Universidade Estadual de Campinas
Instituto de Computação

Paula Jeniffer dos Santos Viriato

Humanization Evaluation in Chatbots

Avaliação da Humanização em Chatbots

CAMPINAS
2022

Paula Jeniffer dos Santos Viriato

Humanization Evaluation in Chatbots

Avaliação da Humanização em Chatbots

Dissertação apresentada ao Instituto de Computação da Universidade Estadual de Campinas como parte dos requisitos para a obtenção do título de Mestra em Ciência da Computação.

Dissertation presented to the Institute of Computing of the University of Campinas in partial fulfillment of the requirements for the degree of Master in Computer Science.

Supervisor/Orientador: Professor Doutor Julio Cesar dos Reis
Co-supervisor/Coorientador: Professor Doutor Leandro Aparecido Villas

Este exemplar corresponde à versão final da Dissertação defendida por Paula Jeniffer dos Santos Viriato e orientada pelo Professor Doutor Julio Cesar dos Reis.

CAMPINAS
2022

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca do Instituto de Matemática, Estatística e Computação Científica
Ana Regina Machado - CRB 8/5467

V819h

Viriato, Paula Jeniffer dos Santos, 1996-
Humanization evaluation in chatbots / Paula Jeniffer dos Santos Viriato. – Campinas, SP : [s.n.], 2022.

Orientador: Julio Cesar dos Reis.
Coorientador: Leandro Aparecido Villas.
Dissertação (mestrado) – Universidade Estadual de Campinas, Instituto de Computação.

1. Chatbot - Avaliação. 2. Interação humano-computador. 3. Robôs - Programação. I. Reis, Julio Cesar dos, 1987-. II. Villas, Leandro Aparecido, 1983-. III. Universidade Estadual de Campinas. Instituto de Computação. IV. Título.

Informações Complementares

Título em outro idioma: Avaliação de humanização em chatbots
Palavras-chave em inglês:
Chatbot evaluation
Human-computer interaction
Robots - Programming
Área de concentração: Ciência da Computação
Titulação: Mestra em Ciência da Computação
Banca examinadora:
Julio Cesar dos Reis [Orientador]
Islene Calciolari Garcia
Lara Schibelsky Godoy Piccolo
Data de defesa: 30-09-2022
Programa de Pós-Graduação: Ciência da Computação

Identificação e informações acadêmicas do(a) aluno(a)
- ORCID do autor: https://orcid.org/0000-0003-0900-1686
- Currículo Lattes do autor: http://lattes.cnpq.br/2868807972894398

Universidade Estadual de Campinas
Instituto de Computação

Paula Jeniffer dos Santos Viriato

Humanization Evaluation in Chatbots

Avaliação da Humanização em Chatbots

Banca Examinadora:

- Professor Doutor Julio Cesar dos Reis
  Universidade Estadual de Campinas (UNICAMP) - Presidente
- Professora Doutora Islene Calciolari Garcia
  Universidade Estadual de Campinas (UNICAMP) - Membro Titular Interno
- Doutora Lara Schibelsky Godoy Piccolo
  The Open University - Membro Titular Externo

A ata da defesa, assinada pelos membros da Comissão Examinadora, consta no SIGA/Sistema de Fluxo de Dissertação/Tese e na Secretaria do Programa da Unidade.

Campinas, 30 de setembro de 2022

Acknowledgements

First, I would like to thank Professors Julio Cesar dos Reis, Leandro Aparecido Villas, and post-doctoral fellow Rafael Roque de Souza. For all the guidelines, both as a student and as a researcher, and also as a future professor. I looked up the word teacher in the dictionary, the word *professor*, and I found that it is *the one who professes, who shows us the future*. I also looked up the word *advisor* in the dictionary and found that it is *the one who directs, leads, and guides in a direction*. In the three, I found these exact meanings, human people and patients who showed me the possibilities of the future every day and guided me towards what is best for me.

I thank my family. As I have a large family, I need to thank my parents, Paulo and Enildes, my brothers Enilse and Jefferson, and my sister that life gave me, Raula, married to my other brother Jefferson. I also take this opportunity to thank the people who inspired me with creativity and enthusiasm every day, my nephews Vitor and Gabriel.

I also thank the two people who will forever live in my heart, Rodrigo Oliveira and Christiane Cantagalli, who have always supported me in continuing my academic life. Furthermore, I also take this opportunity to thank my best friends, Murilo Oliveira, Júlia Nunes, and Gustavo Salmerón, who took care of me and were always around.

I cannot fail to thank those to whom I have devoted faith and inspiration. Our Lady, who always inspired me with patience and resilience, Saint Francis of Assisi, who always inspired charity at any cost, and Saint Jude Thaddeus, who inspired me to follow my vocation even when the weight was heavy.

Finally, I am immensely grateful for the partnership between CI&T and Unicamp, managed by FUNCAMP, which is a partnership of great dedication and knowledge generation in which I participated throughout my master's degree. I want to take this opportunity to thank the 60 students of the MC750A discipline who participated in the use case presented in this work and the doctoral student André Regino, who helped me by monitoring the class with full dedication.

Resumo

Chatbots são softwares que simulam tarefas de conversação humana. Eles podem ser usados para diversos fins, incluindo atendimento ao cliente e conversas terapêuticas. Um desafio encontrado no desenvolvimento de chatbots é torná-los mais humanizados. Um chatbot pode ser humanizado quando realiza uma conversa fluida e agradável com o usuário, demonstrando empatia e personalidade. Outros aspectos são relevantes para a percepção de humanidade, como o visual da plataforma de conversação e a credibilidade passada pelo chatbot analisado pelo usuário. Esta proposta de mestrado visa estudar e desenvolver um método de avaliação que indique o nível de humanização de um chatbot em análise. Nosso método consiste em dois objetivos e questionários breves, que podem ser aplicados para estabelecer uma avaliação adaptável aos diferentes objetivos no uso de chatbots (comerciais ou terapêuticos, por exemplo). O primeiro questionário refere-se ao grau de relevância dos fatores de humanização para o funcionamento de um chatbot avaliado. Este questionário visa ponderar os fatores de acordo com os objetivos específicos do chatbot; o segundo questionário refere-se à eficácia do chatbot analisada quanto aos fatores de humanização. Combinamos esses dois questionários para gerar métricas de avaliação, que fornecem uma pontuação de humanização do software avaliado. Nossa proposta pode ajudar os designers a identificar fatores específicos que afetam a experiência dos usuários interagindo com chatbots.

Abstract

Chatbots are software that simulates human conversational tasks. They can be used for various purposes, including customer service and therapeutic conversations. A challenge encountered in developing chatbots is making them more humanized. A chatbot can be humanized when it performs a fluid and pleasant conversation with the user, demonstrating empathy and personality. Other aspects are relevant to the perception of humanity, such as the look of the conversation platform and the credibility passed by the chatbot analyzed to the user. This MSc Thesis aims to study and develop an evaluation method that indicates a chatbot's humanization level under analysis. Our method consists of two objectives and brief questionnaires, which can be applied to establish an assessment adaptable to the different objectives in using chatbots (commercial or therapeutic, for example). The first questionnaire refers to a degree of relevance regarding humanization factors for the functioning of an evaluated chatbot. This questionnaire aims to weigh the factors according to the chatbot's specific objectives; the second questionnaire refers to the chatbot effectiveness analyzed regarding the humanization factors. We combine these two questionnaires to generate assessment metrics, which provide a humanization score from the evaluated software. Our proposal helps designers identify specific factors affecting users' experience interacting with chatbots. We conducted a case study in applying our framework and reveled key findings regarding its applicability.

List of Abbreviations and Acronyms

NLP Natural Language Processing

FAQs Frequently Asked Questions

AI Artificial Intelligence

SUS System Usability Scale

UEQ User Experience Questionnaire

CUQ Chatbot Usability Questionnaire

TODS Task-Oriented Dialog Systems

ECA Embedded Conversation Agents

ML Machine Learning

AM Analogical Modeling

NLP Natural Language Processing

NLU Natural Language Understanding

Contents

Chapter 1

Introduction

Chatbots and platforms to develop them are increasingly common in the market. They aim to meet the high demands of solving problems arising from current activities, mainly related to customer service. Such software can perform standard and medium complexity tasks, making the human workforce specialized in solving highly complex tasks. These agents assist humans in performing tasks by using Natural Language Processing (NLP) with text and/or voice interactions. In literature, it is observed that the use of these agents is directed only to typical and medium complexity tasks, becoming powerful strategic tools. An example of a low complexity task, such as questions on Frequently Asked Questions (FAQs) pages; examples of medium complexity tasks are those that involve database manipulation, which requires a higher level of integration and customization.

Several chatbots help in business activities: optimization, proactivity, and protection. Optimization chatbots can develop practical activities, which require querying the company's databases and providing services to users. Such chatbots develop services similar to SAC (customer service), considered tasks of medium complexity. They require a certain level of personalization with user information. On the other hand, proactive chatbots monitor the users' activities in the system. These chatbots perform tasks by guiding the user with possible questions and are considered simple to develop software. Protection chatbots serve a similar purpose as FAQs, answering simple questions and, when necessary, referring more complex questions and problems to human attendants [56].

Social chatbots help closed domain systems with specific tasks by interacting or mediating in virtual social groups, conducting voting, playing music, or assisting in joint decision making, for example. Conversely, conversational chatbots aim to talk to users like humans in a typical chat. Conversation-based bots have the function of responding to various things to maintain an ongoing conversation. Typically, open domain systems fall into the conversational category [47]. This categorization includes chatbots with more closed dialogues that use static responses. There are chatbots with more open dialogues with NLP models. The latter should better classify user messages and offer more natural, and fluid responses, similar to the responses humans would give. That is why increasingly sophisticated chatbot development methods are being researched with Artificial Intelligence (AI) resources [34]

All the described categories of chatbots have a common goal: to appear more human and maintain or increase the ability to solve possible tasks. These people have proven to be

efficient in performing their tasks. We observed in studies that users had little interaction with the chatbot and were unsatisfied with the service due to the lack of humanization [45]. The task of humanizing a chatbot, which is to make its behavior similar to that of a human being, is not trivial. Human behavior is complex and varies according to the nature of situations [20]. Each person has his or her perception of humanity. We know how to differentiate typical machine behavior from human behavior, but explaining such differences is complicated. Users tend to have a better dialog with human attendants rather than chatbots. This situation justifies the increasing humanization of chatbots to improve user satisfaction with such systems.

This MSc. thesis aims to investigate and establish a methodology to assess the level of humanization in chatbots. We aim to respect the peculiarities and system's purposes (nature) and offer ways to identify aspects of humanization unsuitable for the analyzed chatbot. Our proposed methodology is the basis for generating a humanization metric that can help designers and developers refine these aspects of the system.

1.1 Problem Characterization

The concept of humanization does not have a simple definition, and personal perceptions of humanity are pretty subjective [19]. About chatbots, the perception of humanity should be linked to the resolution of the proposed tasks. In this context, a chatbot should meet users' expectations for their respective purposes. There is a challenge in building chatbots, which is to establish ways to keep users willing to develop a dialogue with the chatbot, respond to requests for information, and remain confident about the effectiveness of the software in solving their problems. Humans, due to the feeling of empathy, tend to trust human attendants more than machines [1]. It seems that humans are crucial to the perceived credibility and satisfaction of the end consumer towards the chatbot [36].

A chatbot is expected to build meaningful, personalized, and friendly dialogues through a conversational system. Such a conversational system should evolve with each interaction and offer alternative answers to similar questions, answering the interlocutor's questions. Another human person would meet the user's expectations more naturally and fluidly. The chatbot should be able to engage cordially with the consumer, making them feel comfortable, satisfied, and valued [4]. Each chatbot has its characteristics, purposes, and target audience. The perceptions of humanity related to the reality of the chatbot should be prioritized. For example, it is not polite in society to tell jokes in grieving situations; neither would it be appropriate to use colloquial language in chatbot legal advice. Thus, humanization assessments should adapt to the different contexts of needs and chatbots.

Humanizing a chatbot is a gradual and constant task. It is crucial to evaluate the level of humanization of a chatbot. Evaluation helps to improve and make comparisons to propose and apply changes that cause positive effects on the results. A humanization metric can be an essential quantitative tool in developing a chatbot to improve the user experience, allowing an evolutionary comparison of such metrics in humanization of different software versions, and making it possible to diagnose the strengths and weaknesses of the released versions. There is a consensus that the humanization of chatbots should be constantly

improved, with each chatbot being evaluated for its level of humanization, considering its particular characteristics [15]. Chatbot users should be the central characters in this kind of evaluation. It is their perceptions that affect the success of a chatbot.

1.2 Objectives

This master thesis aims to study and develop a methodology for evaluating the level of humanization in chatbots. Our goal is to adopt such a methodology for different purposes and possible target audiences. Our methodology generates metrics that indicate "how close to a perfect level of humanization" the chatbot is. We investigate a guide of factors that affects humanization results in chatbots. This investigation aims to achieve the following specific objectives:

1. Mapping and compiling the broadest possible set of characteristics that impact the perception of humanity that users have towards a chatbot (impact factors).
2. Develop a quantitative methodology to assess the level of humanization in chatbots.
   (a) A methodology that allows application in generic contexts.
   (b) A methodology that is adaptable to different target audiences of chatbots.
   (c) A methodology that is adaptable to different sets of humanization impact factors.
3. Propose and analyze metrics in humanization for quantitative and qualitative evaluation of chatbots.

1.3 Study Context

This master thesis is part of a larger project, which aims to develop a framework for building humanized chatbots supported by the company CI&T. It aims to develop a framework for building humanized chatbots, which consists of a high-level chatbot creation tool based on a pre-trained model using the end-to-end UBAR architecture [55]. This model can perform small experimental learning on a small set of annotated dialogues to quickly deploy a functional chatbot. This rapid adaptation is made possible by the multi-sequence training regime, which allows the end user to quickly create state-of-the-art Task-Oriented Dialog Systems (TODS) [7] at reduced costs, achieving a better user experience than current solutions available on the market [9]. We have designed our methodology inspired and contextualized in this project context. We conducted a case study in a distinct context (undergraduate course at IC/UNICAMP) where student groups created and evaluated chatbots based on our framework here developed.

1.4 Organization of the Dissertation

The remaining of this document is organized as follows:

- **Chapter 2:** presentation of the Theoretical Reference, comprising the Fundamental Concepts (Section 2.1) and the Related Work (Section 2.2). A discussion is conducted (in Section 2.3) on correlated investigations concerning the objectives of our work; we provide final remarks indicating the contributions our study to the state of the art (in Section 2.4).
- **Chapter 3:** This Chapter presents our proposed Principles of Humanization. First, we present the methodology (Section 3.1) used to identify which characteristics (impact factors) present in literature and describe the concept of humanization in the context of chatbots. A compiled set of impact factors was found and presented (Section 3.2). We present a brief assessment regarding the relevance of these impact factors to three chatbots available from the market (Section 3.3). Section 3.4 discusses the results achieved and we conclude how such a study should be developed through an alternative evaluation methodology (Section 3.5).
- **Chapter 4:** presentation of our Chatbots Humanization Assessment Methodology, proposing a quantitative and adaptive assessment methodology. First, we report on the process of identifying the set of impact factors (in Section 4.1); we present how the assessment instrument was developed (in Section 4.2); and finally we explain the ideal and adequate application of the developed assessment instrument (in Section 4.3). We discussed possible variations of the methodology (in Section 4.4), and concluded (in Section 4.5) how the methodology contributes to the construction of a new baseline in chatbots humanization evaluation.
- **Chapter 5:** This presents the conducted Case Study, in which the methodology proposed was applied during a course on Human-Computer Interaction (HCI). Initially, we describe the context of the discipline in which the methodology was applied and the participants (in Section 5.1). We present the methodology for applying the evaluation simultaneously with the discipline itself (in Section 5.2). Section 5.3 presents the results achieved with this case study. Section 5.4 discusses how the results achieved prove the validity of the hypotheses made regarding the methodology. Finally, we conclude (in Section 5.5) how the results obtained allow further advances in the context of humanization assessments.
- **Chapter 6:** This chapter provides the Conclusion, composed of the contributions of this MSc. dissertation (Section 6.1), limitations that still need to be overcome (Section in 6.2), and the advances that can still be made in future work (Section 6.3).

Chapter 2

Theoretical Reference

This chapter presents the fundamental concepts involved in our research and the current state-of-the-art in humanization of chatbots.

2.1 Fundamental Concepts

Chatbots are software systems capable of interacting with humans in natural language in a given domain [38]. There are currently several nomenclatures for this system: Embedded Conversation Agents (ECA), Conversation Systems, Agents, Chatterbots, or just bots. [27]. Many of these agents are designed to use NLP so that users can type or write to or agent that would do them to a human. The agent can then analyze the input and respond appropriately in a conversational manner, as examples of systems that use artificial intelligence to develop better human-computer interaction by simulating conversations between human users [5, 24].

Adamopoulou and Moussiades, based on the definitions of Bansal and Khan [5] and Khanna *et al.* [24], defined *chatbots* as examples of systems that use artificial intelligence to develop a better human-computer interaction, simulating conversations between human users. Currently, chatbots communicate through text or voice and use advanced techniques in NLP to understand, classify and generate conversations with human beings [2].

Currently, chatbots communicate through text or voice and use advanced NLP techniques to understand, classify and generate conversations with humans. For Go and Sundar [17], the primary function of chat agents is to interact with users by answering their questions and resolving their requests. The authors explained that the experience provided by these agents is considered better than static information delivery, such as a list of FAQs, since agents provide information more interactively.

As for interaction design, Følstad, Skjuve, and Brandtzaeg [12] classify chatbots into four categories: chatbots for customer support, personal assistant chatbots; content curation chatbots; and coaching chatbots. Our methodology proposed in this study can be applied to any chatbots presented by Følstad, Skjuve, and Brandtzaeg [12]. The examples presented in this paper focus on chatbots for customer support.

Chatbots based on AI [23] can understand natural language and not only predefined commands. In their development, they present enough skills to interact with users.

Moreover, they can maintain different contexts of conversations and provide the user with richer and more engaging conversations. Because of this, concepts of virtual agents and speech recognition techniques used in virtual agents, such as Apple Siri, Amazon Alexa, and Google Assistant, have emerged. Thus, one can define NLP as an area of AI, which aims the study and creation of techniques that enable the analysis and understanding of human language through a computational system [33].

The central goal of NLP is to enable humans and machines to communicate naturally, without users having to learn a new language [25]. These systems use NLP and Machine Learning algorithms. One often finds a technique that uses variables divided into intentions, entities, and contexts in their internal structure. The use of these concepts aims to represent the information obtained from the user in a structured way, to generate a coherent answer [41]. Natural Language Understanding (NLU) is a subarea of NLP that uses syntactic and semantic analysis of text or speech to classify the meaning of a sentence. Syntax refers to the grammatical structure of a sentence, while semantics concerns its meaning [6, 18].The Analogical Modeling (AM) is a subarea of AI that allows the creation computer programs with the ability to learn and perform tasks [13].

Machine Learning (ML) can be used in Chatbots to help diagnose diseases. During the conversation between the ChatBot and the user, the symptoms are identified by the ChatBot, which can recognize the disease and recommend the appropriate treatment[29]. The development of AM algorithms made it possible to advance in NLP. A large data set can create a trained model; the algorithms use this data to learn patterns and apply them to new inputs. It can do its job automatically [10]. Several techniques ML are currently applied; some of the most used classifications are supervised learning algorithms, unsupervised and neural networks.

