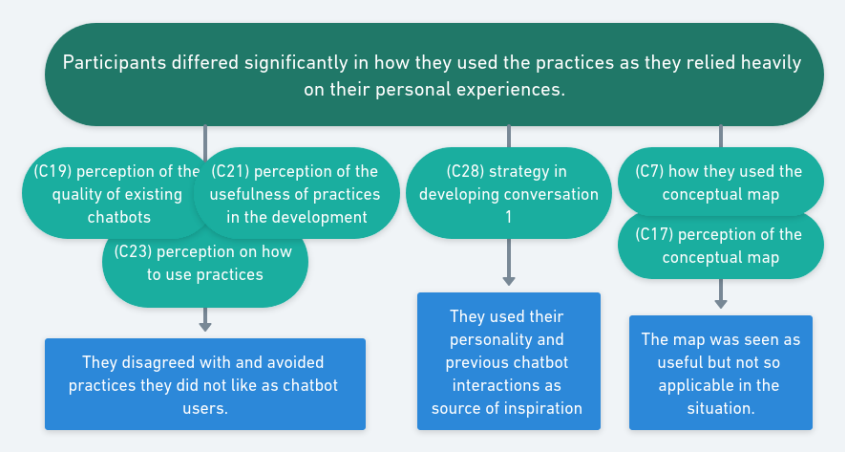
P5: [00:14:48] I think the main thing is precisely the fluidity of the text. I think that after I redid it based on the guide, the conversation became more fluid, more natural, you know? At first, no, it was very robotic, like, it was like this: answer one, two or three, do this or do that. So I think that in the second, if the intention really is for you to have this conversation, for the person to talk, it's okay that he knows it's a chatbot, but to be able to have the conversation in a more natural way, without having a lot of pre-defined things. I think that was the big change. I think that using the guide I managed to evolve a lot in this sense. My conversation became more fluid, almost as if it were two people talking.

P6: [00:04:24] Working with API integration, I have contact with the people who take care of messages, error messages for the user, that kind of thing. So that helped me a little bit, but at the same time it got in the way, because I think it was from this experience that this more mechanical thing came.

P6: [00:14:36] It's still not ideal, but I managed to make the conversation more natural, more fluid, in a way that the person doesn't feel like they were machine, and that they can connect even knowing that they are talking to the machine and could relate in a simpler, more fluid way.

P9: [00:10:15] One thing I had done when I had made the chatbot with WhatsApp was that we had to type a command, back-slash and its name, and at least for me it was intuitive, but not for the user, right? And the chatbot proposal that is inside the guide, I thought it was very easy and that it will make it easier for the user, but I know that it will make it difficult for people who are developers, but I think this was the main reason for me to have changed the txt that I had given you.

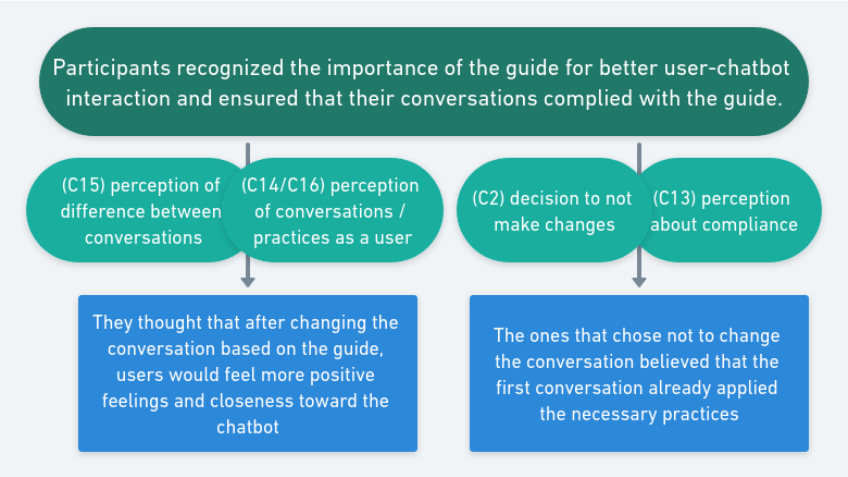
Figure 6.2: Thematic Map 2 (TM2): Those who focused a lot on the technical aspect of implementation had difficulties in applying naturalness on conversation 1 since they prioritized strategies that would be easier for developers to implement.



Transcripts excerpts supporting theme (not extensive)

P2: [00:12:48] Well, there are some here in "What to avoid", right? And then I think the worst of all here, I don't know, I'll see the worst ones here. Repetitive messages. I hate. I hate when this happens to me in chatbot. The exaggeration, which I don't like either.
P2: [00:14:36] So, for example, the small talk part is something that worries me sometimes, the chatbot, for example, wanting to talk to you a lot and not doing what you want it to do.
P2: [00:11:35] I see a lot of humor when it's, for example, a newspaper or when it's dense content that you want to lighten. But I, particularly, don't like humor at all. I'm a humorous person, but sometimes I think that if you use humor too much, it forces you too much. It seems that the person wants to create an intimacy that does not exist. Then maybe I would leave it for longer or more complex conversations with the chatbot towards the end, because then the person has already spent time with you for you to use humor.
P3: [00:10:49] I remember that I took the humor and the self-presentation issue a lot into account, and the issue of, if I'm not mistaken, spelling errors as well.
P4: [00:11:04] Here is, topic three, which is emotionality. I think it is very interesting for this case, mainly because, depending, if it is a person who is doing this for the first time, meditation, in this case, he is probably doing it because he thinks it will solve something, he will be calmer, it will help, it will help her with anxiety or something. So she's probably going through something that she wants to work out, you know? In terms of feelings, probably feeling emotions, so it's interesting. The chatbot tests this point, in this characteristic, of showing that... Empathy, right, it's not really empathy because it's a chatbot, but as if it were showing that it understands the user. And I think that was really important.
P4: [00:12:34] Yeah, which is transparency. I think sometimes it can feel like it's actually padding things out or whatever. It's my feeling when I talk to a chatbot.
P5: [00:15:44] So, as I told you, the part that most touched me was the naturalness part. The transparency part where I thought it was very important, but in a way I had already tried to use it. I think that as a chatbot user, I wouldn't want to be talking to a robot without knowing that he is a robot. I think this issue of transparency is essential. So, I kind of had already internalized that. So, that didn't make me change that part. Emotionality I'm kind of complicated on this issue. I'm not very... this emotional issue is not really my area. So, this naturalness part really and these suggestions of naturalness were really the part that most affected my changes.
P7: [00:06:36] I think I worked a lot on the practice of emotionality both after reading and empirically. And the issue of naturalness, I also felt it was important. Like it or not, I work a lot with the user, I work a lot with the user interface, so I already had that in mind. Maybe not theoretically. But I had it in practice and I need to make it sound like a person at the very least like someone who wants to talk to you and not just a reactive question-and-answer form.

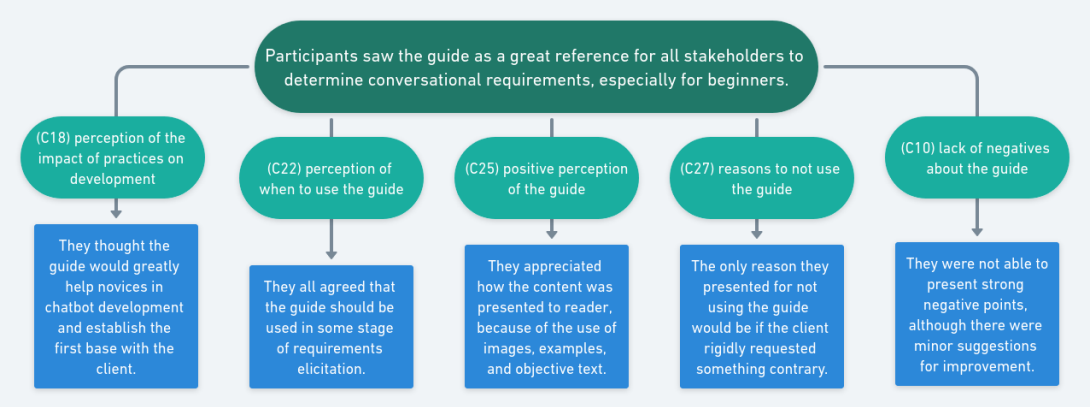
Figure 6.3: Thematic Map 3 (TM3): Participants differed significantly in how they used the practices as they relied heavily on their personal experiences.