2.1.1 Assessment Perspectives on Chatbots

Given the studies analyzed, we observed that evaluations in chatbots are carried out qualitatively, observing the users' perception of the systems. Peras [40] addresses the evaluation of chatbots as an activity from five perspectives: user experience, information retrieval; linguistics; technology; and business. The user experience perspective consists of usability, performance, affectivity, and user satisfaction concerning the chatbot. The information retrieval perspective focuses on the accuracy, accessibility, and efficiency of information delivery. On the other hand, the linguistic perspective looks at concepts such as quality, quantity, relatedness, manner, and grammatical accuracy presented in conversations with chatbots. The technological perspective refers to the humanity of the chatbot, which we will delve into later (in Chapter 3). The business perspective proposes more specific metrics for chatbot effectiveness: number of users, duration of chatbot conversation, number of conversations, number of agents included in a conversation, number of failed conversations, number of inappropriate responses, and number of repeated consultations.

In our investigation, the focus is to measure qualitative aspects more precisely. We aim to convert them to quantitative values in metrics to promote systems' evolution (design refinement) in such aspects. According to Peras [40], among the five perspectives,

only the information retrieval perspective involves only quantitative aspects. Two of the perspectives cited are strongly guided by qualitative aspects: the linguistic and technological perspectives.

2.1.2 Humanity in Chatbots

The technological perspective investigated by Peras [40] was based on the analysis of the humanity of chatbots. Humanity is described in this context as the ability of the chatbot to express human behavior. For this purpose, the chatbot must process, understand and generate natural language. Humanity can be measured both qualitatively and quantitatively. Qualitative assessments are much more used in evaluating humanity, being necessary for providing feedback for the advancement of chatbots.

For Peras [40], a chatbot demonstrates humanity when it presents: naturalness, ability to maintain a thematic discussion; ability to answer specific questions; and ability to understand natural language. Some of our already obtained results (cf. Section 3.2) indicate that key factors that are considered in the conception of humanity vary significantly from one study to another. Therefore, one of the specific objectives of this study is to map from literature factors considered relevant to the perception of humanity that the user has about a chatbot (cf. Chapter 3).

Regarding the humanization of chatbots, Reeves and Nass [43] and Rhim *et al.* [45] observed that users react socially to a computer that exhibits human-like behavior. Users feel that the system behavior is similar to that of other humans, even though they know they are interacting with a computer.

We can, therefore, define *humanization impact factors* as characteristics that strongly influence the perception of humanity that users have about chatbots. Still, in this context of humanization impact factors, Go and Sundar [17] carried out a study to analyze the effects of visual, identity, and behavioral clues on the perception of humanity that users have about chatbots. Such classification of impact factors (visual, identity, and behavioral) is quite interesting in development. Graphical interfaces can influence visual factors, identity factors can influence by personalization, and behavioral factors can be influenced by modeling artificial intelligence. In this way, this MSc. dissertation aims to map the impact factors in this classification, which can help evolve conversational systems. All humanization impact factors cited in this proposal is further described in section 3.2.

The scientific literature considers several factors impacting the humanization of chatbots to leave these agents with more humanistic characteristics to be noticeable by their users. In this context, social presence is a commonly discussed element in studies related to chatbots. Adding this kind of impact factor to the agent means incorporating "sensitive human contact" because when interacting with a chatbot, users have the opportunity to make social presence attributions at first [50]. The more socially present the interactions are, the more engaging the interface will be; however, the more human the interface is, the higher expectations the user will have from the agent [32]. In addition to greeting, language choices are crucial in humanizing chatbots. For example, using a more polite, informal, or social language can help induce anthropomorphic perceptions and perceptions of social presence.

It is with these linguistic features that developers help impose a sense of social presence and further promote humanization in their chatbot [46]. The linguistic communication features, as humanization impact factor employed by both researchers and practitioners in dialogue delays. From one perspective, delays could be interpreted as the chatbot not functioning as expected. However, when implemented correctly, minor delays that are dynamic to the amount of text can dictate levels of persuasiveness and personality perceptions of the chatbot. This type of feature can make the agent more real and humanized because humans do not instantly read and respond to messages sent via text media [16].

In addition to these aspects, humor is a key factor in humanizing chatbots. Humor has been shown to introduce feelings of common ground between two communicating social actors. Like human interactions, humor can be an effective way to personify systems and create a more engaging interaction. Furthermore, humor in task-oriented communication has increased the number of individuals with higher satisfaction with chatbots. Humor in both business and customer service interactions requires a more nuanced approach. Whereas humor in an e-service encounter can help in some situations where the process is to your liking, when the process is not to your liking, the addition of humor exacerbates the negative feelings associated with the service experience. In this sense, it is necessary to have a middle ground when it comes to humanizing the chatbot, looking specifically at its applied context [35].

2.1.3 Psychometric Analysis and Likert Scale

One of the forms of qualitative assessment most suggested by Peras [40] was the Likert scale, a widely used psychometric technique. Pasquali [39] defines *psychometry* as the theory and technique of measuring mental processes, and this is primarily applied in the fields of psychology and education. Psychometrics would transform qualitative aspects, commonly expressed in ordinary language, into quantitative metrics, which can be measured and compared. Also, Pasquali [39] explained that, in a general sense, psychometrics aims to explain the meaning of the responses given by the subjects in a series of tasks usually called items. We present these concepts because they are relevant for the construction of our framework.

According to Joshi *et al.* [22], which was based on the work of Edmondson [11] and McLeod [30], the Likert scale is a psychometric technique that was developed to measure 'attitude' in a scientifically accepted and validated way. An *attitude* can be defined as a preferential behavior in a specific circumstance. Attitude is usually rooted in the lasting organization of beliefs and ideas (around an object, a subject, or a concept) acquired through social interactions [37].

The original Likert scale is a set of statements (items) offered for a real or hypothetical situation under study. Participants are asked to show their levels of agreement (from strongly disagree to agree strongly) with the given statement (items) on a metric scale. Here all the statements in combination reveal the specific dimension of the attitude towards the problem, therefore, necessarily interlinked with each other [48].

In the case of evaluation of humanity in chatbot, the concepts evaluated are factors

that, in an integrated way, make up the perception of humanity that human beings have regarding the system. Although factors are subjective and abstract, psychometric scales help quantify the perception of these characteristics. Humanization impact factors are the characteristics that most influence our perception of humanity in a chatbot. Examples of humanization impact factors are naturalness and empathy (cf. Chapter 3). That is why psychometric analysis and the Likert scale play a key role in our study.

2.1.4 Metrics for Likert Scale Analysis

Croasmun [8], based on the book of Mills [31], explains that when using Likert-type scales, it is essential that the researchers calculate and report Cronbach's α coefficient for internal consistency reliability. Internal consistency reliability refers to the extent to which items in an instrument are consistent among themselves and with the overall instrument; Cronbach's α estimates the internal consistency reliability of an instrument by determining how all items in the instrument report to all other items and the actual instrument.

Like Cronbach's α coefficient, other coefficients, such as Revelle's β and McDonald's, index internal psychometric properties of scale scores applicable with effect indicator scales when we are interested in sampling fluctuations resulting from the sampling of items [58]. Revelle [44] proposed an index labeled coefficient beta (β) that he showed equals the proportion of variance in scale scores accounted for by a general factor under more general conditions. McDonald's coefficient (ω) is computed as ratios of the variance due to the common attribute (i.e., factor) to the total variance [42].

The main difficulty in using these coefficients in the context of our study is that they are applied in single questionnaires and not in correlated questionnaires, as in our case. Such coefficients are not sufficient tools for the correlation between two psychometric questionnaires. Due to this fact, it was necessary to adapt Cronbach's α coefficient to our context as further explained in Subsection 4.2.2.

2.2 Related Work

We observe from literature that there are different ways of evaluating the chatbot in the context of humanization.

In the literature, we can see that there are ways and forms of evaluating the chatbot in the context of humanization. For Go and Sundar [17], the following features are relevant to the construction of dialogue humanity in a chatbot: social presence, homophily, contingency, dialogue, expertise, friendliness, and human resemblance. Their paper sought to investigate the influence of visual, identity, and behavioral factors on users' perception of humanization about the chatbot. Their study aimed to identify how certain changes in a specific chatbot influenced users' perceptions. However, a measurement of users' perceptions of humanness towards the chatbot is not performed.

As for Nordheim, Følstad and Bjørkli [36], the main aspects of the humanization of chatbots are expertise, including correct-concrete-response, interpretation, responsiveness, predictability, human-like, ease of use, and absence of marketing. Their study conducted an experimental protocol in two phases: exploratory and quantitative. The exploratory

analysis aimed to freely identify, in a non-induced way, which factors most influenced the users' perception of reliability regarding different types of chatbots. Their quantitative analysis induced research on which factors most influenced the perception of reliability, using the factors initially identified in the exploratory analysis. Ultimately, their research ponders each identified factor's influence in building users' perception of credibility regarding chatbots. Their study, and mainly the quantitative analysis performed, was the primary reference for the proposal of the weighting questionnaire of our proposed methodology (cf. Chapter 4).

In the study conducted by Følstad and Brandtzaeg [12], the aspects of the interactive system are separated into desirable and undesirable. Desirable aspects are help and assistance, information and updates, entertainment, novelty and inspiration, and human likeness. The undesirable aspects are interpretation issues, inability to help, repetitiveness, strange or rude responses, unwanted events, and boring attitudes. Their study sought to identify factors that lead users to have positive and negative experiences with chatbots in a qualitative analysis of such experiences. Users participating described their interactions using free text, which was later analyzed. The use of free text searches is too expensive, especially when the objective is the acquisition of a quantitative metric, which makes it impossible to use their methodology [12] in our proposal.

Westerman, Cross and Lindmark [51] mentioned only two aspects of humanization: humanity and social attraction. Similar to the study by Go and Sundar [17], the study by Westerman, Cross and Lindmark [51] sought to investigate the influence of typographical errors on the perception of humanity that users have regarding a particular chatbot. However, the chatbot was executed by a human being who simulated the behavior of a chatbot, without the users participating in the experiment knowing that it was a chat between human beings, which differs from our proposal, which seeks to evaluate real chatbots developed with computational resources. The factors used were relatively generic; therefore, the entire dimension of what perceived humanity was not evaluated.

Balaji [4] considered the following aspects in the humanization of chatbots as being relevant: initiating conversation, communication effort, content relevance, response clarity, reference to service, graceful breakdown, speed, privacy, accessibility, ease of starting a conversation, flexibility of linguistic input, communication quality, response quality, expectation setting, ability to maintain themed discussion, recognition, and facilitation of users' goal and intent, understandability, and credibility. Their work sought to map as many impact factors as possible in evaluating humanization in chatbots, taking into account user satisfaction regarding using a chatbot. To this end, several qualitative assessment questionnaires were analyzed and finally synthesized into a final questionnaire. Volunteers were asked whether or not each impact factor was relevant, along with a justification. Interestingly, descriptions of impact factors were included in the questionnaire, which makes the survey clearer for volunteers.

Kondylakis *et al.* [26] investigated usability assessment of a specific chatbot (the R2D2) using two generic questionnaires for evaluating the quality of software products: the ISO/IEC 25,000 series in conjunction with the System Usability Scale (SUS). Although such questionnaires are not specific for evaluating humanization in chatbots, they are adaptable for usability evaluation, in a way not adapted to the contexts of different types

of chatbots.

Holmes *et al.* [21] assessed how personalization affects usability metrics. Usability in the case was evaluated through three methodologies: SUS, User Experience Questionnaire (UEQ) and Chatbot Usability Questionnaire (CUQ). The CUQ is not adaptable to different contexts and does not use the orientation to impact factors. Despite the minor detail of the questionnaires used during the experiment of Holmes *et al.* [21], the metrics generated with the questionnaires' results were well explained.

2.3 Discussion

We discuss the need for a better-established protocol for evaluating humanization in chatbots. We did not find in literature a humanization assessment methodology focused on the acquisition of metrics. This is what our proposed investigation aims to innovate. Our goal is to obtain a method that better evaluates the evolutionary process of chatbot design and development. This might be an essential instrument for mapping chatbot's improvements within its specific contexts. Table 2.1 provides a comparison among the related studies from literature and our proposal. We considered five critical criteria: evaluation of humanization, generic context, adaptability, mapping of impact factors, generation of metrics.

The *evaluation of humanization* criterion considers whether users' perception of humanization regarding a chatbot is evaluated. Nordheim, Følstad and Bjørkli [36] focused on the perception of trust. Westerman, Cross and Lindmark [51] evaluated humanization in a very generic way, not considering the various characteristics that make up the perception of humanization. And both Kondylakis *et al.* [26] and Holmes *et al.* [21] assessed usability in general, not the perception of humanization.

The *generic context* criterion verifies whether the study was conducted analyzing a specific chatbot context or chatbots in general, with only Nordheim, Følstad and Bjørkli [36], Følstad and Brandtzaeg [12], and Balaji [4] analyzed chatbots generically. Completing the *generic context* criterion, the *adaptability* criterion examines whether studies can be adapted to different contexts, and only the study by Kondylakis *et al.* [26] cannot be adapted by focusing on the development of a specific chatbot.

The *mapping of impact factors* criterion shows which studies cited and described the humanization impact factors that were analyzed. The study by Westerman, Cross and Lindmark [51] superficially evaluates humanization, not addressing the factors that impact the perception of humanity that users have regarding the chatbot used. Meanwhile, studies by Kondylakis *et al.* [26] and Holmes *et al.* [21] use usability questionnaires adapted to the reality of chatbots, and the questions of such questionnaires are not mapped to humanization impact factors.

The last criterion we analyzed is the *generation of metrics*. The studies by Go and Sundar [17], Følstad and Brandtzaeg [12], and Balaji [4] did not aim to analyze humanization quantitatively, focusing only on qualitative analysis. The investigations of Kondylakis *et al.* [26], and Holmes *et al.* [21] generated quantitative results, but these results are not related to the humanization of the analyzed chatbots. The main value of

the quantitative results generated by Kondylakis *et al.* [26] and Holmes *et al.* [21] concerns the methodology for converting psychometric questionnaires into quantitative metrics.

Table 2.1: Comparison among Related Work

Study	**Evaluation***	**Generic Context**	**Adaptability**	**Mapping****	**Metrics *****
Go and Sundar [17]	✓	-	✓	✓	-
Nordheim *et al.* [36]	-	✓	✓	✓	✓
Følstad *et al.* [12]	✓	✓	✓	✓	-
Westerman *et al.* [51]	-	-	✓	-	✓
Balaji [4]	✓	✓	✓	✓	-
Kondylakis *et al.* [26]	-	-	-	-	✓
Holmes *et al.* [21]	-	-	✓	-	✓
Our Proposal	✓	✓	✓	✓	✓

* Evaluation of Humanization
** Mapping of Impact Factors
*** Generation of Metrics

2.4 Conclusion

We conclude that none of the analyzed studies cover our proposal developed in this MSc. dissertation. We found that none of them simultaneously evaluates humanization in generic contexts in an adaptive way. We explicitly map the impact factors generating a metric in humanization. Specifically, many of investigated studies generated metrics mainly based on the idea of usability; and it is inferred that usability is an internal concept to humanity. A humanized chatbot has high usability, but a chatbot with high usability is not necessarily humanized. The generation of metrics on humanization and the generation of a methodology for evaluating humanization is necessary. We observe that many of the methods for humanization assessment are content with qualitative assessment, not reaching quantitative assessment. We consider that qualitative assessment of humanity makes it very difficult to verify the evolution of the chatbots development process. Our proposed methodology aims to verify the evolution of the chatbot humanization in development process. This turns possible to verify that the chatbot is becoming further humanized. We investigate how we can extract sets of impact factors that are as generic as possible; that is, the broadest possible set of impact factors. On this basis, we study a methodology for evaluating humanization in chatbots that is as adaptive as possible to different contexts in which chatbots may be inserted.

Chapter 3

Principles of Humanization

Chatbots and platforms for developing them are increasingly common in the market. They aim to meet the growing demands for solving problems arising from current activities mainly related to customer services. Such software can perform standard and medium-complexity tasks, making the human workforce specialized in solving highly complex tasks [54]. Low and medium-complexity tasks are frequent in call centers, whereas high complexity tasks are time-consuming. This situation creates a bottleneck of activities that reduces the quality of services to the user. Chatbots assisting in answering tasks, even if only for ordinary and medium-complexity tasks, become powerful strategic tools [52]. An example of a low-complexity task is answering simple questions, such as questions on FAQs. Examples of medium-complexity tasks are those involving database manipulation, which requires a greater level of integration and customization [52]. Already examples of highly complex tasks involve general maintenance, mainly system maintenance, in which the intervention of specialized human agents is necessary [52, 53].

All described categories of chatbots have a common goal: appear more and more human, maintaining or increasing the resolution ability of possible tasks. The task of humanizing a chatbot, which is to make its behavior similar to that of a human being, is not trivial. Human behavior is complex and varies according to the nature of situations. Each person has his/her perception of humanity. We know how to differentiate typical machine behavior from human behavior, but explaining such differences is a hard task. Users tend to maintain a better dialogue with human attendants instead of chatbots.

This situation justifies increasing the humanization of chatbots to improve users' satisfaction with such systems [17]. When it comes to chatbots, the perception of humanity must be linked to the resolution of the proposed tasks. In this context, a chatbot must satisfy users' expectations for its respective purpose [17]. Some characteristics are generally expected of a chatbot, such as naturalness. A chatbot is expected to build meaningful, personalized, and friendly dialogues through a conversational system. Such a conversational system must evolve with each interaction and offer alternative answers to similar questions, answering the interlocutor's questions as another human person would meet user expectations more naturally and fluidly. The chatbot must be able to relate cordially with the consumer, making them feel comfortable, satisfied, and valued [4].

From this perspective, humanization impact factors aspects regarding chatbots are essential. This Chapter provides an original study to collect and organized a set of

humanization impact factors. We assess their manifestation in an adhoc analysis of market chatbots. We obtained two main results: the broad set of impact factors in humanization; and the analysis of three chatbots presented in the market based on this broad set of identified impact factors.

3.1 Methodology

We aim to identify impact factors related to humanization in chatbots via a bibliographic analysis. We call impact factors those characteristics that contribute to users' perception of humanity regarding a chatbot. Our first task refers to the compilation of a complete set of humanization impact factors possible. Our conducted methodology is presented in subsection 3.1.1. Section 3.3 presents the obtained impact factors.

To demonstrate how broad impact factors can be used, we performed an analysis of which of the 36 resulting impact factors in the evaluation of humanization in three chatbots present in the market (health, retail, and education sectors). For such identification of the perception impact factors, the purpose of the chatbot, the field of activity, and the target audience were taken into account. Subsection 3.1.2 presents the result analysis.

3.1.1 Identifying Impact Factors on Humanization

We observed that existing investigations aiming to study humanization in chatbots selected different sets of characteristics to be analyzed in chatbots (impact factors). There is no simple convergence between these sets of impact factors. One of the possible reasons is that each work studies or focuses on specific types of chatbots. Our selected set of characteristics corresponds to what is highly desired. Therefore, the objective is to carry out a complete compilation of these sets of attributes and generate a more extensive and complete set of impact factors.