Transcripts excerpts supporting theme (not extensive)

P1: [00:05:38] I thought the guide was very explanatory. I really liked it, especially the part where it focuses on what you shouldn't do in a chatbot. I think I managed to follow the closest to this, so I even opted not to edit my chat.
P1: [00:06:44] I believe so. In accordance with the guide.
P2: [00:16:42] I'm going to make a projection here now, because I don't know so many users of meditation, but I imagine that the person who is looking for meditation, for some reason, wants a calm, peaceful scene. She wants to think for a while and relax. So she wants to generate good feelings, that's why she's there. And I think conversation 1 doesn't get into the vibe, let's say. It's as if I had the meditation app and it was going to make me calmer. But I don't feel it when I'm talking to the chatbot. I will start to have this feeling only when the meditation starts. And, in conversation 2, I'm being, as it were, treated well from the beginning. So, the chatbot ends up giving me the feeling even before I start the meditation.
Q3: [00:15:22] Let's say I added some identity to the bot, so it's not just the bot that says "Hi, how can I help you?" He turned it into a bot that has a name, which I named him. And he kind of introduces himself and says he's available there and tries to help the user and also when the chat is over he says he's available for when he needs it again. So, if he communicates in a more human way, who is he sending messages to, who is he talking to. I think that was the main point.
P4: [00:09:15] I went through the topics, then I kind of saw: "ah this here, I tried to do it, this here I tried to do it". [...] But otherwise I think it turned out pretty similar, trying to bring more naturalness, those things.
P5: [00:17:20] I think that, like, after I did both conversations, I would feel a better conversation experience in conversation two. For sure.
P6: [00:14:36] It's still not ideal, but I managed to make the conversation more natural, more fluid, in a way that the person doesn't feel like they were machine, and that they can connect even knowing that they are talking to the machine and could relate in a simpler, more fluid way.
P7: [00:11:14] Especially when a user asks me in requirement B, when you suggest a type of meditation for a specific type of mood. I think it makes sense from a chatbot perspective, you comfort the user in negative emotions and you kind of kind of coddle the user in positive moods, like encourage them maybe.
P8: [00:08:53] I think he would be more comfortable in [conversation] 2.
P9: [00:12:43] I would follow. I think guides are useful, right? We always have to emphasize this, because for people who develop, many times we are dependent on things that are inside our heads. So we don't have a guide to follow. Well, this is something that I find very complicated, very difficult, and I see that it would be useful, I would go on without any problems.

Figure 6.4: Thematic Map 4 (TM4): Participants recognized the importance of the guide for better user-chatbot interaction and ensured that their conversations complied with the guide.



Transcripts excerpts supporting theme (not extensive)

P1: [00:10:10] Look, I didn't see any point initially that any change would be necessary. I think the guide is well summarized, it is well exemplified, it has a lot of examples. I, at least at first, think that if I had a slightly more complex chatbot development experience maybe I could give you an idea of something to improve. But at first, I see the guide as a very solid foundation for you to develop a good chatbot. The strong point would be exactly that, being something summarized, not too extensive, having examples and well focused on instructions for creating the chatbot.

P2: [00:20:22] Look, overall I really liked the guide. I think it's well organized and short, which is an important thing. Maybe I would change the introduction a little bit, because I don't really know if that was the intention, like, make a more academic introduction.

P3: [00:21:17] Well, the strengths, I think those examples you have for each of the points, I think it helps a lot. So if possible even have two examples. I don't know, instead of just one, it can also help a lot until the person can think better how to improve their creativity a little, you know? Have a better starting point. Of the weaknesses, I think that everything is always susceptible to be improved, so I don't think anything is perfect. It is possible that over time new ideas will emerge that may be being added there to better define how to make a chatbot a little more human, so to speak. But other than that I thought it was pretty cool.

P4: [00:16:03] I would use it. I don't think I would stop using it, because it's very practical and these are things that you can see what you need with little text. So I would use, yes.

P5: [00:20:27] So the guide for me is all strong points. I thought it was well structured. I already think he addresses all the necessary points. I don't know everything about a chatbot, so I wouldn't be able to say exactly if any point was left out. But for the experience I had, all points were well addressed and well exemplified. So, yes, his strong point would be that, a very didactic presentation, with examples, although short examples, but very punctual and well directed towards the explanation, right there on the side. I thought it was very good, very didactic. [...] That experience I have, right, which is little, but I thought it's already a very good starting point for those who have never worked with this, if they need to do work in this sense.

P6: [00:17:32] I think the strong point is that the way it is structured, especially the three characteristics, I think they are very clear and we can understand very well how to apply them. The only thing that left me a little confused was the conceptual map. I don't know if it was the reading I did at the beginning, because after I read the rest of the content I was able to understand it better. But maybe if you could, I don't know, putting it after the three characteristics would help a little bit more.

P7: [00:05:39] I think it's pretty intuitive. I was a little apprehensive and I don't know how to solve it, but it was something that I felt as a user, that it would be a very extensive guide. When actually not, it is very lean, it is very nice. But the landing page, the first idea I had when I saw it, like several topics, several things, I said "man, I'm going to spend half an hour here, calmly". So, maybe the homepage is not as inviting as the rest of the guide.

P10: [00:12:30] Really, there is no case that I wouldn't use, because as I don't have much experience, it was something that made me understand very well why this is important in the chatbot. And if I, as a user, were using the application of someone who used it, maybe I would also like what I would receive.

Figure 6.5: Thematic Map 5 (TM5): Participants saw the guide as a great reference for all stakeholders to determine conversational requirements, especially for beginners.

human-like since the first conversation. Their interviews also reveal a greater concern about how they should make users feel with their conversations. This finding is in line with the work of Clemmensen et al. [99], in which it was found that usability professionals worry more about user-related constructs, subjective UX, and emotion-related aspects of system use than developers and users. This reflects the discrepancies between design and implementation, which cause breakdowns between designers and developers in software development [100].

Many participants were very concerned with how their conversation would be technically implemented, which affected their design decisions. Looking through the conversations delivered in Stage 1, it is possible to notice a strong use of guided interactions, in which the chatbot gives a format the user should use to respond or numbered options. Although participants do not have experience as chatbot developers, they understand that a free-text interaction requires more complex algorithms than a menu-based approach. This is due not only to their concerns about ease of implementation, but also their emphasis on objectivity, as mentioned previously.

Participants' initial preference for menu-based interactions and objectivity raise concerns about the proposed guide because these interactions are usually not human-like, which goes against the proposed practices. Interestingly, this relates to a part of the interview of P7: "[...] it has its uses, when well applied, they are very well applied, but when not, you look at that there and say: man, I don't know if this was needed", meaning that chatbots are not a solution for all cases and sometimes a form is a better fit. Although participants unanimously confirmed that the guide's practices are an important asset for conversational design, as seen in other works, users crave more humanness in chatbot interactions [101], its effectiveness depends on the correct choice of using a chatbot in a given context.

The practices also depend on the designer's consciousness about the chatbot domain, context, and audience. According to these external factors, the guide was conceived as a menu of practices that should be used wisely. As envisioned, participants relied heavily on their previous experiences and knowledge to select the best options for this meditation chatbot. For example, P2 noted that small talk was not a wise decision, given that the conversation sample was short. Therefore, she understood that the addition of this practice would not contribute to the conversation in this case. The use of the designer's personal experience was noticed by other works as well [102, 103], and it is recognized as a valid and important design strategy.

Looking at the conversations delivered after reading the guide, the most used practice was self-presentation. We infer that this practice was used a lot because it would be easy to implement, it is very much domain-independent, and it is unrelated to personality traits.

Since it usually happens on the first message, it does not depend on conversation context, database query or natural language processing. Moreover, it is not a controversial practice due to personal preferences, as we have seen P2 and P5 reported not enjoying humor and small talk so much, respectively. Still, it is very alarming that such a basic practice was not used by many of them in the first conversation since it is the one responsible for giving an overview of the chatbot to the user and essential for shaping users' satisfaction and intent to engage based on how their expectations were met [16]. Therefore, it reinforces the importance of the guide.

Generally, participants reported positive feelings toward the guide, stating their interest in using it if they were to develop a chatbot in the future. The most praised point was the presentation of the content, which was seen as easy to absorb, specially with the help of the guide's use case, something mentioned by many participants. Looking broadly, the guide can be considered as a type of software documentation and, in this context, it is known that regarding readability, documentation clarity is the issue perceived as most important by practitioners, and it should be tested by someone with little domain knowledge [104]. As for support for newcomers in chatbot development, after studying Stack Overflow posts about chatbot development, Abdellatif et al. [105] concluded that there is a lack of proper chatbot introductory documentation. The main findings that validate that the guide fulfilled its purposes are:

1. Only 2 of 10 participants opted not to change their conversations due to perceived compliance with the guide, and the ones who did changes were based on their perception that their conversations could be improved, which indicates that the guide successfully instructed them in how to design more natural and pleasant conversations;

2. When directly asked, all of them affirmed that they would use the guide in chatbot development, showing that the guide is seen as useful from the perspective of developers with little experience to those with extensive experience, accommodating the concerns of practitioners that pay more attention to the quality of chatbot UX while also awakening the need to prioritize user needs in developers who tend to prioritize ease of development;

3. Their conversations delivered in stage 2 clearly show that the participants understood and were able to apply the practices. However, the application decisions were influenced by several factors, in a positive and foreseeable way, given that the guide was conceived to be used flexibly and respecting the decisions that designers may make depending on the domain, their experiences, and external requirements.

6.7 Guide Improvements

As done with the survey, the case study results were used to improve the guide to serve chatbot development better. Remarkably, it was seen that the participants very well received the changes made after the survey. In highlight, the use case was the change most mentioned in the interviews, always in a very positive way and proving that it is valuable for understanding how to apply the guide. Less notoriously, some participants also unpromptedly mentioned the importance of the list of bibliographic references for them to place more trust in the guide.

The interviews allowed participants to freely express their opinions, which encouraged them to suggest in detail how the guide could be changed to improve its usability and understandability. Table 6.4 presents the suggestions extracted from the interviews which had a specific intervention idea for the guide and also presents how these suggestions were addressed. The new pages generated after changes can be seen in our Zenodo repository with the complete third version of the guide [32], which is the final version of the guide in the scope of this work.

6.8 Chapter Summary

This chapter detailed the methodology and results of GCCD's case study validation. We invited software development practitioners to participate in an experiment in which they were asked to develop conversation samples for a fictitious meditation chatbot. The first conversation was done without the guide and later they were asked to read the guide and optionally change their conversations if they saw an opportunity for improvement. Finally, interviews were conducted with each participant and the transcripts were examined with objective, narrative, and thematic analysis. Participants were unanimous in approving the use of the guide and it was seen that the guide has a flexibility capable of meeting the different approaches of each participant according to their knowledge and personal experiences. In consonance with the survey, the guide's strengths are related to content presentation and it was seen that technical details for implementation are still a limitation since developers had problems using practices while also thinking about how it should be implemented. In addition, participants contributed with some suggestions that were incorporated into the final version of the proposed guide.

Suggestion	Measure
P2: [00:20:22] Maybe I would change the introduction a little bit, because, actually, I don't know if that was the intention, like, make a more academic introduction. And then the part of the map, the naturalness, etc., make it really close to the chatbot, because the chatbot language is really much more informal than the language of a guide. Anyway, I don't know if it was the intention, but maybe I would leave the language of the guide all uniform. As the chatbot already has a more informal language, I would leave the language at the beginning also more informal. Because when you open the page, you are already faced with a text. P7: [00:05:39] P7: [00:05:39] I was a little apprehensive and I don't know how to solve it, but it was something that I felt as a user, that it would be a very extensive guide. When actually not, it is very lean, it is very nice. But the landing page, the first idea I had when I saw it, like several topics, several things, I said "man, I'm going to spend half an hour here, calmly". So, maybe the homepage is not as inviting as the rest of the guide. P8: [00:10:48] Yes, this, the difference between the types of chatbots, to say that it is only for this type of chatbot that it is the most suitable. That's basically it.	The introduction and the conceptual map pages were rewritten to become shorter and less academic. Moreover, the new text reinforces the type of chatbots the guide is aimed at.
P4: [00:18:12] When you're on the homepage, you have the conversational design and you start reading, then there's this "read here before proceeding to the pages". And since it's in bold, my eye went straight to it and I kept trying to click on "here" and then I understood [it was not a link].	Since the introduction was rewritten, this exact text no longer exists. The term "here" was not used to avoid such misinterpretations.
P6: [00:17:32] The only thing that left me a little confused was the conceptual map. I don't know if it was the reading I did at the beginning, because after I read the rest of the content I was able to understand it better. But maybe if you could, I don't know, putting it after the three characteristics would help a little bit more. P10: [00:13:29] On the conceptual map issue, maybe I got a little lost in the first column of "chatbots with a purpose". Maybe I don't know if I understood how to get to the second stage, but that's because I don't have much experience either, I think. P10: [00:14:47] [About putting the conceptual map after the practices] It could be a little more understandable, yes, although in the end, reaching the three, the three main points are very clear on how to use it, how it is done, why. But before that, maybe it is a little more confusing.	The conceptual map was replaced after the pages about the practices, as suggested by participants. The text was rewritten to become shorter, objective and focused on explaining the columns of the map.
P8: [00:10:48] And since you're doing it with a web page, like putting in some links. Today I was looking at it on my cell phone and when I get to the end of a page like that, for example, there could be a link to the next one. For example, there is naturalness, which already has a link to emotionality. P8: [00:11:30] And a little arrow to go back, one to go, those things like that, to leave more... To navigate easier.	Navigation links were added to the end of each page, establishing a navigation in the order in which the content should be read.

Table 6.4: Suggestions for improvement made during the interviews and the measures taken to address them.

Chapter 7

Discussion

This chapter discusses the results of the execution of this work such as implications, contribution, and threats to validity.

7.1 Theoretical Contribution

As seen in Section 2.3, the interaction of conversational agents has been an object of study in several papers, but with different methodologies and focus. After presenting the results of this article, it is possible to establish how it contributes to what has been presented by current literature. Table 7.1 shows the conversational design practices mentioned in similar works. However, in some of them, the practices are not explicitly mentioned or are part of a broader recommendation.

Our SLR results produced a ready-to-use guide that summarizes the results found in a way accessible to professionals in the field, adding a step further to pure SLR studies [24, 64, 25]. Moreover, these SLR studies differ from this work mainly because they focus more on the social characteristics of the chatbots and present broader discussions rather than straightforward guidelines. On the other hand, in our SLR, we are concerned with finding software requirements ready to be implemented.

The works that derived guidelines from user studies had a final result closer to our proposed guide [65, 26, 66]. However, their list is different regarding the practices presented because of their sample's different focus or limitations. The SLR approach enabled us to take advantage of many user studies conducted with a diverse sample, making our conversational design practices broader than those elicited from one user study. Moreover, since we considered the user impacts reported in the literature when deriving the proposed design practices, our study resulted in a guide that is guaranteed to propose practices that awaken positive feelings in users, which would not be guaranteed with a simple assembling of practices from related works.

Table 7.1: Conversational design practices that were recommended by related works.

Naturalness	[24]	[64]	[25]	[65]	[26]	[66]	[67]	[68]	[69]	[70]
Self-introduction		✓	✓			✓		✓		✓
Address user by name		✓								
Small talk	✓	✓	✓			✓		✓		
Echoing responses			✓						✓	
Casual language	✓	✓	✓	✓	✓					✓
Emotionality	**[24]**	**[64]**	**[25]**	**[65]**	**[26]**	**[66]**	**[67]**	**[68]**	**[69]**	**[70]**
Exclamatory feedback			✓							
Graphical media	✓		✓							
Empathic messages	✓		✓				✓			
Humor	✓		✓			✓		✓		✓
Transparency	**[24]**	**[64]**	**[25]**	**[65]**	**[26]**	**[66]**	**[67]**	**[68]**	**[69]**	**[70]**
Present capabilites	✓			✓	✓			✓		✓
Acknowledge limitations										✓
Make suggestions	✓	✓	✓	✓				✓		✓
Ask for clarification		✓	✓				✓	✓		
What to avoid	**[24]**	**[64]**	**[25]**	**[65]**	**[26]**	**[66]**	**[67]**	**[68]**	**[69]**	**[70]**
Repetitive messages										
Exaggeration										✓
Hiding true identity	✓	✓	✓							✓
Machine-like typeface			✓			✓				
Forcing errors		✓	✓					✓		

7.2 Future Concerns

Social media platforms only allowed developers to create chatbots in 2016 [5]. Up until then, chatbots were only being developed and presented to the general public by big companies, such as Google and Apple. The oldest paper selected for the SLR is from 2017, which aligns with the public release of chatbot development platforms for developers. Therefore, we can infer that the advent of these tools changed how chatbots are developed and behave, consequently impacting aspects of user experience, which is the focus of our research. In this sense, if chatbot development suffers from a significant change, it can impact the user experience in the future as well as the applicability of our guide.

Since natural language interaction is a field that is constantly evolving, it may evolve to the point that text-based interactions become outdated, and voice-based interactions become the main form of interaction since it is more natural and practical in general. Although this is an assumption, this could make our guide less influential for the field and reinforce the need to repeat the study for voice-based chatbots or see if the guide's design practices are as positive for voice-based interactions as they are for text-based interactions.

Still in the same line of reasoning, if natural language interactions evolve enough to become indistinguishable from human interactions, it is necessary to review some design practices, such as revealing the chatbot's identity. However, this will raise ethical concerns regarding deceiving humans about whom they are talking to as well as the limitations of how far the humanness of these chatbots can go. Besides ethical concerns, data protection laws can also impact some design practices, such as collecting the user name. Therefore, ethical and privacy principles can be one of the future directions for evolving this guide.

7.3 Threats to Validity and Limitations

This work suffers from some common threats in SLRs [106]. The first threat is the use of an automatic search only, which can result in missing primary studies. Moreover, the limited number of authors may introduce a bias in the selection of studies since there are only two researchers to discuss and reach a consensus on the inclusion or exclusion of a paper. Lastly, since the first step was excluding papers by abstract, relevant papers could be excluded due to poorly written or incomplete abstracts that do not adequately convey the work that has been done.

The survey validation of the guide suffers from threats seen in surveys, such as the sincerity of responses and respondents' commitment to reading the whole guide carefully since their reading was unsupervised. On the other hand, the occurrence of ill-considered

answers was reduced since participation was voluntary, and no incentives or gains were linked to the survey response. Furthermore, we added mandatory questions that made respondents confirm that they had followed the instructions before moving on to the next stages to mitigate these effects.