Figure 3.1 presents the first step of our study (defined as step A) as exploratory bibliographic research, using platforms for academic research such as Google Scholar, Scopus, IEEEXplore, and Web of Science. Some examples of key terms used in this search were: "chatbots"; "humanization assessment"; "chatbot assessment methodologies"; "qualitative assessments of chatbots"; and "humanized chatbots". The selected studies must necessarily deal with chatbots and were filtered according to the year (preferably selecting the most up-to-date ones) and the type of research carried out (selecting qualitative analysis works). We believe the selected studies should discuss aspects of humanization, humanization evaluation, or usability. Such aspects are analyzed qualitatively but not always in a psychometric way. The preference was for works that performed qualitative analysis using psychometric methods for evaluation. However, as such works are rare, the simple indication of which qualitative aspects can be evaluated has already helped construct a more robust set of factors that impacted users' perception of humanity about chatbots. We seek to answer six main questions:

- What is humanizing a chatbot?
- How to humanize a chatbot?

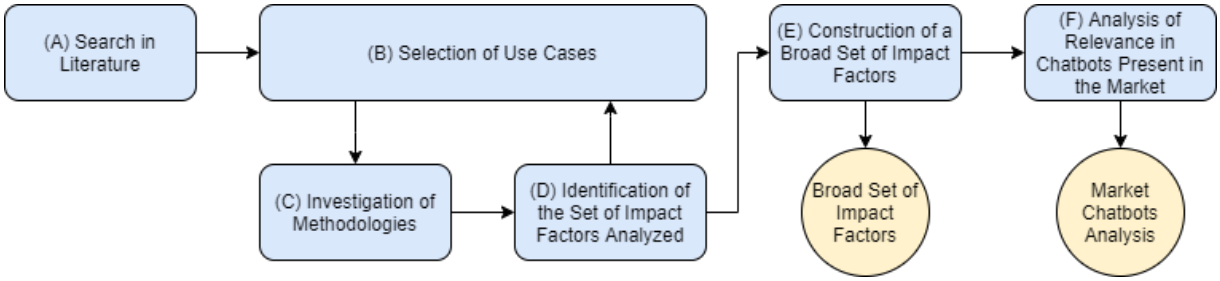

Figure 3.1: Methodology for Identifying Impact Factors on Humanization

- How to assess that a chatbot is humanized?
- What should be evaluated in the humanization of chatbots?
- What aspects are qualitatively analyzed and evaluated in chatbots?
- What factors affect users' perception of humanity towards a chatbot?

In step B of our methodology (cf. Figure 3.1), cases involving the specific evaluation of humanization in chatbots were carried out. In this search, we ruled out cases in which usability assessments were used to assess the humanization of chatbots, as is the case of studies by Kondylakis *et al.* [26] and Holmes *et al.* [21]. Here, we need to consider how the analysis of the usability of a chatbot is included in the idea of humanization since a humanized chatbot necessarily has high usability. However, a chatbot with high usability is not necessarily humanized.

We verified that Go and Sundar [17]; Nordheim, Følstad, and Bjørkli [36]; Følstad and Brandtzaeg [12]; Balaji [4] meet the criteria of specifically evaluating humanization in chatbots. The concept of humanization is broader and more complex than usability since humanity includes all the characteristics that differentiate men from other beings and things.

After selecting the four use cases (step B), to select the humanization impact factors that will consider in our study, that is, which qualitative aspects will jointly evaluate. The methodology used for evaluating humanity in chatbots was analyzed in each case. The context in which such methods were applied (step C in Figure 3.1). This step aims to assess how generic or adaptable the cases are in the four selected studies, seeking the reproducibility of the studies.

At step D, the set of characteristics were verified - evaluated by each methodology (step D), it is at this step that impact factors are actually identified and compiled into a single qualitative set. It is already demonstrated that the groups of factors found to vary significantly from context to chatbot context and therefore differ from study to study.

In step E (cf. Figure 3.1), the sets of characteristics evaluated by each case are compiled into a single, broader set of characteristics. Here, we already called these characteristics impact factors in humanization. Those that have a very similar meaning are united in a single impact factor, although the terminology may be different. Therefore, we already have the broadest set of impact factors identified in the four selected articles at this stage. However, there may still be duplicates and impact factors that may or may not necessarily have the same names but have descriptions and meanings close enough to be gathered in a single concept. The coexistence of impact factors explains the need for this phase in more

than one article. We got our first result (presented in circle 1 in Figure 3.1), a set of 36 impact factors and their descriptions according to the cases in which they were analyzed.

3.1.2 Market Chatbots Analysis

Based on the results achieved during the application of the methodology to Identify Impact Factors in Humanization (Subsection 3.1.1), and as shown in step (F) of Figure 3.1, we carried out the first experiment to analyze the perception that users had about each of the impact factors in three chatbots on the market. This experiment took place in the context of the CI&T company, using chatbots suggested by the company and with wide dissemination to employees. We applied the methodology presented in Figure 3.2 to the employees of the CI&T company who volunteered to participate in the experiment. Next, we will explain each of the steps of the methodology of this experiment.

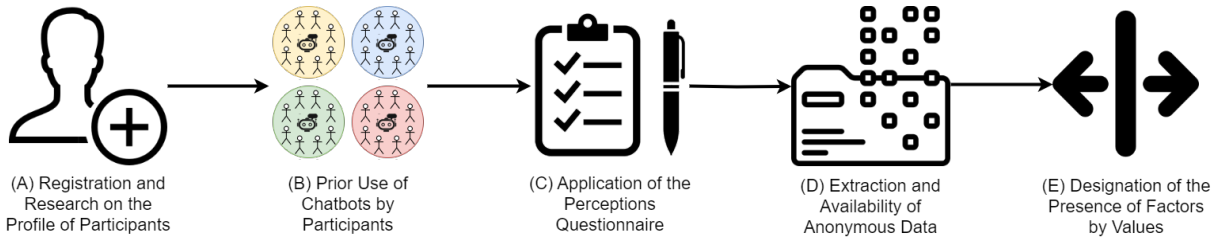

Figure 3.2: Methodology for Market Chatbots Analysis

(A) Registration and Research on the Profile of Participants

The objective of this research was to know better the public that participated in the analysis of chatbots present in the market, being this public composed exclusively of employees of the CI&T company, who volunteered to participate in the experiment proposed in this subsection.

After completing this survey, instructions for participating in the experiment were sent to each participant's email. All recorded data were protected and not shared with people not involved in the research. Below we present each of the questions, their objectives and their possible answers.

1. **Have you ever interacted with a chatbot?**
   - **Objective:** to know if the participant understands the basics of chatbots.
   - **Answers:** Yes or No
2. **Do you have ease with technological resources?**
   - **Objective:** to know if the participant is a person who has difficulties with technological resources in general, since chatbots are technological resources.
   - **Answers:** Yes or No
3. **Are you a person who likes technology?**
   - **Objective:** to know if the participant is a person potentially fearful of technology, or averse to technology, seeing all negative technology.

- **Answers:** Yes or No

4. **Do you have knowledge in any programming language?**
   - **Objective:** Programming knowledge puts the participant at an average level of knowledge of the limitations of chatbots.
   - **Answers:** Yes or No
5. **Have you ever participated in the development of any software?**
   - **Objective:** Participants who have already participated in a software development process have a medium to high level of knowledge of the limitations of chatbots.
   - **Answers:** Yes or No
6. **Have you ever participated in the development of a chatbot?**
   - **Objective:** Participants who have already participated in a software development process have a high level of knowledge of the limitations of chatbots.
   - **Answers:** Yes or No
7. **Position / Profession**
   - **Objective:** To verify if the position or profession is directly linked with technology, more specifically with chatbot design. It is also important to understand which types of professionals are interested in the topic, so that the questionnaire is as aggregating as possible.
   - **Answers:** Open
8. **Training Level**
   - **Objective:** Another way to verify the level of familiarity with the proposed theme, as well as an opportunity to adapt the language of the questionnaire for the best possible understanding of all those present.
   - **Answers:** Incomplete Elementary School, Complete Elementary School, Incomplete High School, Complete High School, Incomplete Higher Education, Complete Higher Education, Incomplete Specialization, Complete Specialization, Incomplete Master's, Complete Master's, Incomplete Doctorate or Complete Doctorate

(B) Prior Use of Chatbots by Participants

For the selection of chatbots, we used recommendations from the company CI&T, which sent a spreadsheet with 100 chatbot options. Initially, we chose six chatbots with the criteria of belonging to different segments, being active, and being available on Whatsapp, a viral, simple, and widely used conversation platform in Brazil. Below we present Table 3.1, where we present the six selected chatbots, along with the respective segments and numbers for access via Whatsapp.

After the first selection of the six chatbots, we agreed with part of the CI&T team to run the experiment with only three to acquire more data per analyzed chatbot. The CI&T team chose the three chatbots, and we present them below.

Table 3.1: First Selected Chatbots

Segment	Name	Whatsapp Number
Health	Fleury Medicine and Health	(11) 3179 0822
Education	MeBote na Conversa	(11) 93456 5026
Entertainment	ZapFlix (Netflix)	(11) 99653 5902
Foods	McDonald's	(11) 3230 3223
Retail	Helô (Riachuelo)	800 772 3555
Services	ConectCar	(11) 3003 4475

Chatbot A:

- **Name:** Fleury Medicina e Saúde
- **Tracking:** Health
- **Message:** Fleury: health and well-being for the full realization of people.
- **About:** Fleury Medicina e Saúde is one of the most respected medical and health organizations in the country, recognized by the medical community and public opinion as excellence in quality, innovation and customer service. Learn more about at link.
- **Whatsapp Number:** (11) 3179 0822
- **Whatsapp Link**
- **Usage Video Link**
- Tasks:
    1. Search for tests for COVID-19 and influenza (by numerical menu).
    2. Know the addresses and opening hours of the units (by text).
    3. Search for coverage of agreement or plan (by text).
    4. Know about budgets (by numerical menu).
    5. Know how to be a Fleury supplier (by text).
    6. Talk about other subjects (by numeric menu).

Chatbot B:

- **Name:** Helô da Riachuelo
- **Tracking:** Retail
- **Message:** Hey! Welcome to Riachuelo's WhatsApp!
- **About:** Helô is the persona of digital service and the name of the brand's chatbot and now appears to the public. She is responsible for helping the retailer's customers on various topics, and also forwards to sales via WhatsApp. Helô knows everything about digital and social networks. She is a columnist for Blog Riachuelo and an expert in giving online shopping and fashion tips. To learn more about Riachuelo, access link.
- **Whatsapp Number:** 800 772 3555
- **Whatsapp Link**
- **Usage Video Link**
- Tasks:
    1. Find stores near your address (by numeric menu).

2. Learn more about the RCHLO App (by text).
3. Ask about exchanges and returns in physical stores (by numerical menu).
4. Learn more about exchange vouchers in physical stores (by text).
5. Request an exclusive Riachuelo card, but do not download the application (by numerical menu).
6. Trying to understand about the Riachuelo Cards (by text).

Chatbot C:

- **Name:** MeBote na Conversa
- **Tracking:** Education
- **Message:** A tool to help you understand all the terms and acronyms used in the corporate world.
- **About:** Meta — the new name of Facebook Inc., owner of WhatsApp — joined the Indique Uma Preta collective and the MOOC agency to create the Me Bote Na Conversa project, a bot for WhatsApp that explains English expressions used in the corporate world. The idea is to make communication in a more inclusive work environment. Learn more about MeBote in Conversation on link.
- **Whatsapp Number:** (11) 93456 5026
- **Whatsapp Link**
- **Usage Video Link**
- Tasks:

  1. Ask what a Brainstorm is.
  2. Ask what a Deadline is.
  3. Ask what a Briefing is.
  4. Ask what a Workshop is.
  5. Ask what a Hands-On is.
  6. Ask what an MVP is.
  7. Ask what a Freelancer is.

We randomly selected one of the selected chatbots for each participant who volunteered and responded to the Profile Survey (step A). We sent an email to each of the participants indicating the link to access the chatbot, what is the purpose of such a chatbot, a video of previous use of the chatbot under analysis, and a list of suggested tasks to be performed in the chatbot, being the participant free to use the chatbot as they wish.

(C) Application of the Perceptions Questionnaire

The perceptions questionnaire aims to assess how much users perceived the presence of each of the impact factors in the chatbot used. Along with the weighting questionnaire, the perceptions questionnaire verifies the level at which characteristics considered positive or negative are perceived in the chatbot under analysis. Such positive and negative characteristics and perceptions are variable according to the context in which they are analyzed.

For the development of this questionnaire, statements are generated about the user's perception of each of the impact factors in the analyzed chatbot. The objective is to

identify a level of agreement with the proposed statement. The possible answers follow one of the Likert scale patterns, ranging from "strongly disagree" to "strongly agree". The following is an example of claim generation:

Impact Factor: Appropriate Language Style
Affirmation: The chatbot has an appropriate language style.
Impact Factor Explanation: Ability of the chatbot to use the appropriate language style for the context.
Alternatives: strongly disagree, disagree, somewhat disagree, neither agree nor disagree, somewhat agree, agree, strongly agree.

After using the chatbot selected for each participant, as presented in step B, the participant was instructed to access the following link to access the Perceptions Questionnaire hosted on the Google Forms tool and answer it according to their interaction with the chatbot. The Perceptions Questionnaire has a brief description of the research objectives, an explanation of the Likert scale, and fields to fill in the participant's e-mail and to indicate the chatbot previously used, in addition to the 36 questions related to each impact factor identified in Subsection 3.1.1. The collection of e-mails was requested by the company CI&T to monitor the employees who collaborated with the research.

(D) Extraction and Availability of Anonymous Data

In this methodology step, we anonymized the data collected with the Perceptions Questionnaire (in step C), erasing the only sensitive data collected: the participants' e-mail. After anonymization, the results were made available through a Google Spreadsheet at the following link that only reads the data, allowing the free reproduction and conference of the calculations performed in the next step.

(E) Designation of the Presence of Factors by Values

The objective of this last step of the methodology is to define, based on the anonymized responses acquired in step D, whether or not a factor is present in each of the analyzed chatbots. The first step was to calculate the arithmetic mean of the values of the responses obtained on each chatbot, and the possible perceptions indicated in step C were converted to psychometric values following the rules in Figure 3.3.

As we see in Figure 3.3, an impact factor is only perceived at some level in the analyzed chatbot if it has a score above 4, and it is not perceived if it has a score equal to or below four. In this way, we evaluate the presence or absence of each impact factor in each chatbot according to the rule below:

- **PRESENT:** If the average value of the impact factor is above 4.0.
- **ABSENT:** If the average value of the impact factor is equal to or below 4.0.

The results of applying the five steps of this methodology are available in Section 3.3.

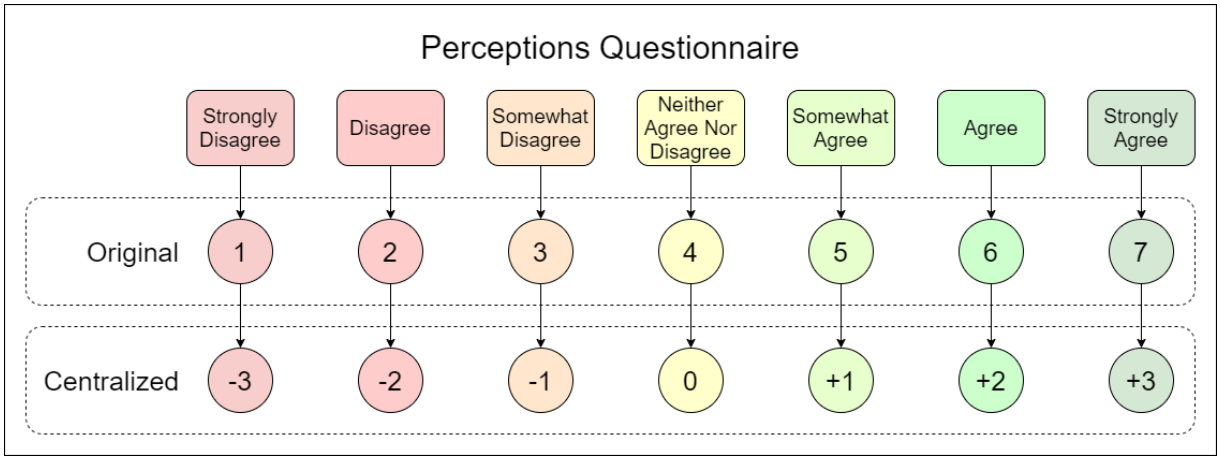

Figure 3.3: Rule for Converting Perceptions into Psychometric Values

3.2 Results of the Identified Impact Factors

We present the 36 impact factors obtained from our literature analysis.

- **Social presence** [17][12]: Social presence is formally defined as the degree of salience of the other person in the interaction. It is also defined as the feeling of being with another in a mediated environment.
- **Homophily** [17]: Perceived homophily is defined as the amount of similarity two people perceive themselves as having, since human-like figure of the agent is likely to be perceived as being more similar to the user than a bubble figure.
- **Contingency** [17]: Perceived contingency occurs when the conversation is considered fully interactive or responsive, individuals will perceive greater levels of dialogue emerging from threaded message exchanges with a chat agent during computer-mediated communication.
- **Dialogue** [17]: Dialogue perception may make online interactions feel like face-to-face conversations, in turn creating positive attitudes toward the agent.
- **Expertise** [17][36]: Perceived expertise occurs when the participants perceived the agent as intelligent, knowledgeable, competent, and informed.
- **Friendliness** [17]: Perceived friendliness occurs when the participants perceived the agent as empathetic, personal, warm, emotionally invested, willing to listen, careful, and open.
- **Human-likeness** [17][36][12]: Human-likeness suggest the presence of a human figure attached to a chat agent, that is, a “other person” in the interaction.
- **Predictability** [36]: Users’ perceptions of the consistency with which the interactive system behaves.
- **Ease of use** [36][4]: The ease or simplicity with which the interaction with the system is accomplished.
- **Absence of marketing** [36]: Absence of marketing, and a sense of the chatbot putting the customer first.
- **Help and assistance** [12]: The chatbot is reported to provide customer support or training, personal assistance, or help with a particular task at hand. Efficiency and

ease of access were often highlighted.