Furthermore, there are four other main threats to the validity of the case study: interviewer effects, courtesy bias, limitation of study settings, and analysis bias. The threat of interviewer effects is about the possibility of inducing preferred answers by the writing of interview questions or by the interviewer giving unconscious clues. This threat was mitigated by carefully producing questions that asked for broad perceptions instead of using pre-defined options for the interviewee to choose from. Moreover, the interview was conducted remotely with cameras off to avoid unconscious clues from the interviewer's behavior or expressions.

The courtesy bias is about the feeling participants may have in pleasing and being courteous with the interviewer, leading to a lack of sincerity when answering questions. This threat was mitigated by directly asking if they had any negative perceptions and reinforcing the need for honest and sincere participation in the instructions document. Moreover, we had cameras off during the remote interview to make participants more comfortable and avoid them feeling confronted at any moment.

The limitation of study settings concerns the simulated environment and chatbot requirements that our case study provides, which is only a small sample of what developers may face in a real chatbot development scenario. Moreover, since this is a qualitative study executed by the two authors, this work is not free from bias in the analysis of interview transcripts. However, both the case study and the interview analysis followed a systematic protocol, supported by open data [32], in the hope of mitigating these threats. Still, this should be considered when interpreting this study's results.

Concerning the limitations and coverage of this work, we only cover conversational design practices for text-based chatbots, which may be applied or adapted to speech interfaces. Nevertheless, the impacts may be different from those presented here. The coverage of the impacts is also limited because we only considered positive outcomes in our search string since including negative keywords would make the string too big. Moreover, the guide only summarizes the results of the selected papers, and there may be other conversational design practices and strategies that positively impact users for each purpose that were not listed. Lastly, the guide only covers recommendations for establishing chatbot requirements, but it does not address technical implementation and viability of practices.

7.4 Chapter Summary

The customized protocol of the SLR led us to a broader set of conversational design practices than the set of related works. Moreover, we were able not only to assemble a set of these practices but also their impacts on users. However, our guide is subject to rapid changes in the field of virtual assistants, and how users prefer to interact with such assistants can impact the applicability of GCCD. Still, the guide can evolve to meet these new interaction trends since the methodology is replicable. Furthermore, the guide is not free from threats of studies based on SLRs, surveys, and case studies, but our multi-method approach helps mitigate these threats and reinforces the credibility of the overall results.

Chapter 8

Conclusion

In this work, we proposed guidelines for text-based chatbot conversational design, considering the impacts caused on users by using some conversational practices. These guidelines culminated in the creation of the guide Guidelines for Chatbot Conversational Design (GCCD). This guide was built upon the analyses of the results of an SLR conducted for this purpose and was validated through a survey.

The SLR returned a total of 1101 papers, but after applying the protocol, we selected 40 papers from different contexts and with various practices being tested with users to evaluate how they feel about the presence or absence of these practices. The joint analysis of selected papers revealed some patterns in chatbot design that were attached to the chatbot purpose. For each purpose, papers generally focused on a group of impacts and tested practices to enhance positive impacts. These patterns were added to our conceptual map, the starting point for creating the guide. The map establishes that these relationships can be enforced through some conversational practices, which were grouped into three objectives: naturalness, emotionality, and transparency. Moreover, it presents some practices that should be avoided.

The guide was developed as a web page that exposes, explains, and exemplifies each conversational practice with an accessible language and presentation. It was validated with a survey and a case study. The survey was shared with technology practitioners to gather their opinions about it and assess the guide's quality through a TAM questionnaire, which revealed satisfactory scores for the guide's usability and understandability. Moreover, developers' comments have shown that the guide's main strengths are objectivity and clarity. The case study consisted of participants proposing conversation samples first without the guide and later having the opportunity to improve the conversation after being presented to the guide.

In the case study, we invited practitioners to participate in an experiment in which they were asked to develop conversation samples for a fictitious meditation chatbot. The first

conversation was made without the GCCD guide, and later they were asked to read the guide and optionally change their conversations if they saw an opportunity for improvement. Finally, interviews were conducted with each participant, and the transcripts were examined with objective and thematic analysis. The case study confirmed the usefulness of the guide given that participants' reported intention to use it and the guide's objectivity and clarity according to their extensive perceptions given through the conducted interviews. It also provided evidence that the guide is very flexible to be used according to the designer's own convictions since most participants had negative experiences with chatbots, which greatly influenced how they viewed some practices. Furthermore, their experiences in specific software development positions also influenced their design and adoption of practices. Still, implementation aspects are a limitation for the guide's use, which was evidenced in both validations.

The results achieved are promising and show that the guide is helpful for practitioners with different levels of experience, and it is generic enough for use in various domains and by professionals with different backgrounds. However, it is important to notice that the guide is limited to text-based chatbots, and the analysis that provided the creation of the guide is based on selected studies with current trends in human-chatbot interaction, which is a rapid-evolving field. The most needed future work based on our findings would be to analyze the implementation strategies for the guide's practices in real chatbot frameworks, as well as see how practitioners cope with the guide in a real situation of chatbot development.

References

[1] Khan, Rashid and Anik Das: *Introduction to chatbots*. In *Build Better Chatbots: A Complete Guide to Getting Started with Chatbots*, pages 1–11. Apress, Berkeley, CA, 2018. vii, 1

[2] Xu, Yingzi, Chih Hui Shieh, Patrick van Esch, and I Ling Ling: *Ai customer service: Task complexity, problem-solving ability, and usage intention*. Australasian Marketing Journal (AMJ), 28(4):189–199, 2020, ISSN 1441-3582. https://www.sciencedirect.com/science/article/pii/S1441358220300240. vii, 1

[3] McTear, Michael F.: *Conversational AI: Dialogue Systems, Conversational Agents, and Chatbots*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, Toronto, 2020. https://doi.org/10.2200/S01060ED1V01Y202010HLT048. vii, 1

[4] Brandtzaeg, Petter Bae and Asbjørn Følstad: *Why people use chatbots*. In *International conference on internet science*, pages 377–392. Springer, 2017. vii, 1

[5] Adamopoulou, Eleni and Lefteris Moussiades: *Chatbots: History, technology, and applications*. Machine Learning with Applications, 2:100006, 2020, ISSN 2666-8270. https://www.sciencedirect.com/science/article/pii/S2666827020300062. xvii, 7, 8, 21, 84

[6] Soni, Neha, Enakshi Khular Sharma, Narotam Singh, and Amita Kapoor: *Impact of artificial intelligence on businesses: from research, innovation, market deployment to future shifts in business models*. arXiv preprint arXiv:1905.02092, 2019. 1

[7] Weizenbaum, Joseph: *ELIZA - a computer program for the study of natural language communication between man and machine*. Commun. ACM, 9(1):36–45, 1966. https://doi.org/10.1145/365153.365168. 1, 7

[8] Laumer, Sven, Fabian Tobias Gubler, A. A. Racheva, and Christian Maier: *Use cases for conversational agents: An interview-based study*. In *AMCIS*, 2019. 1

[9] Androutsopoulou, Aggeliki, Nikos Karacapilidis, Euripidis Loukis, and Yannis Charalabidis: *Transforming the communication between citizens and government through ai-guided chatbots*. Government information quarterly, 36(2):358–367, 2019. 1

[10] Petriv, Yulia, Regina Erlenheim, Valentyna Tsap, Ingrid Pappel, and Dirk Draheim: *Designing effective chatbot solutions for the public sector: A case study from ukraine*. In *International Conference on Electronic Governance and Open Society: Challenges in Eurasia*, pages 320–335. Springer, 2019. 1

[11] Lommatzsch, Andreas: *A next generation chatbot-framework for the public administration.* In *International Conference on Innovations for Community Services*, pages 127–141. Springer, 2018. 1

[12] Miner, Adam S, Liliana Laranjo, and A Baki Kocaballi: *Chatbots in the fight against the covid-19 pandemic.* NPJ digital medicine, 3(1):1–4, 2020. 1

[13] Zhang, Juliana JY, Asbjørn Følstad, and Cato A Bjørkli: *Organizational factors affecting successful implementation of chatbots for customer service.* Journal of Internet Commerce, pages 1–35, 2021. 2

[14] MarketsandMarkets: *Conversational ai market size, share and global market forecast to 2026*, Oct 2021. https://www.marketsandmarkets.com/Market-Reports/conversational-ai-market-49043506.html?gclid=CjOKCQjw0PWRBhDKARIsAPKHFGh6B5mjeTgniXBOCG6HkceFBDGMDxe0O9HbeqJEWG0scGsKD6UQd1MaAqwcB. 2

[15] Følstad, Asbjørn and Petter Bae Brandtzæg: *Chatbots and the new world of HCI.* Interactions, 24(4):38–42, 2017. https://doi.org/10.1145/3085558. 2

[16] Zamora, Jennifer: *I'm sorry, dave, i'm afraid I can't do that: Chatbot perception and expectations.* In Wrede, Britta, Yukie Nagai, Takanori Komatsu, Marc Hanheide, and Lorenzo Natale (editors): *Proceedings of the 5th International Conference on Human Agent Interaction, HAI 2017*, pages 253–260, Bielefeld, Germany, 2017. ACM. https://doi.org/10.1145/3125739.3125766. 2, 79

[17] Google: *What is conversation design?*, Feb 2021. https://developers.google.com/assistant/conversation-design/what-is-conversation-design. 2