- **Information and updates** [12]: The chatbot is reported to provide updates and general information - often sought on a routine basis - such as news, weather and online searches.
- **Entertainment** [12]: The chatbot interaction is described in words reflecting engagement and enjoyment.
- **Novelty and inspiration** [12]: The novelty of chatbots, or the inspirational value of the interaction, is accentuated.
- **Absence of interpretation issues** [12]: The chatbot must avoid interpretation issues, which occur when the chatbot is reported to misinterpret requests or input or provide an answer that does not fit the question.
- **Absence of inability to help** [12]: The chatbot must avoid inability to help, which occurs when the chatbot is reported to be unable to assist the participant in solving a particular task or to be unable to provide help in general.
- **Absence of repetitiveness** [12]: The chatbot must avoid repetitiveness, which occurs when the chatbot is reported to ask the same questions or repeatedly provide the same line of answers, which is experienced as obstructing the user from getting help or assistance.
- **Absence of strange or rude responses** [12]: The chatbot must avoid strange or rude responses, which occurs when the chatbot is reported to give improper or embarrassing responses.
- **Absence of unwanted events** [12]: The chatbot must avoid unwanted events, which occur when the chatbot is reported as the source of unwanted contact, actions, or content.
- **Absence of boring attitudes** [12]: The chatbot must avoid boring attitudes, which occurs when the chatbot interaction is reported to be boring - either immediately or after a period of use.
- **Initiating conversation** [4]: How easy it is for the user to start interacting with the chatbot, including not only accessibility but also how simple it feels to actually start the conversation i.e. to start typing.
- **Communication effort** [4]: How easy it is for the user to successfully (or not) convey his or her information-retrieval goal to the chatbot.
- **Content relevance** [4]: The extent to which the chatbot's response addresses the user's request.
- **Response clarity** [4]: How easy it is for the chatbot's response to be understood by the user.
- **Reference to service** [4]: The ability of the chatbot to provide useful and relevant hyperlinks or automatic transitions either in lieu of or in addition to its response to the user's request.
- **Graceful breakdown** [4]: The appropriateness of the manner in which the chatbot responds if and when it encounters a situation in which it cannot help the user.
- **Speed** [4]: How quickly the chatbot responds to each input the user gives.
- **Privacy** [4]: How secure the entire interaction feels as a consequence of revealing potentially personal information to the chatbot.

- **Flexibility of linguistic input** [4]: How easily the chatbot understands the user's input.
- **Communication quality** [4]: How easy it is for the user to communicate his or her information-retrieval goal.
- **Response quality** [4]: The overall quality of the chatbot's response once the user has provided some form of input to the chatbot.
- **Expectation setting** [4]: The extent to which the chatbot sets expectations for the interaction with an emphasis on what it can and cannot do.
- **Ability to maintain themed discussion** [4]: The ability of the chatbot to maintain a conversational theme once introduced and keep track of context.
- **Recognition and facilitation of users' goal and intent** [4]: Ability of the chatbot to understand the goal and intention of the user and to help them accomplish these.
- **Understandability** [4]: Ability of the chatbot to communicate clearly and is easily understandable.
- **Credibility** [4]: The extent to which the user believes the chatbot's responses to be correct and reliable.

Figure 3.4 shows the intersections between the sets of humanization impact factors analyzed by different studies. The studies of Go and Sundar [17]; Følstad and Brandtzaeg [12]; and Nordheim, Følstad and Bjørkli [36] have some impact factors in common, whereas the work of Balaji [4] has impact factors in common only with Nordheim, Følstad and Bjørkli [36]. We pooled all these impact factors by removing duplicates in our study covering the four studies cited above.

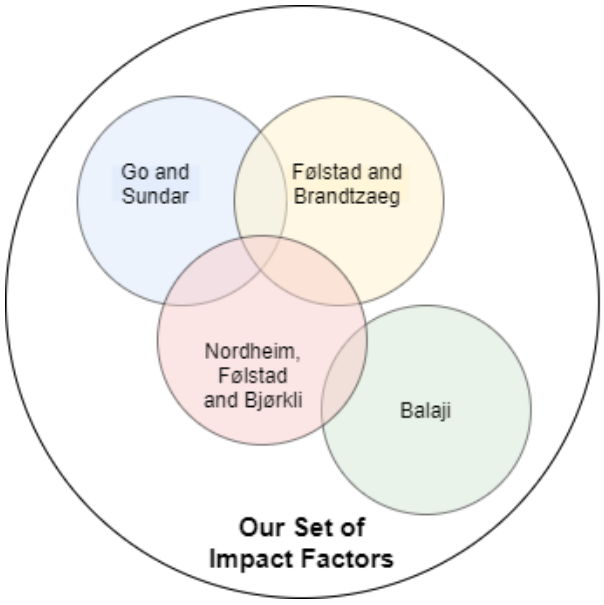

Figure 3.4: Intersection Between the Sets of Impact Factors

Table 3.2 shows more specifically in which studies each impact factor is present. Table 3.2 helps to understand Figure 3.4, showing which impact factors are unique to particular articles (such as the Homophily impact factor in Go and Sundar [17]) and which coexist in more than one work (such as the impact factor Social Presence which is studied both in Go and Sundar [17] and in Følstad and Brandtzaeg [12]).

Table 3.2: Presence of Impact Factors in Reference Articles

Impact Factors	[17]	[36]	[12]	[4]	**Our Set**
Social presence	✓	-	✓	-	✓
Homophily	✓	-	-	-	✓
Contingency	✓	-	-	-	✓
Dialogue	✓	-	-	-	✓
Expertise	✓	✓	-	-	✓
Friendliness	✓	-	-	-	✓
Human-likeness	✓	✓	✓	-	✓
Predictability	-	✓	-	-	✓
Ease of use	-	✓	-	✓	✓
Absence of marketing	-	✓	-	-	✓
Help and assistance	-	-	✓	-	✓
Information and updates	-	-	✓	-	✓
Entertainment	-	-	✓	-	✓
Novelty and inspiration	-	-	✓	-	✓
Absence of interpretation issues	-	-	✓	-	✓
Absence of inability to help	-	-	✓	-	✓
Absence of repetitiveness	-	-	✓	-	✓
Absence of strange or rude responses	-	-	✓	-	✓
Absence of unwanted events	-	-	✓	-	✓
Absence of boring attitudes	-	-	✓	-	✓
Initiating conversation	-	-	-	✓	✓
Communication effort	-	-	-	✓	✓
Content relevance	-	-	-	✓	✓
Response clarity	-	-	-	✓	✓
Reference to service	-	-	-	✓	✓
Graceful breakdown	-	-	-	✓	✓
Speed	-	-	-	✓	✓
Privacy	-	-	-	✓	✓
Flexibility of linguistic input	-	-	-	✓	✓
Communication quality	-	-	-	✓	✓
Response quality	-	-	-	✓	✓
Expectation setting	-	-	-	✓	✓
Ability to maintain themed discussion	-	-	-	✓	✓
Recognition and facilitation of users' goal and intent	-	-	-	✓	✓
Understandability	-	-	-	✓	✓
Credibility	-	-	-	✓	✓

3.3 Results of the Market Chatbots Analysis

Table 3.3 presents the results regarding how the collected impact factors are manifested in the analyzed chatbots, according to the methodology presented in Subsection 3.1.2.

3.4 Discussion

3.4.1 Market Chatbots Analysis

We observe from Table 3.3 that some impact factors are desirable in the three chatbots analyzed, such as homophily, help and assistance, and absence of strange and rude responses. This means that some of the impact factors are always desirable in chatbots. However, some impact factors are desirable in only one of the three chatbots, such as the dialog impact factor, which is only desirable in chatbot C, and the speed impact factor, which is only desirable in chatbot A.

We assume that certain groups of impact factors are unanimously desirable in any chatbot and others are not. For example, some impact factors are unnecessary for any chatbot, such as entertainment. This means that in the chatbots that were evaluated, in this case, entertainment was not necessary; but it does not mean that it will never be necessary for any chatbot. We can, for example, imagine chatbots with a humorous purpose, in which entertainment will be highly relevant.

The main aspect we need to verify is that the need to assess or not the impact factor in a chatbot does not add to every problem. Another question is how much is necessary to evaluate an impact factor. In other words, how desirable is an impact factor for the chatbot being analyzed, or how undesirable is the impact factor for the chatbot being analyzed. In the examples given, from the analysis performed on chatbots A, B, and C, entertainment did not appear as relevant in any of them. Nevertheless, this could mean that entertainment is highly undesirable for some of them, and how can we represent that?

We observe that the relevance of an impact factor is not something binary but something continuous. How necessary an impact factor is should be a continuous value. After all, one impact factor may be more relevant, but both are relevant, with one more relevant than the other. Social presence may be relevant to chatbots B and C. However, it may be more critical to one of these chatbots than the other, so a lasting value for relevance becomes necessary. The use of real values allows ranking both the level of perception that users had regarding each impact factor and ranking the relevance of each impact factor in the conception of the chatbot's idea of humanity, which is not possible using only binary values.

3.4.2 Limitations

The main limitation of this investigation of the impact factors in humanization is that there is no way to define that the set of impact factors we currently identify is static. It will likely be modified over time according to the new needs of chatbots in the future. Chatbots will be applied in new contexts, addressing new characteristics that determine whether a

Table 3.3: Analysis of Impact Factors in Different Types of Chatbots

Impact Factors	Chatbot A		Chatbot B		Chatbot C	
	Scores*	P**	Scores*	P**	Scores*	P**
Social Presence	5,0	✓	6,0	✓	4,3	✓
Homophily	5,0	✓	4,7	✓	3,7	-
Contingency	4,0	-	4,3	✓	3,3	-
Dialogue	4,0	-	5,3	✓	4,7	✓
Expertise	4,0	-	5,3	✓	2,3	-
Friendliness	5,0	✓	5,7	✓	4,0	-
Human-Likeness	4,0	-	3,3	-	3,0	-
Predictability	5,0	✓	5,7	✓	4,3	✓
Ease of Use	5,0	✓	5,7	✓	3,7	-
Absence of Marketing	5,0	✓	6,0	✓	3,0	-
Help and Assistance	5,0	✓	3,7	-	3,0	-
Information and Updates	6,0	✓	5,7	✓	3,3	-
Entertainment	5,0	✓	6,0	✓	3,7	-
Novelty and Inspiration	5,0	✓	5,3	✓	3,3	-
Absence of Interpretation Issues	4,0	-	5,0	✓	2,7	-
Absence of Inability to Help	5,0	✓	4,7	✓	2,0	-
Absence of Repetitiveness	4,0	-	3,3	-	3,0	-
Absence of Strange or Rude Responses	6,0	✓	6,7	✓	5,3	✓
Absence of Unwanted Events	6,0	✓	5,7	✓	4,7	✓
Absence of Boring Attitudes	6,0	✓	4,3	✓	4,7	✓
Initiating Conversation	6,0	✓	4,0	-	4,3	✓
Communication Effort	5,0	✓	5,0	✓	3,3	-
Content Relevance	5,0	✓	6,3	✓	3,7	-
Response Clarity	5,0	✓	5,7	✓	5,0	✓
Reference to Service	5,0	✓	5,7	✓	3,0	-
Graceful Breakdown	4,0	-	4,0	-	5,3	✓
Speed	3,0	-	7,0	✓	6,0	✓
Privacy	6,0	✓	4,7	✓	4,3	✓
Flexibility of Linguistic Input	3,0	-	3,7	-	3,3	-
Communication Quality	4,0	-	4,3	✓	3,7	-
Response Quality	6,0	✓	5,7	✓	4,0	-
Expectation Setting	4,0	-	6,0	✓	4,0	-
Ability to Maintain Themed Discussion	4,0	-	4,7	✓	2,0	-
Recognition and Facilitation of Users' Goal and Intent	3,0	-	2,7	-	3,7	-
Understandability	5,0	✓	5,7	✓	4,3	✓
Credibility	6,0	✓	6,3	✓	3,0	-

* Average Score of the Perception Evaluation
** Presence

chatbot is humanized. In addition, some studies on the humanization of chatbots were evaluated, which are limited, taking into account some specific chatbots as models.

3.4.3 Threats to Validity

The main threat to the validity of this study on humanization principles is the limited amount of studies on humanization and chatbots. Specifically, there is also the threat of the constant use of usability evaluation methods for humanization evaluation cases. It is seen that humanization is a more abstract and broader concept than the concept of usability; and usability is a concept that is contained in humanization. By definition, every humanized chatbot must have high usability, but not necessarily a high usability chatbot is humanized.

In this sense, further study regarding humanization in chatbots is necessary to establish a more adequate and generic set of impact factors on humanization in the specific case we are analyzing chatbots. We can also evaluate the sense of humanization of human-computer interfaces as a whole.

Another threat to validity is that the concept of humanization and humanity are much studied in philosophy and psychology. Hence, articles and studies from psychology and philosophy need to be aggregated in current work so that we can assess what humanity is more generally.

3.5 Conclusion

The concept of humanity is abstract and cannot be described as a simple and specific characteristic inherent to human beings; but, as a set of characteristics that makes humans different from other beings, whether animals or objects. In this sense, the great challenge was to find which characteristics are the key to describing what humanity is in the specific case we are working on (chatbots). It was also a great challenge to discover which characteristics are the key to defining what humanity is in the HCI context.

We were able to acquire an initial set of humanization impact factors, that is, those characteristics that influence the view that users have about the chatbot's humanity. This set is not fixed and is subject to change over time and according to with the new contexts in which chatbots will be used. In our study regarding the evaluation of chatbots from the market, we observed how the simple binary assessment of the relevance of impact factors in the analysis of chatbots is not enough to reveal how relevant each of these impact factors is.

We conclude that an assessment with continuous values is necessary to establish the context. To this end, it is necessary to use psychometric scales, such as the Likert scale. In the next Chapter, our study proposes the generation of a methodology for evaluating humanization in chatbots, capable of continuously establishing the relevance of each of the impact factors within different chatbot contexts (*i.e.*, capable of continuously defining how desirable or undesirable each of the impact factors is within a chatbot context).

Chapter 4

Chatbots Humanization Assessment Methodology

This investigation proposes to develop and apply a method to assess the level of humanization in chatbots, using the perceptions of real users. The method must adapt to different types of chatbots, with different purposes and target audiences, through a dynamic weighting system for the different humanization impact factors present in chatbots.

We assume that metrics generated through the application of the developed method can portray the evolution of aspects (factors) related to the humanization of chatbots. Figure 4.1 presents our proposed method, as well as its development and application. Our proposal goes through three major phases: bibliographic analysis, the methodology's development, and the evaluation method's application.

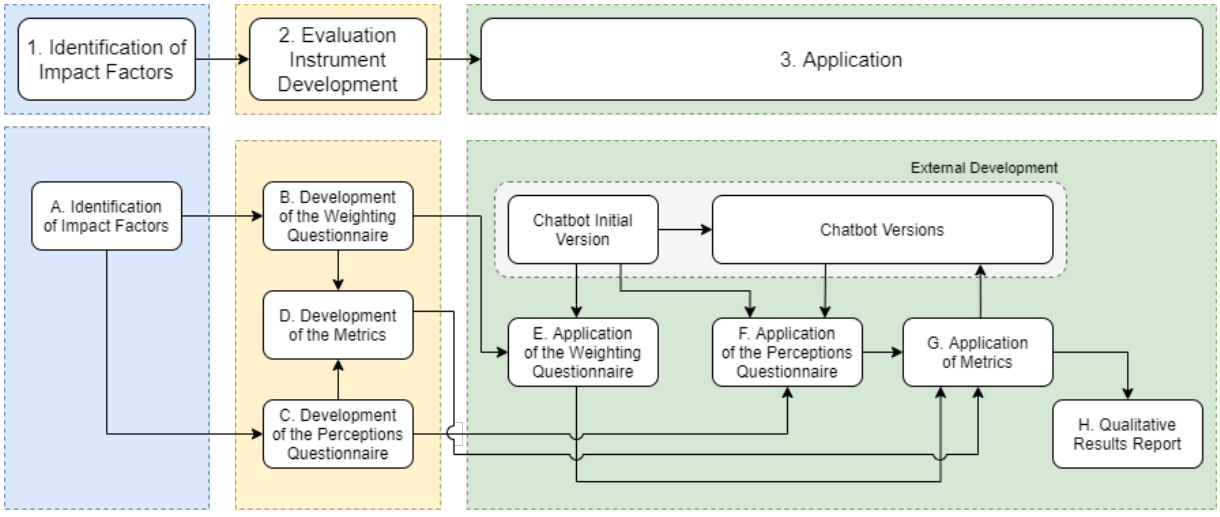

Figure 4.1: Method for Evaluating Humanization in Chatbots

4.1 Identification of Impact Factors

We aim to identify impact factors related to humanization in chatbots via literature analysis. We call impact factors those characteristics that contribute to users' perception of humanity regarding a chatbot. Therefore, our first task refereed to the compilation of a

complete set of humanization impact factors possible (cf. Section 3.2 presents the impact factors identified in the literature).

We observed that existing investigations aiming to study humanization in chatbots select different sets of characteristics to be analyzed in chatbots (impact factors). There is no simple convergence between these sets of impact factors. One of the possible reasons is that each work studies or focuses on specific types of chatbots (cf. Chapter 3). The selected set of characteristics corresponds to what is highly desired. Therefore, the objective at this stage was to carry out a complete compilation of these sets of characteristics and generate a more extensive and more complete set of impact factors (we developed our set in Chapter 3). In our proposal, we aim to weigh them according to the needs of each chatbot.

Finally, duplicates were removed, in the case of characteristics that have the same or very close meaning (cf. Section 3.2 presents the impact factors identified in literature).

4.2 Evaluation Instrument Development

We propose the development of two questionnaires with different objectives: the weighting questionnaire and the perception questionnaire. These questionnaires intend, respectively, to capture how relevant each impact factor is for the pleasant functioning of the chatbot; and how present they are in the analyzed system.

We present the development of the Perceptions Questionnaire in Chapter 3, specifically in Topic 3.1.2, internal to Subsection 3.1.2 on Market Chatbots Analysis. In Subsection 4.3.2 of this chapter, we will delve into the recommendations for applying the Perceptions Questionnaire.

In addition to the Perceptions Questionnaire, we will present two more assessment tools in this section: The Development of the Weighting Questionnaire (In Subsection 4.2.1) and the Generation of Metrics (In Subsection 4.2.2).

4.2.1 Development of the Weighting Questionnaire

The purpose of the weighting questionnaire is to determine which impact factors are the most important. This questionnaire aims to address which are less critical in a chatbot. Likewise, this questionnaire determines which factors are desirable and which are undesirable in the behavior of a chatbot. In this sense, the purpose of this questionnaire is to adapt the evaluation method to the most diversified realities of existing chatbots.

For the development of this questionnaire, questions are generated about how relevant each of the impact factors is presented in the set of characteristics generated in the Identification of Impact Factors phase. The possible answers follow one of the Likert scale patterns, ranging from “totally irrelevant” to “totally relevant”. The following is an example of question generation:

Impact Factor: Appropriate Language Style
Question: How important is the appropriate language style in the chatbot used?
Impact Factor Explanation: Ability of the chatbot to use the appropriate language

style for the context.
Alternatives: totally irrelevant, irrelevant, a little irrelevant, indifferent, a little relevant, relevant, totally relevant.

4.2.2 Generation of Metrics

We propose a metric to assess the level of humanization of a chatbot. Such metric is obtained with the combination of the two questionnaires applied because the perception of humanity is given by combining all impact factors. This investigation develops two types of metrics: a general metric of the level of humanization; and a performance metric for each of the analyzed impact factors.

The general metric in humanization aims to define how close to adequate humanization a chatbot is, considering the chatbot's context. The metric varies between 0 and 1, with a value of 0 representing that the chatbot was not humanized or that the humanization was performed inappropriately for its context. On the other hand, the value 1 represents a perfectly humanized chatbot with an adequate humanization for its purposes. The value of the proposed metric reflects how close the analyzed chatbot is to optimal humanization.

The general metric in humanization is generated from the results of the initial application of the weighting questionnaire together with the results of the perceptions questionnaire. Through a calculation, which is an adaptation of the formula for Cronbach's α coefficient, a value between 0 and 1 is returned.