[18] Liebrecht, Christine, Lena Sander, and Charlotte van Hooijdonk: *Too informal? how a chatbot's communication style affects brand attitude and quality of interaction.* In *International Workshop on Chatbot Research and Design*, pages 16–31. Springer, 2020. 2

[19] Følstad, Asbjørn, Cecilie Bertinussen Nordheim, and Cato Alexander Bjørkli: *What makes users trust a chatbot for customer service? an exploratory interview study.* In Bodrunova, Svetlana S. (editor): *Internet Science - 5th International Conference, INSCI 2018*, volume 11193 of *Lecture Notes in Computer Science*, pages 194–208, St. Petersburg, Russia, 2018. Springer. https://doi.org/10.1007/978-3-030-01437-7_16. 2

[20] Henman, Paul: *Improving public services using artificial intelligence: possibilities, pitfalls, governance.* Asia Pacific Journal of Public Administration, 42(4):209–221, 2020. 2

[21] Abdellatif, Ahmad, Diego Costa, Khaled Badran, Rabe Abdalkareem, and Emad Shihab: *Challenges in chatbot development: A study of stack overflow posts.* In Kim, Sunghun, Georgios Gousios, Sarah Nadi, and Joseph Hejderup (editors): *MSR '20: 17th International Conference on Mining Software Repositories, Seoul, Republic of*

Korea, 29-30 June, 2020, pages 174–185. ACM, 2020. https://doi.org/10.1145/3379597.3387472. 2

[22] Freed, Andrew: *Conversational AI: Chatbots that work*. Manning Publications, New York, NY, 2021. 2, 3

[23] Meharabian, Albert: *Silent messages: Implicit communication of emotions and attributes*. Wadsworth, California, 1980. 3

[24] Chaves, Ana Paula and Marco Aurelio Gerosa: *How should my chatbot interact? a survey on social characteristics in human–chatbot interaction design*. International Journal of Human–Computer Interaction, 37(8):729–758, 2021. https://doi.org/10.1080/10447318.2020.1841438. 3, 13, 82, 83

[25] Rapp, Amon, Lorenzo Curti, and Arianna Boldi: *The human side of human-chatbot interaction: A systematic literature review of ten years of research on text-based chatbots*. International Journal of Human-Computer Studies, 151:102630, 2021, ISSN 1071-5819. https://www.sciencedirect.com/science/article/pii/S1071581921000483. 3, 13, 21, 82, 83

[26] Yang, Xi and Marco Aurisicchio: *Designing conversational agents: A self-determination theory approach*. In Kitamura, Yoshifumi, Aaron Quigley, Katherine Isbister, Takeo Igarashi, Pernille Bjørn, and Steven Mark Drucker (editors): *CHI '21: CHI Conference on Human Factors in Computing Systems*, pages 256:1–256:16, Virtual Event / Yokohama, Japan, May 8-13, 2021, 2021. ACM. https://doi.org/10.1145/3411764.3445445. 3, 13, 82, 83

[27] Standardization, International Organization for: *Systems and software engineering — Vocabulary*. Standard, International Organization for Standardization, Geneva, CH, September 2017. 3

[28] Barbara, Kitchenham and Stuart Charters: *Guidelines for performing systematic literature reviews in software engineering*. Keele University, UK, 9:1–65, 2007. 5, 16, 18, 22

[29] Wolfswinkel, Joost F, Elfi Furtmueller, and Celeste P M Wilderom: *Using grounded theory as a method for rigorously reviewing literature*. European Journal of Information Systems, 22(1):45–55, 2013. https://doi.org/10.1057/ejis.2011.51. 5, 23, 34

[30] Fowler Jr, Floyd J: *Survey research methods*. Sage publications, 2013. 5

[31] Yin, Robert K: *Case study research: Design and methods*, volume 5. sage, 2009. 5

[32] Silva, Geovana Ramos Sousa and Edna Dias Canedo: *Supplementary Material for Human Factors in the Design of Chatbot Interactions: Conversational Design Practices*, January 2023. https://doi.org/10.5281/zenodo.7538681. 5, 23, 36, 45, 46, 47, 53, 55, 59, 69, 80, 85

[33] Shneiderman, Ben, Catherine Plaisant, Maxine Cohen, Steven Jacobs, and Niklas Elmqvist: *Designing the User Interface: Strategies for Effective Human-Computer Interaction*. Pearson, Boston, 6th edition, 2016, ISBN 978-0-13-438038-4. 7

[34] Cahn, Jack: *CHATBOT: Architecture, design, & development.* Senior thesis, University of Pennsylvania School of Engineering and Applied Science Department of Computer and Information Science, 2017. 7

[35] Adamopoulou, Eleni and Lefteris Moussiades: *An overview of chatbot technology.* In Maglogiannis, Ilias, Lazaros Iliadis, and Elias Pimenidis (editors): *Artificial Intelligence Applications and Innovations*, pages 373–383, Cham, 2020. Springer International Publishing, ISBN 978-3-030-49186-4. 7, 8, 9

[36] Colby, Kenneth Mark, Sylvia Weber, and Franklin Dennis Hilf: *Artificial paranoia.* Artificial Intelligence, 2(1):1–25, 1971. 7

[37] Carpenter, Rollo: *About the jabberwacky ai.* http://www.jabberwacky.com/j2about. 8

[38] Mauldin, Michael L: *Chatterbots, tinymuds, and the turing test: Entering the loebner prize competition.* In *AAAI*, volume 94, pages 16–21, 1994. 8

[39] Wallace, Richard S.: *The Anatomy of A.L.I.C.E.*, pages 181–210. Springer Netherlands, Dordrecht, 2009, ISBN 978-1-4020-6710-5. https://doi.org/10.1007/978-1-4020-6710-5_13. 8

[40] Marcondes, Francisco S., José João Almeida, and Paulo Novais: *Chatbot theory.* In Yin, Hujun, David Camacho, Paulo Novais, and Antonio J. Tallón-Ballesteros (editors): *Intelligent Data Engineering and Automated Learning – IDEAL 2018*, pages 374–384, Cham, 2018. Springer International Publishing. 9

[41] Yeung, Ken: *Facebook opens its messenger platform to chatbots.* VentureBeat. [Online]. Available at https://venturebeat.com/2016/04/12/facebook-opens-its-messenger-platform-to-chatbots/, Apr 2016. 9

[42] Andrade, Guilherme De Guy, Geovana Ramos Sousa Silva, Francisco Carlos Molina Duarte Júnior, Giovanni Almeida Santos, Fábio Lúcio Lopes de Mendonça, and Rafael Timóteo de Sousa Júnior: *Evatalk: A chatbot system for the brazilian government virtual school.* In Filipe, Joaquim, Michal Smialek, Alexander Brodsky, and Slimane Hammoudi (editors): *Proceedings of the 22nd International Conference on Enterprise Information Systems, ICEIS 2020*, pages 556–562, Prague, Czech Republic, 2020. SCITEPRESS. https://doi.org/10.5220/0009418605560562. 9, 12

[43] Singh, Abhishek, Karthik Ramasubramanian, and Shrey Shivam: *Introduction to Microsoft Bot, RASA, and Google Dialogflow*, chapter 7, pages 281–302. Apress, Berkeley, CA, 2019, ISBN 978-1-4842-5034-1. https://doi.org/10.1007/978-1-4842-5034-1_7. 9

[44] Fadhil, Ahmed: *Can a chatbot determine my diet?: Addressing challenges of chatbot application for meal recommendation.* arXiv preprint arXiv:1802.09100, 2018. 9

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

[46] Jakic, Ana, Maximilian Oskar Wagner, and Anton Meyer: *The impact of language style accommodation during social media interactions on brand trust.* Journal of Service Management, 2017. 10

[47] Følstad, Asbjørn and Petter Bae Brandtzæg: *Chatbots and the new world of hci.* interactions, 24(4):38–42, 2017. 10

[48] Følstad, Asbjørn and Petter Bae Brandtzaeg: *Users' experiences with chatbots: findings from a questionnaire study.* Quality and User Experience, 5(1):1–14, 2020. 10

[49] Neff, Gina and Peter Nagy: *Automation, algorithms, and politics| talking to bots: Symbiotic agency and the case of tay.* International Journal of Communication, 10:17, 2016. 10

[50] Harwell, Drew: *The accent gap.* The Washington Post, July 2018. https://www.washingtonpost.com/graphics/2018/business/alexa-does-not-understand-your-accent/, visited on 2021-09-09. 10

[51] Google: *Conversation design*, February 2021. https://developers.google.com/assistant/conversation-desig. 10

[52] IBM: *Conversational ux design*, 2021. https://conversational-ux.mybluemix.net/design/conversational-ux/. 10

[53] Amazon: *Designing for conversation*, 2021. https://developer.amazon.com/ask-resources/guided/conversational-design-workshop/. 10

[54] Lei, S.I., H. Shen, and S. Ye: *A comparison between chatbot and human service: customer perception and reuse intention.* International Journal of Contemporary Hospitality Management, 33(11):3977–3995, 2021. 10, 29