The first step in acquiring this metric is pre-processing the data acquired with the questionnaires. Likert scales of size seven will be used in the questionnaires, and all responses collected on such scales will be converted to numerical values, (cf. Figure 4.2). After extracting the data acquired through the questionnaires, data are centralized and scaled using equation 4.1 and equation 4.2, for weighting and perception data, respectively (Figure 4.2 also shows the transformation of data with centralization).

In Equation 4.1: VW_j is the pre-processed score of the respondent j in the weighting questionnaire, vw_j is the original score of the respondent j, k is the size of the Likert scale.

$$VW_j = \frac{vw_j - \frac{k+1}{2}}{\frac{k-1}{2}} \tag{4.1}$$

In Equation 4.2: VP_j is the pre-processed score of the respondent j in the perceptions questionnaire, vp_j is the original score of the respondent j, k is the size of the Likert scale.

$$VP_j = \frac{vp_j - \frac{k+1}{2}}{\frac{k-1}{2}} \tag{4.2}$$

With the pre-processed data, we obtain the average of user responses for each of the impact factors, separately for each questionnaire. Equation 4.3 presents the average of responses by impact factors for the weighting questionnaire. Equation 4.4 presents this average for the weighting questionnaire. Equation 4.3 defines w_i as the weight value for the impact factor i; n is the number of respondents to the weight questionnaire; VW_j is the score given by the respondent j for the impact factor i.

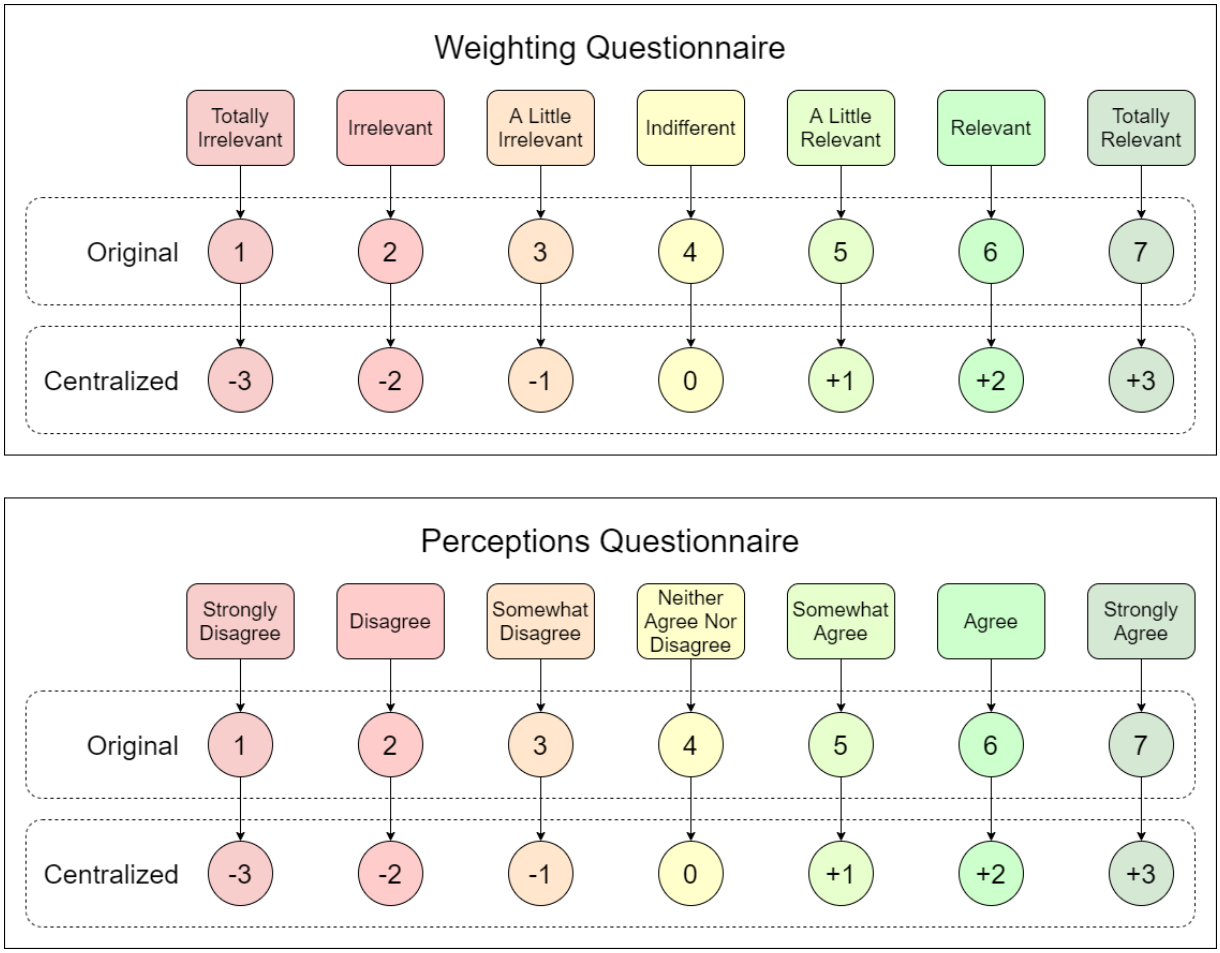

Figure 4.2: Mapping of responses on a Likert scale to numerical values

$$w_i = \frac{\sum_{j=1}^{n} VW_j}{n} \tag{4.3}$$

In Equation 4.4: p_i is the perception value for the impact factor i; n is the number of respondents to the perceptions questionnaire, VP_j is the score given by the respondent j for the impact factor i.

$$p_i = \frac{\sum_{j=1}^{n} VP_j}{n} \tag{4.4}$$

We calculate the available metric in humanization by the equation 4.5. Equation 4.5 performs a Manhattan normalization [3] on the weights acquired with the weighting questionnaire and multiplies the new weights by the results of the perceptions questionnaire, correlating the two questionnaires. Manhattan normalization consists of dividing the weights by the sum of their absolute values, and this operation guarantees that the metric value varies between -1 and 1.

The last step is to perform a horizontal translation (adding 1) and a scaling (dividing by 1/2) of the possible values of the metric, causing the metric to range between 0 and 1, a more common pattern among computational metrics.

In Equation 4.5: M is the final value of the metric; n is the number of identified

humanization impact factors, w_i is the average weight for the impact factor i, p_i is the average perception level for the i impact factor.

$$M = \frac{1}{2}(1 + \frac{1}{\sum_{i=1}^{n} |w_i|} \sum_{i=1}^{n} w_i p_i) \tag{4.5}$$

For the values of the performance metrics of each impact factor, the equation 4.6 is applied on the original data, not centered or scaled. Such metrics will vary between 0 and 1. In this way, the performance metrics will assess users' perceptions of each impact factor separately. In 4.6: F_i is the performance metric for the impact factor i, pi_j is the original perception value of the respondent j for the impact factor i, n is the number of respondents to the perceptions questionnaire, k is the size of the Likert scale.

$$F_i = \frac{\sum_{j=1}^{n} pi_j}{n} \times \frac{1}{k} \tag{4.6}$$

4.3 Application

All artifacts acquired so far, after being developed, become static. They do not need to be modified for different contexts of application of the methodology. Different organizations can apply both the weighting questionnaire and the perception questionnaire, and the results might adapt to the realities and needs of the chatbot being analyzed.

It is expected that the survey participants are not involved in the development process of the analyzed chatbot to avoid any bias in the acquired data. The best scenario for the execution of the questionnaires is that real users of the chatbots can answer the questions. However, it is also known that in some development contexts, chatbots cannot be made available to the public before all the tests are performed.

Another alternative for the research volunteers is the organization's employees developing the analyzed chatbot. The only restriction is that employees involved in any way in the chatbot development process do not participate in the evaluation protocol proposed. The analyzed chatbot developers' participation in the evaluation protocol can cause bias in the results.

4.3.1 Application of the Weighting Questionnaire

Ideally, the weighting questionnaire should only be applied once for each specific chatbot context. This recommendation is explained by the fact that the weights must be kept static between different versions of the same chatbot to make a fair comparison between the level of humanization achieved with each new chatbot model acquired.

After a period of use of the chatbot under analysis, the volunteer must inform how relevant each impact factor is, using the Likert psychometric scale starting from "totally irrelevant" to "totally relevant". In this way, the volunteer informs how relevant he considers each factor for constructing the chatbot's humanness under evaluation. This data refers to how desirable the presence of each factor in the chatbot is. The example in Table 4.1 demonstrates how we extract data from two examples of use case groups

that will be presented in the next chapter (Chapter 5). Table 4.1 also shows the average responses for each of the 12 selected impact factors.

Table 4.1: Example of Weighting Results

Impact Factors	Group 04						Group 21					
	Participants						Participants					
	1	**2**	**3**	**4**	**5**	**Mean**	**1**	**2**	**3**	**4**	**5**	**Mean**
Entertainment	7	7	7	7	5	**6,6**	1	4	5	5	4	**3,8**
Friendliness	7	7	7	7	7	**7,0**	4	3	6	4	6	**4,6**
Information and Updates	4	4	3	1	3	**3,0**	1	7	7	7	7	**5,8**
Homophily	6	7	7	7	6	**6,6**	3	4	4	3	7	**4,2**
Social Presence	6	7	6	6	7	**6,4**	4	6	7	4	7	**5,6**
Predictability	2	5	3	6	2	**3,6**	5	6	6	6	6	**5,8**
Dialogue	7	7	5	4	7	**6,0**	5	4	6	4	6	**5,0**
Novelty and Inspiration	4	7	6	6	4	**5,4**	3	5	7	5	5	**5,0**
Human-Likeness	5	7	5	3	6	**5,2**	4	1	5	2	6	**3,6**
Expertise	3	5	1	5	6	**4,0**	7	7	7	7	7	**7,0**
Credibility	3	5	2	3	3	**3,2**	4	7	7	7	7	**6,4**
Absence of Marketing	6	6	6	7	7	**6,4**	7	1	7	7	6	**5,6**

4.3.2 Application of the Perceptions Questionnaire

The perceptions questionnaire must be applied to each new version of the analyzed chatbot; even in the initial version of the chatbot, without the need to repeat the application of the weighting questionnaire. With the evolution of the chatbot over version of the development, it is expected that the acquired results are better and better since the results of one iteration can contribute to the evolution of the next version.

After a period of use of the chatbot under analysis, the volunteer must inform how present each impact factor is in the chatbot, using the Likert psychometric scale starting from “totally irrelevant” to “totally relevant”. This data refers to how much each factor is present and identifiable within the chatbot. The example in Table 4.2 demonstrates how we extract data from two examples of use case groups that will be presented in the next chapter (Chapter 5). Table 4.2 also shows the average responses for each of the 12 selected impact factors.

4.3.3 Application of the Metrics

Metrics are applied using the responses obtained in the weighting questionnaire (applied in the first iteration of chatbot development); the responses obtained in the weighting questionnaire, and the metrics generation methodology described in section 4.2. In each application of the method for a chatbot, the results are analyzed in two ways: quantitative and qualitative.

Metrics are applied using the responses reached in the Weighting Questionnaire (applied in the first iteration of chatbot development). We obtain such metrics by applying the

Table 4.2: Example of Perception Results

Impact Factors	Group 04 Participants						Group 21 Participants					
	1	2	3	4	5	Mean	1	2	3	4	5	Mean
Entertainment	3	4	7	5	5	**4,8**	3	4	4	7	5	**4,6**
Friendliness	4	4	6	7	5	**5,2**	3	5	3	4	4	**3,8**
Information and Updates	7	3	7	5	5	**5,4**	5	7	4	7	6	**5,8**
Homophily	2	4	6	7	6	**5,0**	3	4	2	5	5	**3,8**
Social Presence	2	6	6	7	6	**5,4**	6	6	4	4	5	**5,0**
Predictability	7	3	7	6	5	**5,6**	7	6	6	7	6	**6,4**
Dialogue	6	2	6	7	5	**5,2**	3	5	7	5	3	**4,6**
Novelty and Inspiration	5	3	7	4	6	**5,0**	4	6	6	6	6	**5,6**
Human-Likeness	1	5	5	5	5	**4,2**	2	5	3	4	4	**3,6**
Expertise	7	2	6	7	5	**5,4**	6	7	7	7	6	**6,6**
Credibility	7	3	6	6	5	**5,4**	6	7	7	7	6	**6,6**
Absence of Marketing	7	7	7	4	6	**6,2**	7	7	7	5	7	**6,6**

metrics generation methodology (described in Section 4.2) to the responses acquired with the Weighting Questionnaire. In each application of the method for a chatbot, we analyze the results in two ways: quantitative and qualitative. Table 4.3 shows how the data pre-processing of both weights and perceptions occurred for group 04, one of the groups in the use case present in the next chapter (Chapter 5).

In quantitative aspects, two types of metrics are generated: a general metric, which indicates how close to the ideal humanization level the evaluated chatbot is; and other metrics related to the chatbot's effectiveness in each of the impact factors separately. The explanation of such metrics was carried out in Section 4.2. As these metrics are quantitative, they can be compared with previous versions of the same chatbot, showing the evolution of the chatbot's humanization level in general and of each impact factor separately. In this sense, developers can confirm whether their efforts to obtain more humanized chatbots affect the metrics' evolution.

In qualitative aspects, the metrics generated for each impact factor separately can reveal which characteristics are better perceived by users and which ones need to be better developed. Developers can be guided by these qualitative results, from the most developed and the most minor developed impact factors, to improve chatbot models punctually in the necessary characteristics. The results are analyzed for each impact factor in separate humanization to identify strengths and weaknesses of the current conversational model and so propose improvements. Table 4.4 shows how all steps of a metric execution case were calculated for perception and weighting data. The data were obtained by group 04, one of the groups in the use case present in the next chapter (Chapter 5). It also presented the general metric in humanization for the data presented (using 12 impact factors) and a glossary of each calculation performed.

Table 4.3: Pre-Processed Data from the Questionnaires for Group 04

Weighting Questionnaire Results (Equation 4.1)						
Impact Factors	Participants					
	1	2	3	4	5	Mean
Entertainment	1,00	1,00	1,00	1,00	0,33	**0,87**
Friendliness	1,00	1,00	1,00	1,00	1,00	**1,00**
Information and Updates	0,00	0,00	-0,33	-1,00	-0,33	**-0,33**
Homophily	0,67	1,00	1,00	1,00	0,67	**0,87**
Social Presence	0,67	1,00	0,67	0,67	1,00	**0,80**
Predictability	-0,67	0,33	-0,33	0,67	-0,67	**-0,13**
Dialogue	1,00	1,00	0,33	0,00	1,00	**0,67**
Novelty and Inspiration	0,00	1,00	0,67	0,67	0,00	**0,47**
Human-Likeness	0,33	1,00	0,33	-0,33	0,67	**0,40**
Expertise	-0,33	0,33	-1,00	0,33	0,67	**0,00**
Credibility	-0,33	0,33	-0,67	-0,33	-0,33	**-0,27**
Absence of Marketing	0,67	0,67	0,67	1,00	1,00	**0,80**
Perceptions Questionnaire Results (Equation 4.2)						
Impact Factors	Participants					
	1	2	3	4	5	Mean
Entertainment	-0,33	0,00	1,00	0,33	0,33	**0,27**
Friendliness	0,00	0,00	0,67	1,00	0,33	**0,40**
Information and Updates	1,00	-0,33	1,00	0,33	0,33	**0,47**
Homophily	-0,67	0,00	0,67	1,00	0,67	**0,33**
Social Presence	-0,67	0,67	0,67	1,00	0,67	**0,47**
Predictability	1,00	-0,33	1,00	0,67	0,33	**0,53**
Dialogue	0,67	-0,67	0,67	1,00	0,33	**0,40**
Novelty and Inspiration	0,33	-0,33	1,00	0,00	0,67	**0,33**
Human-Likeness	-1,00	0,33	0,33	0,33	0,33	**0,07**
Expertise	1,00	-0,67	0,67	1,00	0,33	**0,47**
Credibility	1,00	-0,33	0,67	0,67	0,33	**0,47**
Absence of Marketing	1,00	1,00	1,00	0,00	0,67	**0,73**

4.4 Discussion

It is necessary to discuss when is the best time to apply the weighting questionnaire: if there is already an initial version of the proposed chatbot, or if there is already a well-defined idea of the chatbot under development. Figure 4.1 showed this is when an initial chatbot version already exists. Nevertheless, we are considering a development environment in which some initial chatbot has at least been prototyped to clarify what is desired.

Applying the weighting questionnaire can be carried out as soon as the design time has a clarified idea of the proposed chatbot design problem. This happens when you the equip has enough knowledge of the problem the chatbot must address, which tasks are the responsibility of this chatbot, and which target audience such a chatbot must serve. Having these well-defined design decisions, it is plausible to apply the weighting

Table 4.4: Generation of Metrics in Humanization for Group 04

Impact Factors	Weighting Results				Perceptions Results			
	w	|w|	N1	|N1|	p	|p|	SF	IP
Entertainment	0,87	0,87	0,13	0,13	0,27	0,27	**0,62**	0,0350
Friendliness	1,00	1,00	0,15	0,15	0,40	0,40	**0,70**	0,0606
Information and Updates	-0,33	0,33	-0,05	0,05	0,47	0,47	**0,42**	-0,0236
Homophily	0,87	0,87	0,13	0,13	0,33	0,33	**0,64**	0,0438
Social Presence	0,80	0,80	0,12	0,12	0,47	0,47	**0,69**	0,0566
Predictability	-0,13	0,13	-0,02	0,02	0,53	0,53	**0,46**	-0,0108
Dialogue	0,67	0,67	0,10	0,10	0,40	0,40	**0,63**	0,0404
Novelty and Inspiration	0,47	0,47	0,07	0,07	0,33	0,33	**0,58**	0,0236
Human-Likeness	0,40	0,40	0,06	0,06	0,07	0,07	**0,51**	0,0040
Expertise	0,00	0,00	0,00	0,00	0,47	0,47	**0,50**	0,0000
Credibility	-0,27	0,27	-0,04	0,04	0,47	0,47	**0,44**	-0,0189
Absence of Marketing	0,80	0,80	0,12	0,12	0,73	0,73	**0,79**	0,0889

Final Metric in Humanization (Equation 4.5): **0,6498**
w: Average Weighting (Equation 4.3)
|w|: Absolute Average Weighting (Equation 4.6)
N1: Norm 1 of Weighting (Equation 5.1)
|N1|: Absolute Norm 1 of Weighting (Equation 5.2)
p: Average of Perceptions (Equation 4.4)
|p|: Absolute Average of Perceptions (Equation 4.6)
SF: Score by Impact Factor (Equation 5.3)
IP: Internal Product (Weighting versus Perceptions) (Equation 5.4)

questionnaire.

We clarify that the Weighting Questionnaire aims to verify how desirable the impact factors are in a chatbot under development. Therefore, the goal is for the end-user to have a good experience using this chatbot. If an initial version of the chatbot already exists, the weighting questionnaire can be applied to verify what would be desirable to improve in this initial chatbot. However, if there is already a well-defined idea for developing this chatbot, the weighting questionnaire can also be applied.