[55] Paikari, Elahe and André van der Hoek: *A framework for understanding chatbots and their future.* In Sharp, Helen, Cleidson R. B. de Souza, Daniel Graziotin, Meira Levy, and David Socha (editors): *Proceedings of the 11th International Workshop on Cooperative and Human Aspects of Software Engineering, ICSE 2018*, pages 13–16, Gothenburg, Sweden, 2018. ACM. https://doi.org/10.1145/3195836.3195859. 10

[56] Hill, Jennifer, W. Randolph Ford, and Ingrid G. Farreras: *Real conversations with artificial intelligence: A comparison between human–human online conversations and human–chatbot conversations.* Computers in Human Behavior, 49:245–250, 2015, ISSN 0747-5632. https://www.sciencedirect.com/science/article/pii/S0747563215001247. 10

[57] Moore, Robert J. and Raphael Arar: *Conversational UX Design: A Practitioner's Guide to the Natural Conversation Framework.* Association for Computing Machinery, New York, NY, USA, 2019, ISBN 9781450363013. 10

[58] McTear, Michael: *Conversation modelling for chatbots: current approaches and future directions*. In Berton, André, Udo Haiber, and Wolfgang Minker (editors): *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2018 [Conference on Electronic Speech Signal Processing]*, pages 175–185, Ulm, Germany, 2018. TUDpress, Dresden, ISBN 978-3-959081-28-3. 10

[59] Hobert, Sebastian: *How are you, chatbot? evaluating chatbots in educational settings – results of a literature review*. In Pinkwart, Niels and Johannes Konert (editors): *DELFI 2019*, pages 259–270, Bonn, 2019. Gesellschaft für Informatik e.V. 11

[60] Pérez-Soler, Sara, Esther Guerra, and Juan de Lara: *Model-driven chatbot development*. In Dobbie, Gillian, Ulrich Frank, Gerti Kappel, Stephen W. Liddle, and Heinrich C. Mayr (editors): *Conceptual Modeling*, pages 207–222, Cham, 2020. Springer International Publishing, ISBN 978-3-030-62522-1. 11

[61] Qian, Hongjin, Xiaohe Li, Hanxun Zhong, Yu Guo, Yueyuan Ma, Yutao Zhu, Zhanliang Liu, Zhicheng Dou, and Ji Rong Wen: *Pchatbot: A large-scale dataset for personalized chatbot*. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 2470–2477, New York, NY, USA, 2021. Association for Computing Machinery, ISBN 9781450380379. https://doi.org/10.1145/3404835.3463239. 12

[62] Choi, Yoonseo, Toni Jan Keith Palma Monserrat, Jeongeon Park, Hyungyu Shin, Nyoungwoo Lee, and Juho Kim: *Protochat: Supporting the conversation design process with crowd feedback*. Proc. ACM Hum.-Comput. Interact., 4(CSCW3), jan 2021. https://doi.org/10.1145/3432924. 12

[63] Pichiliani, Mauro, Heloisa Candello, Claudio Pinhanez, Julio Nogima, Sara Vidon, Melina Guerra, and Maira de Bayser: *A user interface with taxonomy features for content curators of chatbots*. In Stephanidis, Constantine, Margherita Antona, and Stavroula Ntoa (editors): *HCI International 2022 Posters*, pages 439–446, Cham, 2022. Springer International Publishing, ISBN 978-3-031-06417-3. 12

[64] Sugisaki, Kyoko and Andreas Bleiker: *Usability guidelines and evaluation criteria for conversational user interfaces: A heuristic and linguistic approach*. In *Proceedings of the Conference on Mensch Und Computer*, MuC '20, page 309–319, New York, NY, USA, 2020. Association for Computing Machinery, ISBN 9781450375405. https://doi.org/10.1145/3404983.3405505. 13, 82, 83

[65] Amershi, Saleema, Daniel S. Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi T. Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz: *Guidelines for human-ai interaction*. In Brewster, Stephen A., Geraldine Fitzpatrick, Anna L. Cox, and Vassilis Kostakos (editors): *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI 2019*, page 3, Glasgow, Scotland, UK, 2019. ACM. https://doi.org/10.1145/3290605.3300233. 13, 82, 83

[66] Feine, Jasper, Ulrich Gnewuch, Stefan Morana, and Alexander Maedche: *A taxonomy of social cues for conversational agents*. International Journal of Human-Computer

Studies, 132:138–161, 2019, ISSN 1071-5819. https://www.sciencedirect.com/science/article/pii/S1071581918305238. 14, 82, 83

[67] Guo, Yunsan, Jian Wang, Runfan Wu, Zeyu Li, and Lingyun Sun: *Designing for trust: a set of design principles to increase trust in chatbot.* CCF Transactions on Pervasive Computing and Interaction, pages 1–8, 2022. 14, 83

[68] Mafra, Malu, Kennedy Nunes, Adailton Castro, Adriana Lopes, Ana Carolina Oran, Geraldo Braz Junior, João Almeida, Anselmo Paiva, Aristofanes Silva, Simara Rocha, *et al.*: *Defining requirements for the development of useful and usable chatbots: An analysis of quality attributes from academy and industry.* In *International Conference on Human-Computer Interaction*, pages 479–493. Springer, 2022. 14, 83

[69] Komatani, Kazunori, Ryu Takeda, Keisuke Nakashima, and Mikio Nakano: *Design guidelines for developing systems for dialogue system competitions.* In *Conversational AI for Natural Human-Centric Interaction*, pages 161–177. Springer, 2022. 14, 83

[70] Stanley, Jeff, Ronna ten Brink, Alexandra Valiton, Trevor Bostic, and Becca Scollan: *Chatbot accessibility guidance: A review and way forward.* In *Proceedings of Sixth International Congress on Information and Communication Technology*, pages 919–942. Springer, 2022. 14, 83

[71] Kitchenham, Barbara: *Procedures for performing systematic reviews.* Keele, UK, Keele University, 33(2004):1–26, 2004. 16

[72] Wohlin, Claes, Per Runeson, Martin Höst, Magnus C. Ohlsson, Björn Regnell, and Anders Wesslén: *Systematic Literature Reviews*, pages 45–54. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012, ISBN 978-3-642-29044-2. https://doi.org/10.1007/978-3-642-29044-2_4. 17

[73] Brereton, Pearl, Barbara A. Kitchenham, David Budgen, Mark Turner, and Mohamed Khalil: *Lessons from applying the systematic literature review process within the software engineering domain.* J. Syst. Softw., 80(4):571–583, 2007. https://doi.org/10.1016/j.jss.2006.07.009. 18

[74] Costal, Dolors, Carles Farré, Xavier Franch, and Carme Quer: *Inclusion and exclusion criteria in software engineering tertiary studies: A systematic mapping and emerging framework.* In Lanubile, Filippo, Marcos Kalinowski, and Maria Teresa Baldassarre (editors): *ESEM '21: ACM / IEEE International Symposium on Empirical Software Engineering and Measurement*, pages 30:1–30:6, Bari, Italy, 2021. ACM. https://doi.org/10.1145/3475716.3484190. 18, 20

[75] Mitchell, Elliot and Lena Mamykina: *From the curtain to kansas: Conducting wizard-of-oz studies in the wild.* In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–6, 2021. 20

[76] Budiu, Raluca and Kate Moran: *How many participants for quantitative usability studies: A summary of sample-size recommendations*, Jul 2021. https://www.nngroup.com/articles/summary-quant-sample-sizes/. 22

[77] Sauro, Jeff and James R. Lewis: *Chapter 6 - what sample sizes do we need? part 1: summative studies*. In Sauro, Jeff and James R. Lewis (editors): *Quantifying the User Experience (Second Edition)*, pages 103–141. Morgan Kaufmann, Boston, second edition edition, 2016, ISBN 978-0-12-802308-2. https://www.sciencedirect.com/science/article/pii/B9780128023082000060. 22

[78] Grimes, G. Mark, Ryan M. Schuetzler, and Justin Scott Giboney: *Mental models and expectation violations in conversational ai interactions*. Decision Support Systems, 144:113515, 2021, ISSN 0167-9236. https://www.sciencedirect.com/science/article/pii/S0167923621000257. 29

[79] Davis, Fred D: *Perceived usefulness, perceived ease of use, and user acceptance of information technology*. MIS quarterly, pages 319–340, 1989. 47

[80] Marangunić, Nikola and Andrina Granić: *Technology acceptance model: a literature review from 1986 to 2013*. Universal access in the information society, 14(1):81–95, 2015. 47

[81] Silveira, Sofia A.M., Luciana A.M. Zaina, Leobino N. Sampaio, and Fábio L. Verdi: *On the evaluation of usability design guidelines for improving network monitoring tools interfaces*. Journal of Systems and Software, 187:111223, 2022, ISSN 0164-1212. https://www.sciencedirect.com/science/article/pii/S016412122200005X. 47, 50

[82] Diwakar, Anita S. and Santosh Noronha: *Usability and usefulness of advice tool experiment design guidelines for virtual laboratories*. In *2018 IEEE Tenth International Conference on Technology for Education (T4E)*, pages 146–149, 2018. 47