Another issue that must be discussed is the set of impact factors identified. We investigated the literature looking for impact factors that have already been investigated in other cases of humanization assessment. This set is neither stagnant nor fixed and can be changeable, mainly depending on the different contexts in which chatbots will be used in the future. The methodology must be able to adapt to new types of impact factors, depending on variations over time, both in the current set and the evolution of the current set of impact factors identified by this study. We observe that one of the relevant aspects of this methodology is that a set of impact factors can be modified over time, mainly according to the changes in contexts in which chatbots are applied.

4.5 Conclusion

In this Chapter, we addressed a methodology for chatbot humanization evaluation capable of adapting to different contexts, for different purposes and target audiences. The methodology is adaptive to the different sets of impact factors being analyzed. This makes our methodology as dynamic as possible, and as adaptive as possible over time. It is expected that the methodology can be a skeleton for evaluating humanization in the long term.

We need to indicate that the assessment of humanity in the context of the methodology is not restricted to chatbots alone. This same methodology can be adapted to several other types of user interfaces, such as websites, human-computer interaction software, voice systems, and even algorithms for identifying facial expressions. Therefore, our proposal in this Chapter is not only a method of evaluating humanization for chatbots, but can also be applicable to evaluating humanization in HCI in general way.

In the future, we can even apply our proposed methodology to other contexts. This must imply unforeseen challenges that require further investigations.

Chapter 5

Case Study

We assessed our proposal to understand its applicability. Our proposed methodology was applied in the first semester of 2022 during an undergraduate course on Construction of Human-Computer Interfaces, whose objective was to teach students how designing and evaluating user interfaces in interactive software systems. The author of this dissertation acted as monitor in the course by assisting in applying the proposed methodology in the case of the discipline in question. Some modifications were proposed regarding the planning of the discipline for considering our proposed methodology in the context of project development by students. Since the methodology aims to evaluate the humanization of chatbots, the proposal was building chatbots as human-computer interfaces.

5.1 Context and Participants

The course objective was to study how to design and evaluate user interfaces in software systems, as well as to study processes for interaction design. One of the learning goals was to obtain sensitivity to the usability of computer systems, that is, to learn human capabilities and limitations, in addition to design principles and standards for usability. The students must acquire a holistic view of software design, learn how to reconcile stakeholders' interests, and understand the importance of well-designed interfaces. In the course, the students must develop low and high fidelity prototypes seeking inclusive and participatory approaches for constructing and evaluating interface solutions for applications in different contexts. In the context of the first semester of 2022, it was proposed that the proposed solutions have the interactive format of chatbots. The following topics were covered in the course:

- Introduction to Human-Computer Interfaces (HCI): History and Evolution;
- Human Factors in HCI (Mechanisms of Human Perception and Memory);
- Paradigms in HCI;
- Design Methods and Techniques (User-Centered Design, Participatory Design);
- Prototyping;
- Interface Assessment;
- Usability;
- Accessibility and Universal Design;

- Environments and Tools for User Interface Specification, Construction, and Evaluation.

Five practical tasks were proposed throughout the course that involved solving design and evaluation problems. The activities were carried out individually or in pairs. In addition to the tasks, a practical project was developed aiming at the transversal application of the HCI concepts and techniques studied throughout the semester. This project was carried out in groups of 5 to 6 students. The projects developed by the groups concern the design, modeling, and prototyping of solutions in user interfaces using the proposed methodological artifacts and tools. In particular, user interfaces should be chatbots. In addition to the specification and prototypes resulting from the project, the groups presented reports describing the project developed.

In this sense, our proposed methodology for evaluating humanization in chatbots was taught to students and applied during the course project execution during the first semester of 2022. The design and prototyping of 11 chatbots thought by the students themselves were monitored by the monitors of the course taught. In the following section (Section 5.2) we describe the timeline on which the project and the evaluation methodology were carried out.

5.2 Methodology

Figure 4.1 presents the application methodology proposed (adapted from the original methodology). We considered that the application of the weighting questionnaire should be carried out before the low prototyping, at the time when ideas about how the chatbot design should be clarified. In the context of the discipline, this makes sense because it is at this point that students must be as transparent as possible about the target audience's intentions and what the chatbot should accomplish. Then the perception questionnaire was applied to each new high-prototyping or reprototyping of the chatbot (cf. Figure 5.1). Figure 4.1 presented in chapter 4 presented a more general version of the application shown in Figure 5.1 regarding the application of the methodology.

5.2.1 Dynamics of the Discipline

The chatbot humanization assessment methodology was applied in the context a discipline on Human-Computer Interaction conducted at IC/UNICAMP. More specifically in the practical project of the discipline, which is mandatory for all students. The application of the methodology took place only during the execution of the practical project (cf. Figure 5.1).

We describe the pre-project phases, which consist of the theme proposition (Subsection 5.2.1.1); the problem selection (Subsection 5.2.1.2); and the formation of the groups (Subsection 5.2.1.3); then, we describe the phases of project execution which consist of the design problem clarification (Subsection 5.2.1.4); the low prototyping (Subsection 5.2.1.6); the high prototyping (Subsection 5.2.1.7); the evaluation (Subsection 5.2.1.8); and the video generation (Subsection 5.2.1.10). We present the optional phases of the

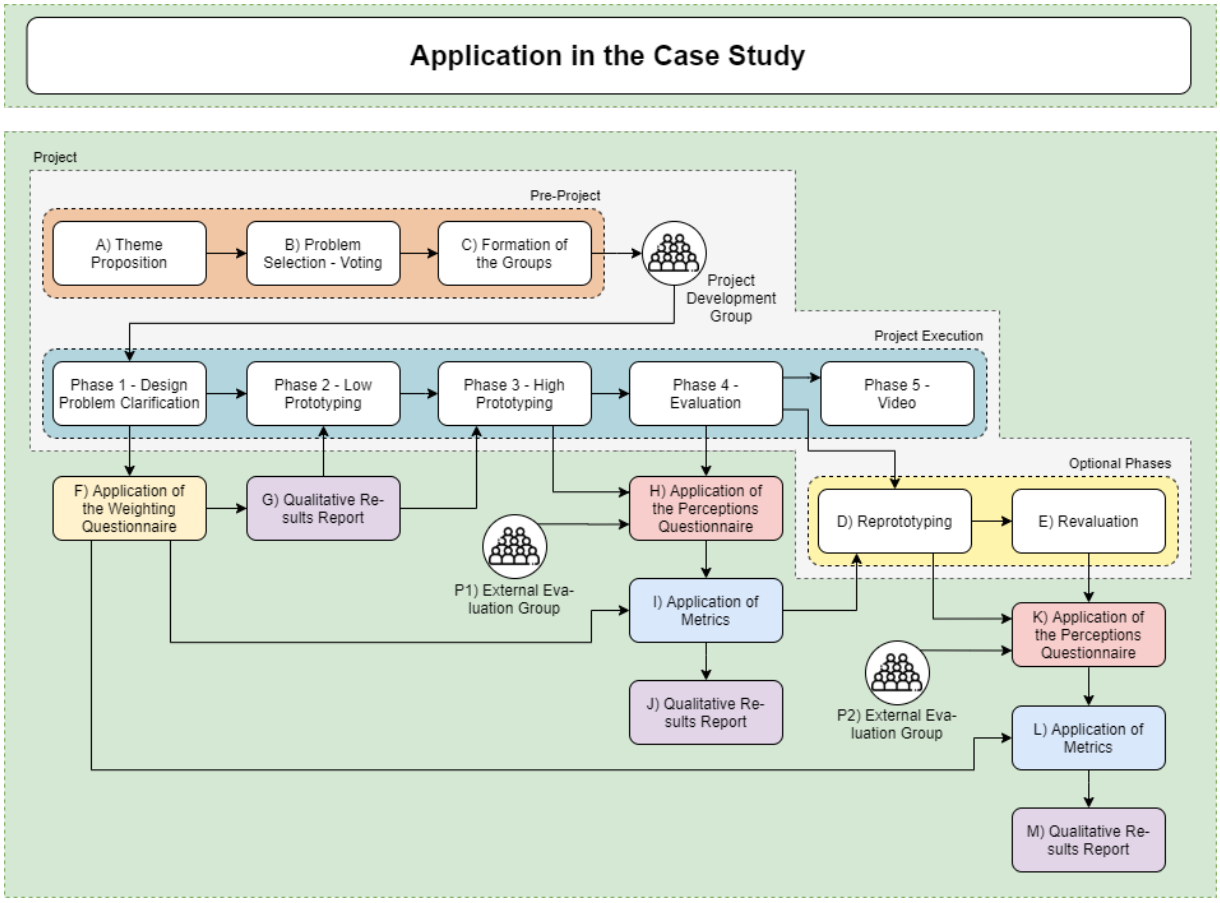

Figure 5.1: Application of the Method for Evaluating Humanization in Chatbots in the MC750A Course

project, which are the reprototyping (Subsection 5.2.1.11) and the reassessment (Subsection 5.2.1.12) of the project.

5.2.1.1 Theme Proposition

Ideas for possible design problems were proposed through a form. In each proposal, the students determined a general title for the problem and a brief description of the idea and its objective. In this idea, there is brainstorming of design problems.

5.2.1.2 Problem Selection - Voting

After collecting student ideas, the themes were made available to all students. An election was held among the discipline's students to select the most promising theme propositions. Only the 12 most voted topics were considered for the group formation stage. Thus, only after this vote, the respective groups were formed.

5.2.1.3 Formation of the Groups

For the development of a design solution for a given problem, groups were formed according to the aptitude of each student for a given topic. The groups were composed of 5 to 6 members. Each student indicated three topics, the first being the most interesting and the

third the least interesting. The order of composition of the groups was given by order of arrival of the students in filling out the online form. Before joining a particular group, the student should look at the groups worksheet to verify that the chosen topic was already complete. If the student entered a topic that was already complete, he/she was reallocated to the second option indicated. If the second option was complete, he/she was reallocated to the third. Students who did not indicate the themes were allocated to groups that were not complete.

5.2.1.4 Phase 1 - Design Problem Clarification

This phase aimed to present the result of the application of instruments and artifacts that allowed to clarify the design problem. At this project stage, it was necessary to understand and clarify the design problem of the interactive system under construction in the role of designers. The final document of this phase was submitted to the online teaching-learning environment.

5.2.1.5 Application of the Weighting Questionnaire

Along with Phase 1 of the project, the application of the Weighting Questionnaire was started since the methodology for evaluating humanization in chatbots can be an excellent instrument for clarifying the needs of the intended chatbot. The purpose of the Weighting Questionnaire is to determine which impact factors are more or less critical; and, similarly, to determine which factors are desirable and which are undesirable in chatbot behavior. At this stage of the project about humanization, the objective was for each group to identify possible users of the proposed chatbots, thus explaining the context and purpose of the respective chatbot for such future users. Such users could even be the participants of the groups themselves; this became an activity for all the participants of the groups as well. Users accessed the questionnaire, selected the evaluated topic, and filled out the form. Moreover, any user could complete this questionnaire, but the responses were restricted to a single e-mail to avoid repeated respondents. Respondents' e-mails were collected but restricted and not shared even with the discipline's students, and the responses were made available through a non-editable link.

5.2.1.6 Phase 2 - Low Prototyping

The objective of this phase was to present the results of applying low-fidelity prototyping techniques and participatory design to obtain design alternatives from different parts of the chatbot. Low-fidelity prototypes (on paper) were built to represent interaction elements in the interfaces. For this, brainwrite techniques were applied to obtain possible usage scenarios [14]; braindraw techniques were conducted to obtain design alternatives that were later consolidated [28]; wireframe proposals were created (as a way to consolidate design proposals) [57]; and design decisions were recorded; elements of interfaces consolidated from the alternatives found via braindraw were documented and justified; storyboarding were defined to illustrate the interaction narrative of prototyped features [49]. The final document of this phase was submitted to the online teaching-learning environment.

5.2.1.7 Phase 3 - High Prototyping

In this phase, the objective was to build high-fidelity prototypes based on the results obtained from storyboarding and participatory design activities carried out in Phase 2. The methodology used to obtain the results is described in detail and discussed. The final document of this phase was submitted to the online teaching-learning environment.

5.2.1.8 Phase 4 - Evaluation

This phase involved the presentation of evaluations conducted to investigate different aspects of the interaction with the prototype created (the result of Phase 3). It was necessary to conduct at least two assessments:

- **Evaluation with The End User**. We are planning and evaluating the design proposal from Phase 3 (high fidelity prototype) with a prospective user. Analyze and document the results carefully.
- **Usability Inspection using Nielsen's ten Heuristics**. Inspection of the main functionalities/screens of the prototype. Analyze detected problems and suggest how they could be solved or mitigated.

5.2.1.9 Application of the Perceptions Questionnaire

The interface evaluation and the Perceptions Questionnaire were applied along with the beginning of Phase 4 of the project. The objective was to evaluate how much users perceived the presence of each of the characteristics in the chatbot used. Such characteristics and positive and negative perceptions vary according to the context in which they are analyzed. In this project phase, concerning humanization, the objective was for people outside the group under analysis to evaluate the effectiveness regarding humanization of the generated chatbot. It could be the classmates themselves evaluating each other's projects. Users accessed the questionnaire, selected the evaluated topic, and filled out the form. Any user could fill in the questionnaire. Respondents' e-mails were collected, but restricted and not shared even with the discipline's students, and the responses were made available to the groups through a non-editable link.

5.2.1.10 Phase 5 - Video

At this stage, the team gave an oral presentation of approximately 20 minutes on the project developed, and the set of slides produced for this purpose was also delivered. A video was produced explaining the slides presenting the project's end-to-end results. The purpose of the video was to clarify what was developed and how throughout the course (going through each phase of the project) and to present the decisions made, the justifications, and the result obtained (final prototype of evaluated interaction).

5.2.1.11 Reprototyping

This phase of the project was optional, mainly as an exercise on the evolutionary development of a project and on how to use the results of the first round of application

of the proposed methodology for the evolution of the project. In this way, the second high-fidelity prototype was built based on the results obtained during the evaluations carried out in Phase 4 (Subsection 5.2.1.8), together with the results of the first application of the chatbots humanization assessment methodology (Subsection 5.2.1.5 and Subsection 5.2.1.9). As this is an optional phase, describing details through formal documentation was not necessary.

5.2.1.12 Reassessment

This project Reassessment Phase was also optional, as well as the Reprototyping Phase, with the main objective of verifying improvements of all humanization metrics regarding the second high-fidelity prototype (developed in the Reprototyping Phase in Subsection 5.2.1.11). Taking into account the second high-fidelity prototype, it was evaluated how much users perceived the presence of each of the characteristics in the chatbot used, through the Perceptions Questionnaire, in the same way as presented in the first application of the Perceptions Questionnaire, presented in 5.2.1.9.

5.2.2 Application of our Methodology for chatbot evaluation

Figure 5.1 presents the application of the methodology co-occurred and using the results of each phase of the execution of the practical project in the discipline, which worked with the context of chatbots. The first questionnaire applied was the weighting (Subsection 5.2.1.5), using the results of Phase 1 of the Project, on design problem clarification (Subsection 5.2.1.4). This is important because, in this phase of clarification of the design problem, we have several artifacts, such as what is expected from the system, what type of interaction is desired, a greater understanding of the problem, and especially who is the likely target audience of the chatbot being designed. On this bases, results were generated based on the Weighting Questionnaire and the group participants being analyzed, i.e., the group designing the chatbot under analysis. As they clarify their ideas about what the chatbot should look like, they can use this idea to evaluate what is most relevant and desirable in the project being developed.

After applying the weighting questionnaire (Subsection 5.2.1.5), qualitative results were generated regarding the weights suggested for the project in progress. This report serves as input for Phase 2, low prototyping (Subsection 5.2.1.6), and Phase 3, high prototyping (Subsection 5.2.1.7). This presents which impact factors are more desirable for the chatbot under development; which impact factors are less desirable; and at what level they are more or less desirable. It was essential to verify how relevant each of these factors were for the development of the chatbot, because even if a given impact factor is very undesirable, it ends up having great relevance in the conception of the chatbot's humanity, as it should be avoided. The relevance and how desirable each of these impact factors is for the project under analysis guides the development, both the low prototype with the high prototype, presenting how the chatbot should proceed when executed by the end user.

The perceptions questionnaire (Subsection) was the next humanization assessment artifact applied to the use case. The perceptions questionnaire was applied during Phase 4, evaluation (Subsection 5.2.1.8), using as input the results of Phase 3, high prototyping

(5.2.1.7), which consists of a chatbot already usable for some external audience. Both the use of actual chatbots and videos of the chatbots working in that prototyping phase were evaluated. The audience that evaluated each chatbot was external to the group. They could be people outside the discipline or even students of the discipline who did not participate in the development group of the project under analysis. After the application of the perceptions questionnaire, an application of the metrics was carried out (Subsection 4.2.2); and finally, a report of qualitative results was generated. In this sense, the final humanization metric was obtained at this stage of the chatbot evolution. We verify the effectiveness of each of the impact factors for the current chatbot design and determine how the evolution should proceed. For example, which impact factors were not so well developed and which are already well matured in the chatbot.

Among the optional phases of the practical project of the discipline, there are the phases of reprototyping (Subsection 5.2.1.11) and reassessment (Subsection 5.2.1.12). As an input to the reprototyping phase, we have the metrics in humanization, seeking to make the chatbot closer to perfect humanization and improve the performance of each chatbot impact factor. At this stage, it is expected that those impact factors that have not performed well can be improved, and those that have already performed well can continue. Thus, during the reassessment, the perception questionnaire (Subsection 5.2.1.9) was again applied to this new prototype obtained, using a group of users external to the group. This was conducted to avoid bias in the results, and the metrics were again applied. The qualitative results report was obtained, so it was possible to verify if humanization has improved in relation to the previous prototype. It is also possible to verify if the impact factors developed in the previous prototype were developed in the current prototype.

5.2.3 Results Analysis

The results were analyzed using the equations presented in Section 4.2.2. In addition to the calculations presented in Section 4.2.2, for the results of the Weighting Questionnaire, the results of Manhattan Normalization (Equation 5.1, where w is the vector with the average of the weights and k is the index of the investigated impact factor) were analyzed for each of the impact factors. Such results demonstrate how desirable each impact factor is for the proposed chatbot. The Manhattan Normalization Absolute Value (Equation 5.2, where w is the vector with the average of the weights and k is the index of the investigated impact factor) for each of the impact factors presents the relevance of each impact factor for projecting the idea of humanization of the chatbot under analysis. The impact factors were ordered for each group according to the Manhattan Normalization Absolute Values to identify which impact factors should be worked on and which are not of great concern.

$$n1_k = \frac{w_k}{\sum_{i=1}^{n} |w_i|} \tag{5.1}$$

$$|n1_k| = |\frac{w_k}{\sum_{i=1}^{n} |w_i|}| \tag{5.2}$$

In the case of the results of the perception questionnaire, a metrics analysis was carried out on each of the impact factors separately to verify the chatbot's performance in each of

them through the equation 5.3 (where p_k is the average of the pre-processed insights of the k-index impact factor, and w_k is the weight of the k-index impact factor). Equation 5.4 (where p_k is the average of the pre-processed insights of the k-index impact factor, and w_k is the weight of the k-index impact factor) between the weighting and perception results to verify how impactful each impact factor score was in generating the final metric in humanization.

$$mp_k = \frac{1}{2}[1 + (p_k \times w_k)] \tag{5.3}$$

$$ip_k = |(p_k \times w_k)| \tag{5.4}$$

5.2.4 Supporting Materials

Video classes were prepared and applied during the course to teach students about our chatbot humanization assessment methodology. Four video-class on the humanization of chatbots and two video-class on the simple construction of chatbots were applied:

- **Humanization Video-Class 1:** What is Humanization?
- **Humanization Video-Class 2:** Weighting Questionnaire
- **Humanization Video-Class 3:** What are Weightings? How Should They be Interpreted?
- **Chatbots Video-Class 1:** How to Build Chatbots in a Simple Way?
- **Chatbots Video-Class 2:** How to Build Chatbots in a Simple Way?
- **Humanization Video-Class 4:** Questionnaire of Perceptions and Generation of Metrics

All this content and a little more, such as examples of forms, video lessons, and tools, are available at the following link. We call this content repository the Toolbox for Humanization Evaluation in Chatbots.

5.3 Results

We present the results of both the Weighting and the Perceptions questionnaire (respectively, Subsections 5.3.1 and 5.3.2) in addition to the Metrics in Humanization (Subsection 5.3.3) and the Evolution in Metrics in Humanization for the three groups that performed the optional reprototyping phase (Subsection 5.3.4). In the four subsections mentioned above, we chose to preferentially display results only using the impact factors that presented the most significant variance between the groups' results, discarding the impact factors with medium and low variance. Using impact factors with high variance allows a clearer view of the results. In Table 5.2, impact factors are ordered according to their variance, and those with high variance are highlighted in red.

The results of a specific project are also presented, in this case, project 04 - Someone who Understands You (Section 5.3.5), using the most relevant impact factors for the specific project 04.

5.3.1 Weighting Results

In the case of the weighting results, we can see that the relevance of some impact factors varies significantly from one chatbot theme to another. Some impact factors are very relevant for some chatbots while others are not. In some chatbots, some impact factors have neither relevance nor minimal relevance. Figure 5.2 shows how each chatbot proposal weight can characterize the problem, target audience, and software objectives.

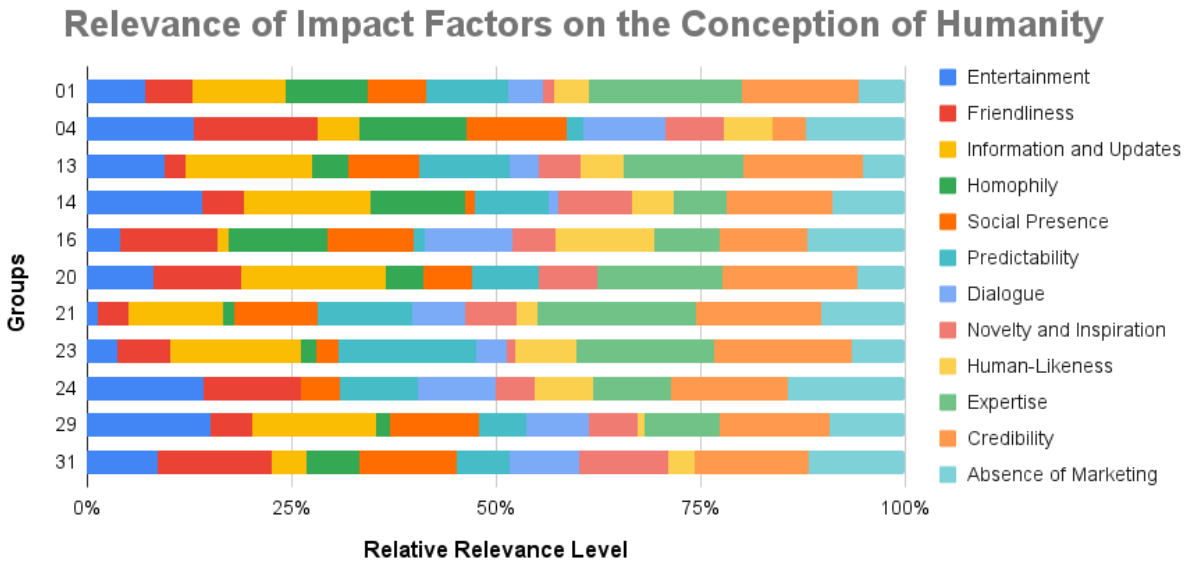

Figure 5.2: Relevance of Impact Factors on the Conception of Humanity in each of the Case Study Chatbots

5.3.2 Perception Results

As for the perception results, we can see that in the first high prototyping, there is a trend of similar results among the impact factors for all analyzed chatbots, as seen in Figure 5.3. Despite the similarity trend, in Figure 5.3 we can already see some very high and deficient scores from some groups. Such variances probably already demonstrate the care the teams took to treat each impact factor with due relevance, presented in Figure 5.2.

In Figure 5.4, we observed that, on average, some impact factors performed better than others. This performance variation can have two explanations. The first is that all chatbots are in the same prototyping phase, probably sharing the doubts and difficulties of the current state of development. The second explanation would be the possible trend of higher or lower weights for some impact factors, according to the chatbots' needs. In Table 5.2 we can verify each impact factor's variance and average weight.

5.3.3 Metrics in Humanization

As for humanization metrics, in most cases, the high prototypes had results above 75% of the ideal humanization, with only one group scoring below 75%, as can be seen in figure 5.5. However, even so, we see the variation between one chatbot and another regarding

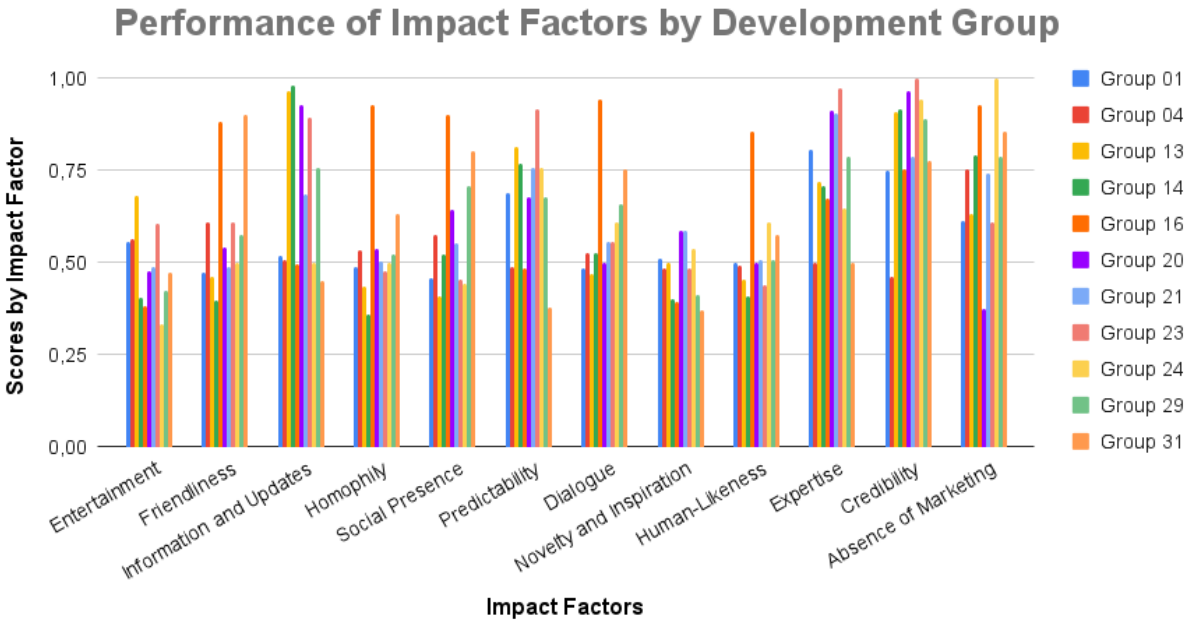

Figure 5.3: Performance of Impact Factors by Development Group according to Perception Results

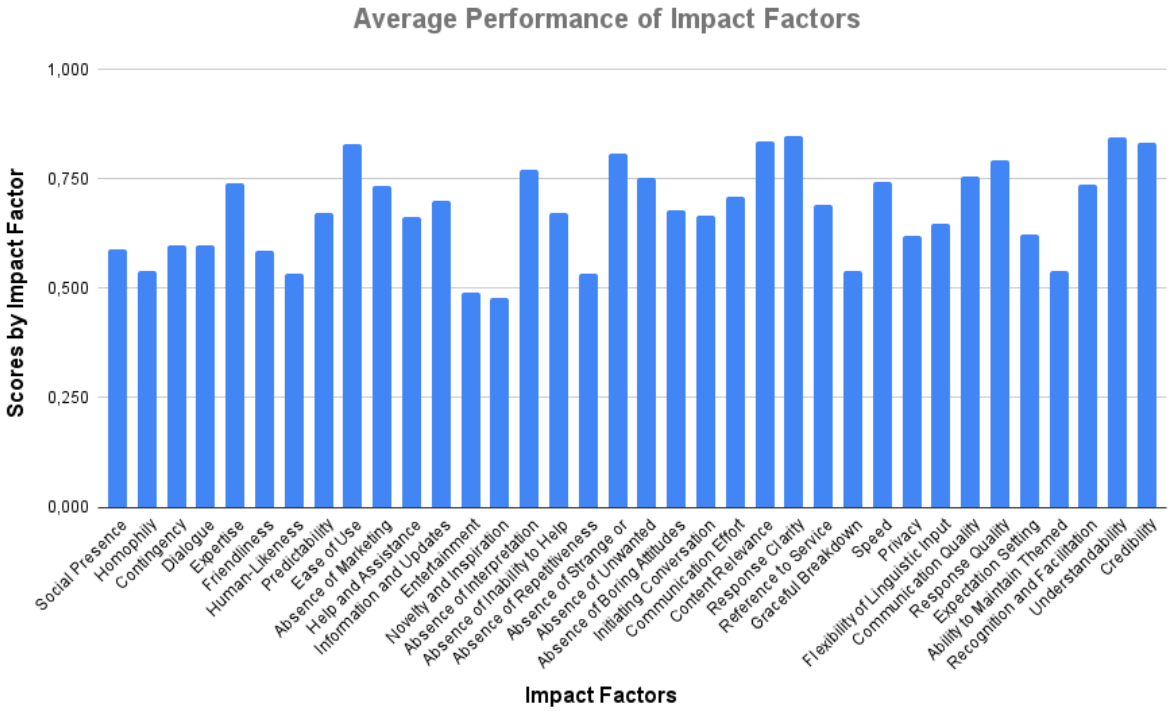

Figure 5.4: Average Performance of Impact Factors according to Perception Results

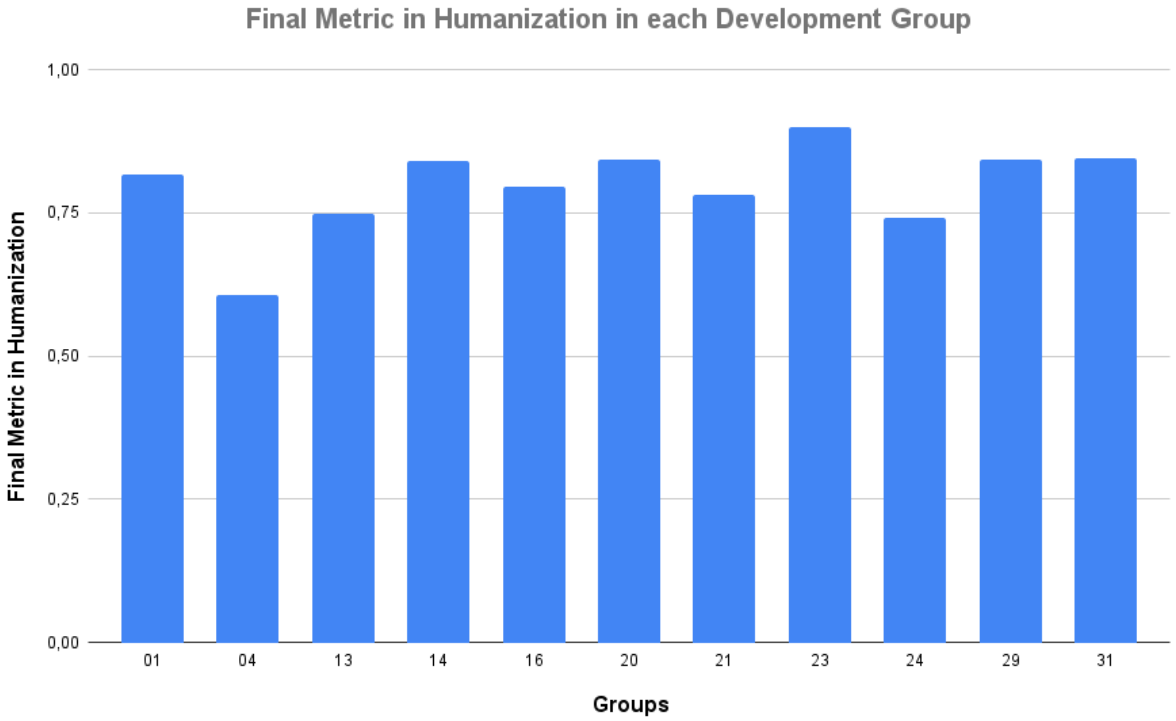

Figure 5.5: Final Metric in Humanization in each Development Group

the level of humanization. This depends a lot on the development of the high prototype itself and the techniques used.

5.3.4 Evolution in Metrics in Humanization

After evaluating the first discharge prototype, in which we performed the first application of the Weighting Questionnaire, we started with two optional phases for the groups: reprototyping (Subsection 5.2.1.11) and reassessment (Subsection 5.2.1.12). Three groups participated in these optional phases, and we verified the effectiveness of the evolution of metrics in humanization when using the methodology presented in this work. The groups that performed the reprototyping and reassessment were: group 04 (*Someone who Understands You*), group 16 (*Mental Health*), and group 23 (*Where to find your Movie/Series?*).

Figure 5.6 shows the advances in the average performance of the impact factors with the highest variance (according to Table 5.2). We observed performance improvement in seven of the 12 factors presented. That is, in most of the presented impact factors. The average performance remained the same in two impact factors, and the performance worsened in three.

In Figure 5.7, we present the progress of metrics in humanization by the project developed. In two of the three projects, there was a clear advance in the humanization metric; in one, the value of the final humanization metric dropped. We can also observe that the group had already obtained a higher grade in the first evaluation round than the other two groups in the two evaluation rounds. The higher the metric value in

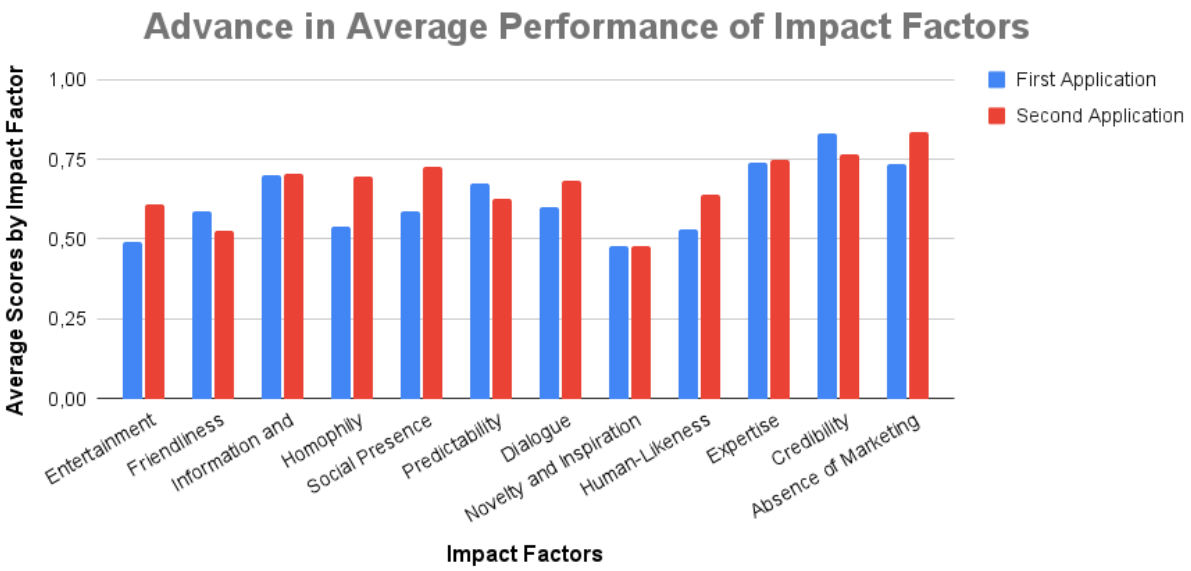

Figure 5.6: Advance in Average Performance of Impact Factors

humanization, the more difficult it is to improve it. At this stage, humanization depends a lot on the details. A possible future work would be the application of the methodology in longer evolutionary processes of software development, where we could verify if it is more challenging to improve the scores in humanization when they are already relatively high.

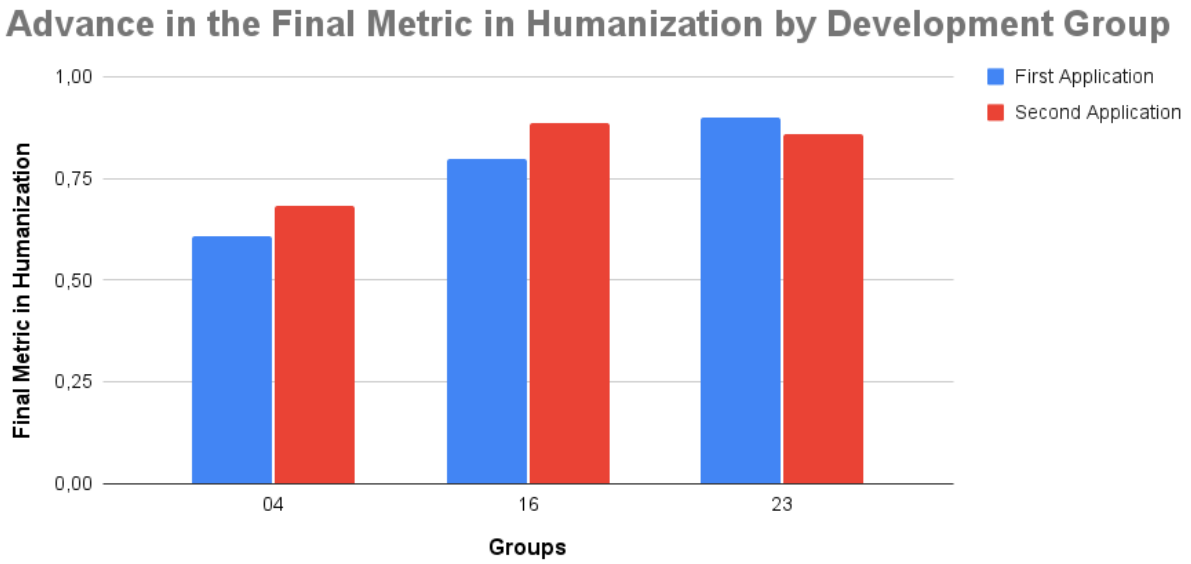

Figure 5.7: Advance in the Final Metric in Humanization by Development Group

5.3.5 Specific Example of Humanization Assessment

In this subsection, we will follow the humanization assessment of the project developed by group 04, one of the groups formed during the MC750A course about Construction of Human-Computer Interfaces. The title of the project of group 04 is “Someone who Understands You”, and below, we present the initial description presented during the

theme proposition phase (presented in Subsection 5.2.1.1):

"Sometimes all you need is to talk to someone. If someone took the last book in front of you at the library, if you got a 4.95 on the test, or if you stood in line for two hours at the university restaurant and found that you forgot to put credit on your card, the chatbot will always be there to comfort you. To give you words of comfort and affection when you need it or to give you a warning when you deserve it."