[83] Schefer, Ricardo Pezzotti, Matheus Sousa Bezerra, and Luciana A Martinez Zaina: *Supporting the development of social networking mobile apps for deaf users: Guidelines based on user experience issues*. In *Proceedings of the 8th International Conference on Software Development and Technologies for Enhancing Accessibility and Fighting Info-Exclusion*, pages 278–285, 2018. 47

[84] Zaina, Luciana A.M., Renata P.M. Fortes, Vitor Casadei, Leornardo Seiji Nozaki, and Débora Maria Barroso Paiva: *Preventing accessibility barriers: Guidelines for using user interface design patterns in mobile applications*. Journal of Systems and Software, 186:111213, 2022, ISSN 0164-1212. https://www.sciencedirect.com/science/article/pii/S0164121221002831. 47

[85] Parizi, Rafael, Marina Moreira, Igor Couto, Sabrina Marczak, and Tayana Conte: *A tool proposal for recommending design thinking techniques in software development*. Journal of Software Engineering Research and Development, 10:3:1 – 3:15, Mar. 2022. https://sol.sbc.org.br/journals/index.php/jserd/article/view/1931. 47

[86] Steinmacher, Igor, Tayana Uchoa Conte, Christoph Treude, and Marco Aurélio Gerosa: *Overcoming open source project entry barriers with a portal for newcomers*. In *Proceedings of the 38th International Conference on Software Engineering*, ICSE '16, page 273–284, New York, NY, USA, 2016. Association for Computing Machinery, ISBN 9781450339001. https://doi.org/10.1145/2884781.2884806. 47

[87] Senarath, Awanthika, Marthie Grobler, and Nalin Asanka Gamagedara Arachchilage: *Will they use it or not? investigating software developers' intention to follow privacy engineering methodologies.* ACM Trans. Priv. Secur., 22(4), nov 2019, ISSN 2471-2566. https://doi.org/10.1145/3364224. 47

[88] Wilder, J Welles: *New Concepts in Technical Trading Systems.* Trend Research, 1978. 50

[89] Vasavada, Navendu: *Fisher's test for exact count data*, 2016. https://astatsa.com/FisherTest/. 50

[90] Canedo, Edna Dias, Fabiana Freitas Mendes, Anderson Jefferson Cerqueira, Márcio Vinicius Okimoto, Gustavo Pinto, and Rodrigo Bonifácio: *Breaking one barrier at a time: how women developers cope in a men-dominated industry.* In Vasconcellos, Cristiano D., Karina Girardi Roggia, Vanessa Collere, and Paulo Bousfield (editors): *35th Brazilian Symposium on Software Engineering, SBES 2021, Joinville, Santa Catarina, Brazil, 27 September 2021 - 1 October 2021*, pages 378–387. ACM, 2021. https://doi.org/10.1145/3474624.3474638. 59

[91] Canedo, Edna Dias, Rodrigo Bonifácio, Márcio Vinicius Okimoto, Alexander Serebrenik, Gustavo Pinto, and Eduardo Monteiro: *Work practices and perceptions from women core developers in OSS communities.* In Baldassarre, Maria Teresa, Filippo Lanubile, Marcos Kalinowski, and Federica Sarro (editors): *ESEM '20: ACM / IEEE International Symposium on Empirical Software Engineering and Measurement, Bari, Italy, October 5-7, 2020*, pages 26:1–26:11. ACM, 2020. https://doi.org/10.1145/3382494.3410682. 59

[92] Ollerenshaw, Jo Anne and John W. Creswell: *Narrative research: A comparison of two restorying data analysis approaches.* Qualitative Inquiry, 8(3):329–347, 2002. https://doi.org/10.1177/10778004008003008. 60

[93] Clandinin, D Jean and F Michael Connelly: *Narrative inquiry: Experience and story in qualitative research.* John Wiley & Sons, 2004. 60

[94] Kiger, Michelle E and Lara Varpio: *Thematic analysis of qualitative data: Amee guide no. 131.* Medical teacher, 42(8):846–854, 2020. 60, 69, 70

[95] Nasheeda, Aishath, Haslinda Binti Abdullah, Steven Eric Krauss, and Nobaya Binti Ahmed: *Transforming transcripts into stories: A multimethod approach to narrative analysis.* International Journal of Qualitative Methods, 18:1609406919856797, 2019. https://doi.org/10.1177/1609406919856797. 61

[96] Følstad, Asbjørn and Marita Skjuve: *Chatbots for customer service: user experience and motivation.* In *Proceedings of the 1st international conference on conversational user interfaces*, pages 1–9, 2019. 72

[97] Brachten, Florian, Tobias Kissmer, and Stefan Stieglitz: *The acceptance of chatbots in an enterprise context – a survey study.* International Journal of Information Management, 60:102375, 2021, ISSN 0268-4012. https://www.sciencedirect.com/science/article/pii/S0268401221000682. 72

[98] Meyer-Waarden, Lars, Giulia Pavone, Thanida Poocharoentou, Piyanut Prayatsup, Maelis Ratinaud, Agathe Tison, and Sara Torné: *How service quality influences customer acceptance and usage of chatbots?* SMR-Journal of Service Management Research, 4(1):35–51, 2020. 72

[99] Clemmensen, Torkil, Morten Hertzum, Jiaoyan Yang, and Yanan Chen: *Do usability professionals think about user experience in the same way as users and developers do?* In *IFIP Conference on Human-Computer Interaction*, pages 461–478. Springer, 2013. 78

[100] Leiva, Germán, Nolwenn Maudet, Wendy Mackay, and Michel Beaudouin-Lafon: *Enact: Reducing designer–developer breakdowns when prototyping custom interactions.* ACM Transactions on Computer-Human Interaction (TOCHI), 26(3):1–48, 2019. 78

[101] Svikhnushina, Ekaterina, Alexandru Placinta, and Pearl Pu: *User expectations of conversational chatbots based on online reviews.* In *Designing Interactive Systems Conference 2021*, pages 1481–1491, 2021. 78

[102] Zhang, Xiao and Ron Wakkary: *Understanding the role of designers' personal experiences in interaction design practice.* In *Proceedings of the 2014 conference on Designing interactive systems*, pages 895–904, 2014. 78

[103] Lin, Yu Tzu and Morten Hertzum: *How do designers make user-experience design decisions?* In *International Conference on Human-Computer Interaction*, pages 188–198. Springer, 2020. 78

[104] Aghajani, Emad, Csaba Nagy, Mario Linares-Vásquez, Laura Moreno, Gabriele Bavota, Michele Lanza, and David C Shepherd: *Software documentation: the practitioners' perspective.* In *2020 IEEE/ACM 42nd International Conference on Software Engineering (ICSE)*, pages 590–601. IEEE, 2020. 79

[105] Abdellatif, Ahmad, Diego Costa, Khaled Badran, Rabe Abdalkareem, and Emad Shihab: *Challenges in chatbot development: A study of stack overflow posts.* In *Proceedings of the 17th international conference on mining software repositories*, pages 174–185, 2020. 79

[106] Zhou, Xin, Yuqin Jin, He Zhang, Shanshan Li, and Xin Huang: *A map of threats to validity of systematic literature reviews in software engineering.* In Potanin, Alex, Gail C. Murphy, Steve Reeves, and Jens Dietrich (editors): *23rd Asia-Pacific Software Engineering Conference, APSEC 2016*, pages 153–160, Hamilton, New Zealand, 2016. IEEE Computer Society. https://doi.org/10.1109/APSEC.2016.031. 84

Primary Studies

[PS1] Candello, Heloisa, Claudio Pinhanez, and Flavio Figueiredo: *Typefaces and the perception of humanness in natural language chatbots.* In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, page 3476–3487, New York, NY, USA, 2017. Association for Computing Machinery, ISBN 9781450346559. https://doi.org/10.1145/3025453.3025919.

[PS2] Araujo, Theo: *Living up to the chatbot hype: The influence of anthropomorphic design cues and communicative agency framing on conversational agent and company perceptions.* Computers in Human Behavior, 85:183–189, 2018, ISSN 0747-5632. https://www.sciencedirect.com/science/article/pii/S0747563218301560.

[PS3] Fadhil, Ahmed, Gianluca Schiavo, Yunlong Wang, and Bereket A. Yilma: *The effect of emojis when interacting with conversational interface assisted health coaching system.* In *Proceedings of the 12th EAI International Conference on Pervasive Computing Technologies for Healthcare*, PervasiveHealth '18, page 378–383, New York, NY, USA, 2018. Association for Computing Machinery, ISBN 9781450364508. https://doi.org/10.1145/3240925.3240965.

[PS4] Gnewuch, U., S. Morana, M.T.P. Adam, and A. Maedche: *Faster is not always better: Understanding the effect of dynamic response delays in human-chatbot interaction.* 2018.

[PS5] Ashktorab, Z., M. Jain, Q. Vera Liao, and J.D. Weisz: *Resilient chatbots: Repair strategy preferences for conversational breakdowns.* 2019.

[PS6] Diederich, S., M. Janßen-Müller, A.B. Brendel, and S. Morana: *Emulating empathetic behavior in online service encounters with sentiment-adaptive responses: Insights from an experiment with a conversational agent.* 2019.

[PS7] Zhou, M.X., G. Mark, J. Li, and H. Yang: *Trusting virtual agents: The effect of personality.* ACM Transactions on Interactive Intelligent Systems, 9(2-3), 2019.

[PS8] Bawa, Anshul, Pranav Khadpe, Pratik Joshi, Kalika Bali, and Monojit Choudhury: *Do multilingual users prefer chat-bots that code-mix? let's nudge and find out!* Proc. ACM Hum.-Comput. Interact., 4(CSCW1), may 2020. https://doi.org/10.1145/3392846.

[PS9] Beattie, A., A.P. Edwards, and C. Edwards: *A bot and a smile: Interpersonal impressions of chatbots and humans using emoji in computer-mediated communication.* Communication Studies, 71(3):409–427, 2020.

[PS10] De Cicco, R., S.C. e Silva, and F.R. Alparone: *Millennials' attitude toward chatbots: an experimental study in a social relationship perspective.* International Journal of Retail and Distribution Management, 48(11):1213–1233, 2020.

[PS11] Diederich, S., A.B. Brendel, S. Lichtenberg, and L.M. Kolbe: *Design for fast request fulfillment or natural interaction? insights from an experiment with a conversational agent.* 2020.

[PS12] Hendriks, F., C.X.J. Ou, A.K. Amiri, and S. Bockting: *The power of computer-mediated communication theories in explaining the effect of chatbot introduction on user experience.* Volume 2020-January, pages 271–278, 2020.

[PS13] Lee, Y. C., N. Yamashita, Y. Huang, and W. Fu: *"i hear you, i feel you": Encouraging deep self-disclosure through a chatbot.* 2020.

[PS14] Liao, Y. and J. He: *Racial mirroring effects on human-agent interaction in psychotherapeutic conversations.* pages 430–442, 2020.

[PS15] Narducci, F., P. Basile, M. de Gemmis, P. Lops, and G. Semeraro: *An investigation on the user interaction modes of conversational recommender systems for the music domain.* User Modeling and User-Adapted Interaction, 30(2):251–284, 2020.

[PS16] Ng, M., K.P.L. Coopamootoo, E. Toreini, M. Aitken, K. Elliot, and A. Van Moorsel: *Simulating the effects of social presence on trust, privacy concerns & usage intentions in automated bots for finance.* pages 190–199, 2020.

[PS17] Sheehan, Ben, Hyun Seung Jin, and Udo Gottlieb: *Customer service chatbots: Anthropomorphism and adoption.* Journal of Business Research, 115:14–24, 2020, ISSN 0148-2963. https://www.sciencedirect.com/science/article/pii/S0148296320302484.

[PS18] Shi, Weiyan, Xuewei Wang, Yoo Jung Oh, Jingwen Zhang, Saurav Sahay, and Zhou Yu: *Effects of Persuasive Dialogues: Testing Bot Identities and Inquiry Strategies*, page 1–13. Association for Computing Machinery, New York, NY, USA, 2020, ISBN 9781450367080. https://doi.org/10.1145/3313831.3376843.

[PS19] Toader, D. C., G. Boca, R. Toader, M. Măcelaru, C. Toader, D. Ighian, and A.T. Rădulescu: *The effect of social presence and chatbot errors on trust.* Sustainability (Switzerland), 12(1), 2020.

[PS20] Xiao, Ziang, Michelle X. Zhou, Wenxi Chen, Huahai Yang, and Changyan Chi: *If I Hear You Correctly: Building and Evaluating Interview Chatbots with Active Listening Skills*, page 1–14. Association for Computing Machinery, New York, NY, USA, 2020, ISBN 9781450367080. https://doi.org/10.1145/3313831.3376131.

[PS21] Ahmad, Rangina, Dominik Siemon, and Susanne Robra-Bissantz: *Communicating with machines: Conversational agents with personality and the role of extraversion.* In *54th Hawaii International Conference on System Sciences, HICSS 2021*, pages 1–10, Kauai, Hawaii, USA, 2021. ScholarSpace. http://hdl.handle.net/10125/71109.

[PS22] Bührke, J., A.B. Brendel, S. Lichtenberg, M. Greve, and M. Mirbabaie: *Is making mistakes human? on the perception of typing errors in chatbot communication.* Volume 2020-January, pages 4456–4465, 2021.

[PS23] Ceha, Jessy, Ken Jen Lee, Elizabeth Nilsen, Joslin Goh, and Edith Law: *Can a Humorous Conversational Agent Enhance Learning Experience and Outcomes?* Association for Computing Machinery, New York, NY, USA, 2021, ISBN 9781450380966. https://doi-org.ez54.periodicos.capes.gov.br/10.1145/3411764.3445068.

[PS24] De Cicco, R., S.C.L.D.C.E. Silva, and F.R. Alparone: *"it's on its way": Chatbots applied for online food delivery services, social or task-oriented interaction style?* Journal of Foodservice Business Research, 24(2):140–164, 2021.

[PS25] Kull, A.J., M. Romero, and L. Monahan: *How may i help you? driving brand engagement through the warmth of an initial chatbot message.* Journal of Business Research, 135:840–850, 2021.

[PS26] Ma, Y., T. Kleemann, and J. Ziegler: *Mixed-modality interaction in conversational recommender systems.* Volume 2948, pages 21–37, 2021.

[PS27] Mozafari, N., W.H. Weiger, and M. Hammerschmidt: *Resolving the chatbot disclosure dilemma: Leveraging selective self-presentation to mitigate the negative effect of chatbot disclosure.* Volume 2020-January, pages 2916–2923, 2021.

[PS28] Mozafari, N., W.H. Weiger, and M. Hammerschmidt: *Trust me, i'm a bot – repercussions of chatbot disclosure in different service frontline settings.* Journal of Service Management, 2021.

[PS29] Pizzi, G., D. Scarpi, and E. Pantano: *Artificial intelligence and the new forms of interaction: Who has the control when interacting with a chatbot?* Journal of Business Research, 129:878–890, 2021.

[PS30] Rana, Kanishk, Rahul Madaan, and Jainendra Shukla: *Effect of polite triggers in chatbot conversations on user experience across gender, age, and personality.* In *30th IEEE International Conference on Robot & Human Interactive Communication, 2021*, pages 813–819, Vancouver, BC, Canada, 2021. IEEE. https://doi.org/10.1109/RO-MAN50785.2021.9515528.

[PS31] Schanke, S., G. Burtch, and G. Ray: *Estimating the impact of "humanizing" customer service chatbots.* Information Systems Research, 32(3):736–751, 2021.

[PS32] Spillner, Laura and Nina Wenig: *Talk to Me on My Level – Linguistic Alignment for Chatbots.* Association for Computing Machinery, New York, NY, USA, 2021, ISBN 9781450383288. https://doi.org/10.1145/3447526.3472050.

[PS33] Tsai, W. H.S., Y. Liu, and C. H. Chuan: *How chatbots' social presence communication enhances consumer engagement: the mediating role of parasocial interaction and dialogue*. Journal of Research in Interactive Marketing, 15(3):460–482, 2021.

[PS34] Wilkinson, Daricia, Öznur Alkan, Q. Vera Liao, Massimiliano Mattetti, Inge Vejsbjerg, Bart P. Knijnenburg, and Elizabeth Daly: *Why or why not? the effect of justification styles on chatbot recommendations*. ACM Trans. Inf. Syst., 39(4), oct 2021, ISSN 1046-8188. https://doi.org/10.1145/3441715.

[PS35] Medeiros, Lenin, Tibor Bosse, and Charlotte Gerritsen: *Can a chatbot comfort humans? studying the impact of a supportive chatbot on users' self-perceived stress*. IEEE Transactions on Human-Machine Systems, pages 1–11, 2021, ISSN 2168-2305.

[PS36] Svikhnushina, Ekaterina and Pearl Pu: *Key Qualities of Conversational Chatbots – the PEACE Model*, page 520–530. Association for Computing Machinery, New York, NY, USA, 2021, ISBN 9781450380171. https://doi-org.ez54.periodicos.capes.gov.br/10.1145/3397481.3450643.

[PS37] Chaves, Ana Paula, Jesse Egbert, Toby Hocking, Eck Doerry, and Marco Aurelio Gerosa: *Chatbots language design: The influence of language variation on user experience with tourist assistant chatbots*. ACM Trans. Comput.-Hum. Interact., 29(2), jan 2022, ISSN 1073-0516. https://doi-org.ez54.periodicos.capes.gov.br/10.1145/3487193.

[PS38] Pecune, F., L. Callebert, and S. Marsella: *Designing persuasive food conversational recommender systems with nudging and socially-aware conversational strategies*. Frontiers in Robotics and AI, 8, 2022. cited By 0.

[PS39] Esmark Jones, C.L., T. Hancock, B. Kazandjian, and C.M. Voorhees: *Engaging the avatar: The effects of authenticity signals during chat-based service recoveries*. Journal of Business Research, 144:703–716, 2022. cited By 0.

[PS40] Rhim, J., M. Kwak, Y. Gong, and G. Gweon: *Application of humanization to survey chatbots: Change in chatbot perception, interaction experience, and survey data quality*. Computers in Human Behavior, 126, 2022.