We will present below the Weighting Questionnaire Results, the Results of the First Application of the Perceptions Questionnaire, the Results of the Second Application of the Perceptions Questionnaire, and the Evolution with the Use of Methodology.

5.3.5.1 Weighting Questionnaire Results

Here, we present two artifacts generated with the application of the Weighting Questionnaire in the case of group 04. The Weighting Questionnaire was applied together with phase 1 on the Design Problem Clarification (as shown in the Subsection 5.2.1.5). Both artifacts show how relevant each impact factor is for the conception of the idea of humanization. They use the Manhattan Normalization Absolute Value (Equation 5.2).

Figure 5.8 presents the pie chart that represents the impact of factors on the idea of humanization. An impact factor relevant to the conception of the humanity of a chatbot does not necessarily mean that such a factor is desired but that greater attention should be paid to it. In the graph of Figure 5.8, we can see that *Friendliness* affects the idea of humanity a lot (5.5%), while *Reference to Service* affects little (1.5%).

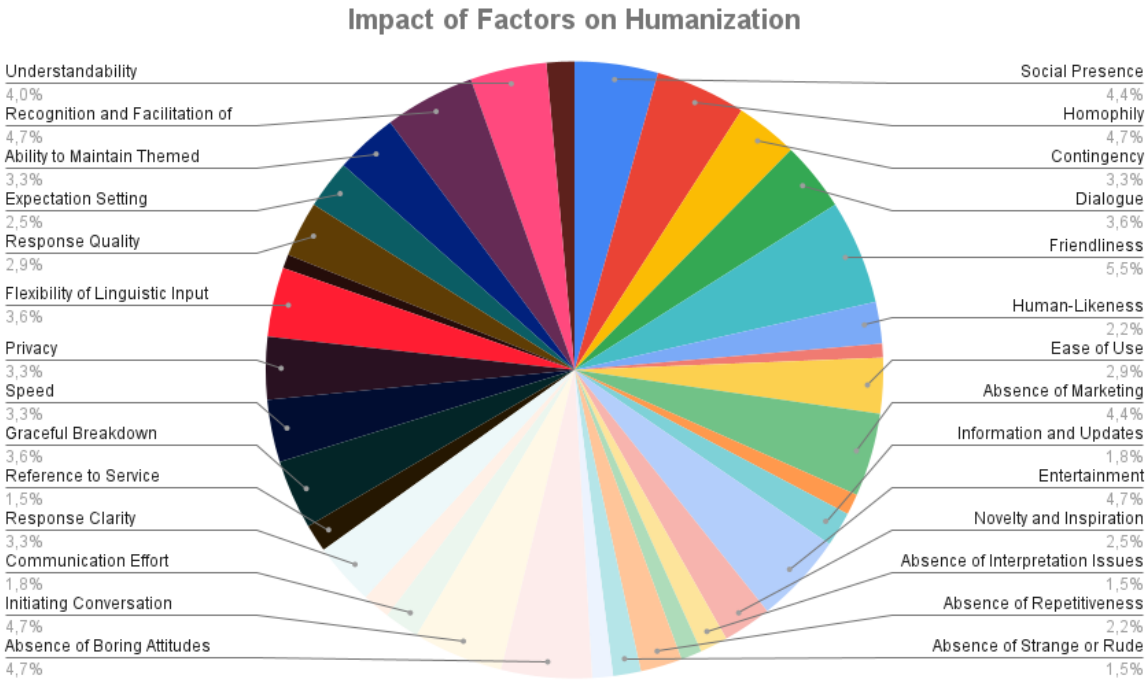

Figure 5.8: Impact of Impact Factors on Humanization according to the Weighting Questionnaire data for Group 04 - Someone who Understands You

The second artifact is Table 5.1, which presents a ranking of impact factors according to the level of relevance for the conception of humanization. Impact factors higher in the table are the most relevant (marked in green), the factors further down are the least relevant (marked in red), and factors with intermediate relevance are marked in yellow. As the Manhattan Normalization Absolute Value (Equation 5.2) is used here, no value is negative, but those used in the final metric can be negative. In the case of group 04, the *Predictability*, *Information and Updates*, *Reference to Service*, and *Credibility* impact factors had negative weights.

5.3.5.2 Results of the First Application of the Perceptions Questionnaire

The first application of the Perceptions Questionnaire occurred during the evaluation of the first high prototype (Subsection 5.2.1.9). After this application, we acquired the first metric in humanization, where group 04 obtained a score of 0.607 (that is, the chatbot had 60.7% of the perfect humanization for the context), and then we presented two acquired artifacts.

Figure 5.9 shows a pie chart of how much the individual performances of the impact factors influence the achievement of the final humanization metric. Here we have that the *Absence of Marketing* factor (10.5%) influenced the humanization metric more than the *Ability to Maintain Themed Discussion* factor (0.9%).

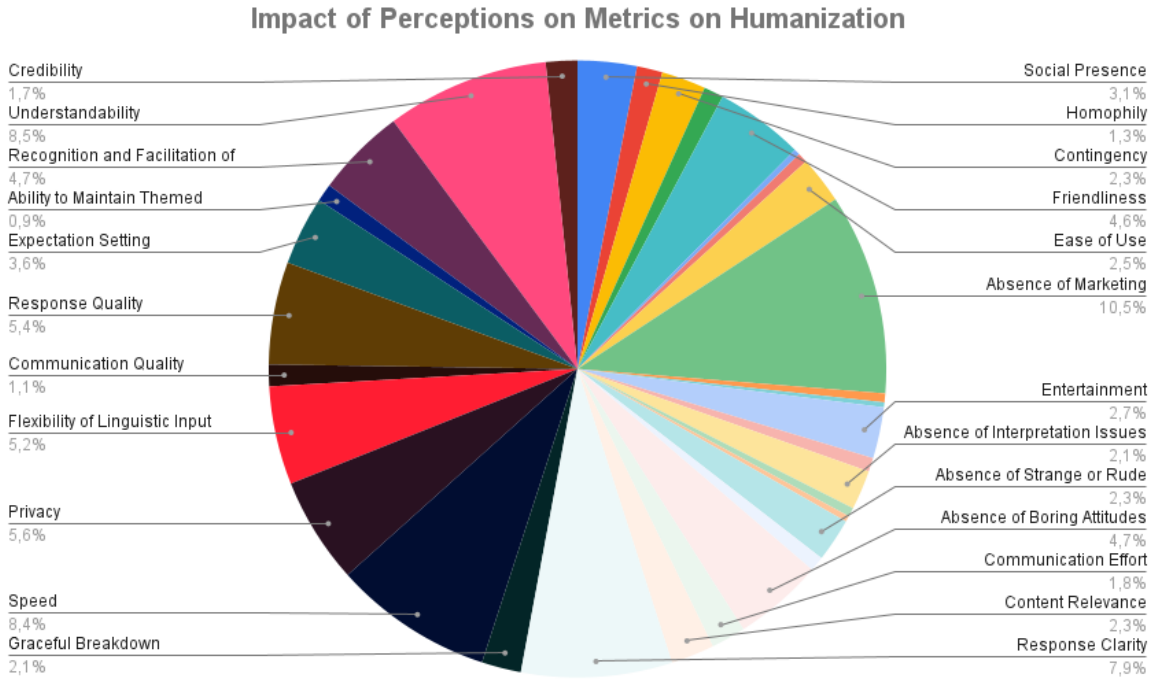

Figure 5.9: Impact of Individual Scores for each Impact Factor in the calculation of the Final Humanization Metric of Group 04

The bar graph of Figure 5.10 complements and confirms the results of the graph of Figure 5.9. Here, the scores obtained by each impact factor in the first round of

Table 5.1: Weight Ranking for Group 04

Weight Ranking	
Impact Factors	**Weightings**
Friendliness	0,05455
Homophily	0,04727
Entertainment	0,04727
Absence of Boring Attitudes	0,04727
Initiating Conversation	0,04727
Recognition and Facilitation of Users' Goal and Intent	0,04727
Social Presence	0,04364
Absence of Marketing	0,04364
Understandability	0,04000
Dialogue	0,03636
Graceful Breakdown	0,03636
Flexibility of Linguistic Input	0,03636
Contingency	0,03273
Response Clarity	0,03273
Speed	0,03273
Privacy	0,03273
Ability to Maintain Themed Discussion	0,03273
Ease of Use	0,02909
Response Quality	0,02909
Novelty and Inspiration	0,02545
Expectation Setting	0,02545
Human-Likeness	0,02182
Absence of Repetitiveness	0,02182
Communication Effort	0,01818
Absence of Interpretation Issues	0,01455
Absence of Strange or Rude Responses	0,01455
Content Relevance	0,01455
Help and Assistance	0,01091
Absence of Inability to Help	0,01091
Absence of Unwanted Events	0,01091
Communication Quality	0,00727
Expertise	0,00000
Predictability	-0,00727
Reference to Service	-0,01455
Credibility	-0,01455
Information and Updates	-0,01818

humanization assessment are presented. Indeed *Absence of Marketing* got the best result (75.18%), but *Ability to Maintain Themed Discussion* (47.78%) was not the worst result, but *Contingency* (44.44%). This is because *Contingency* has greater relevance in the idea of humanization than *Ability to Maintain Themed Discussion*, but received a lower grade.

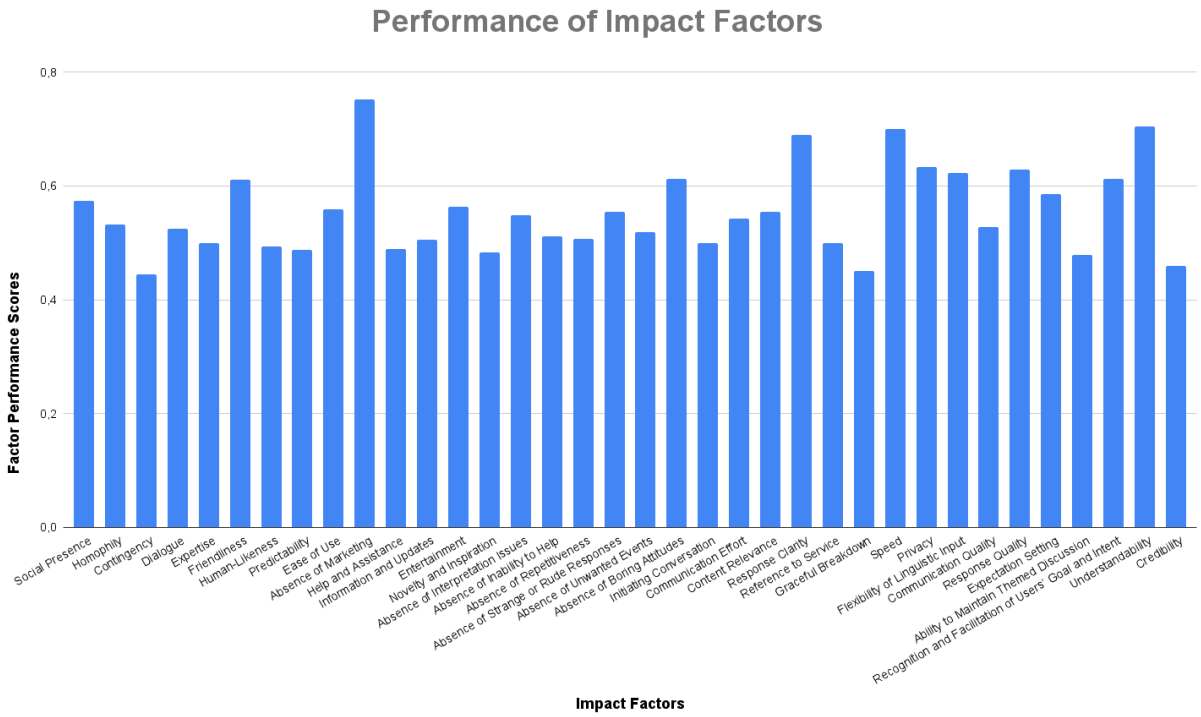

Figure 5.10: Final Performance of each Impact Factor in the High Prototype of Group 04

5.3.5.3 Results of the Second Application of the Perceptions Questionnaire

The second application of the Perceptions Questionnaire, which takes place during the reprototyping assessment (Subsection 5.2.2), was an optional phase, but group 04 participated. It is essential to understand how the chatbot's evolutionary process influences humanization metrics. While group 04 achieved a score of 0.607 in the first round, it has already achieved a score of 0.683 in the second round of evaluation.

As in the first application of the Perceptions Questionnaire, Figure 5.11 shows a pie chart of how much the individual performances of the impact factors influence the achievement of the final humanization metric. The bar graph of Figure 5.12 presents the scores obtained by each impact factor in the second round of the humanization assessment.

5.3.5.4 Evolution with the Use of Methodology

Finally, we arrive at the demonstration of the evolution of the chatbot proposed by group 04. The first clear sign of the evolution of the prototypes is the humanization metrics obtained, being 0.607 in the first evaluation round and 0.683 in the second evaluation round.

However, in addition to the general humanization metrics, the evolution is visible in the bar graph of Figure 5.13, where the scores of only the 12 factors most relevant

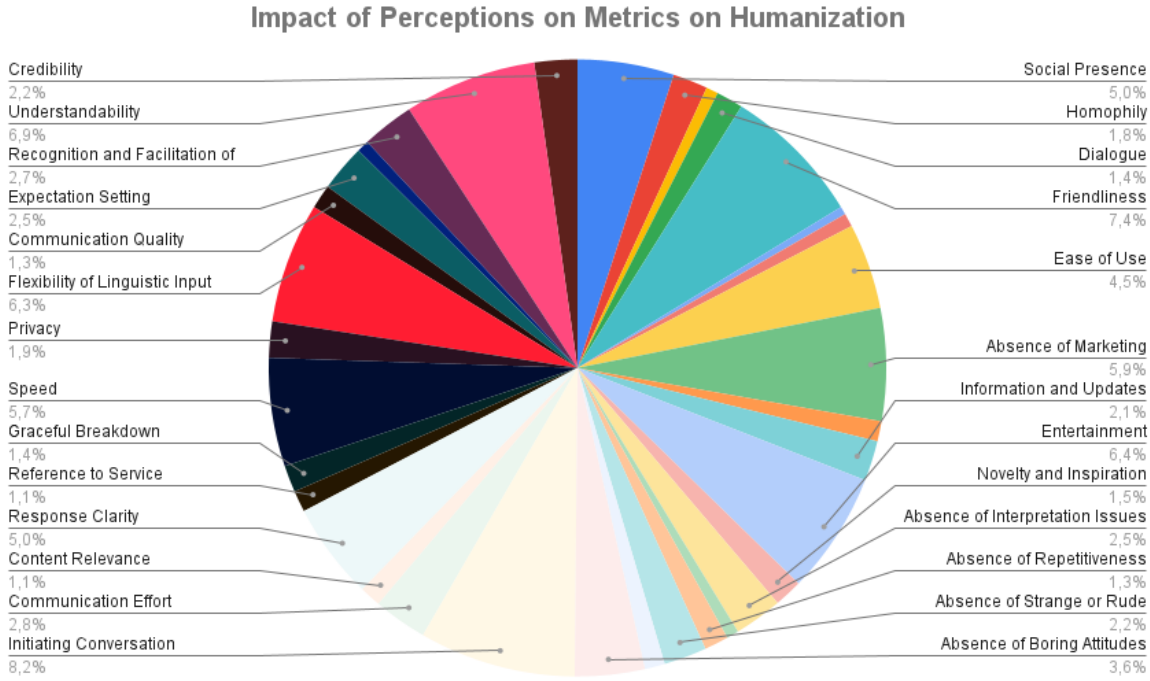

Figure 5.11: Impact of Individual Scores for each Impact Factor in the calculation of the Final Humanization Metric of Group 04

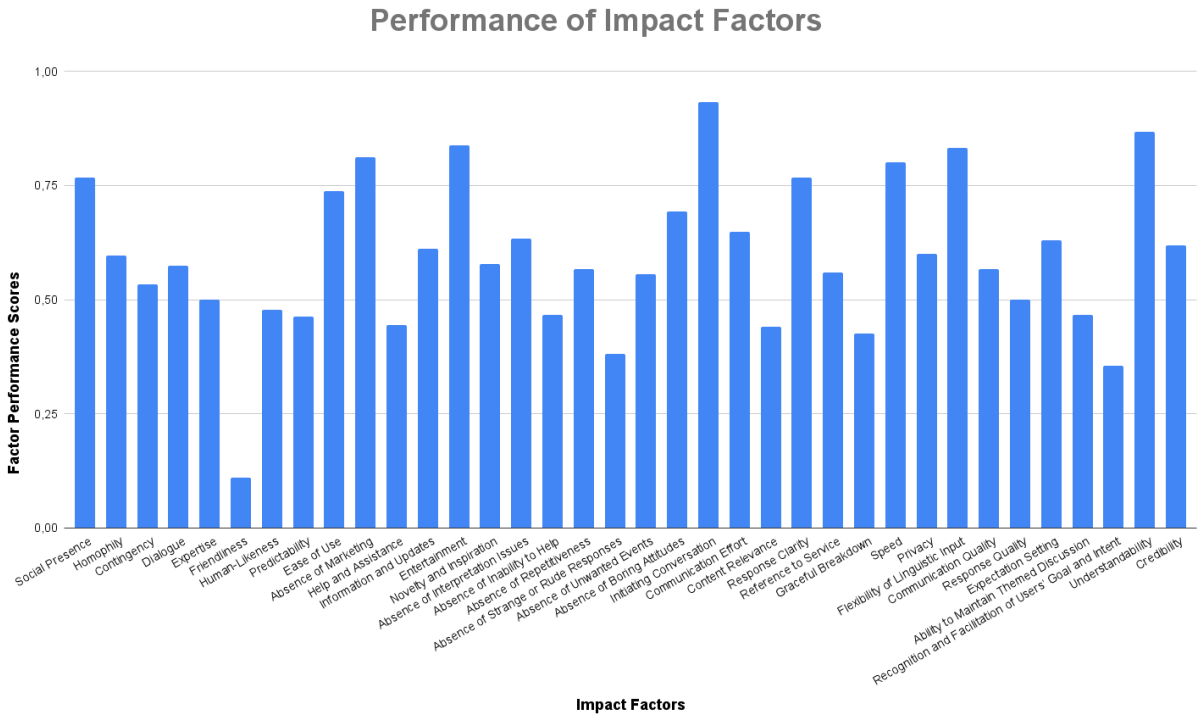

Figure 5.12: Final Performance of each Impact Factor in the High Prototype of Group 04

in constructing the idea of humanity were compared. In nine of the 12 factors, the performance improved (the vast majority of the factors analyzed), and the performance worsened in only 3 of the 12 factors.

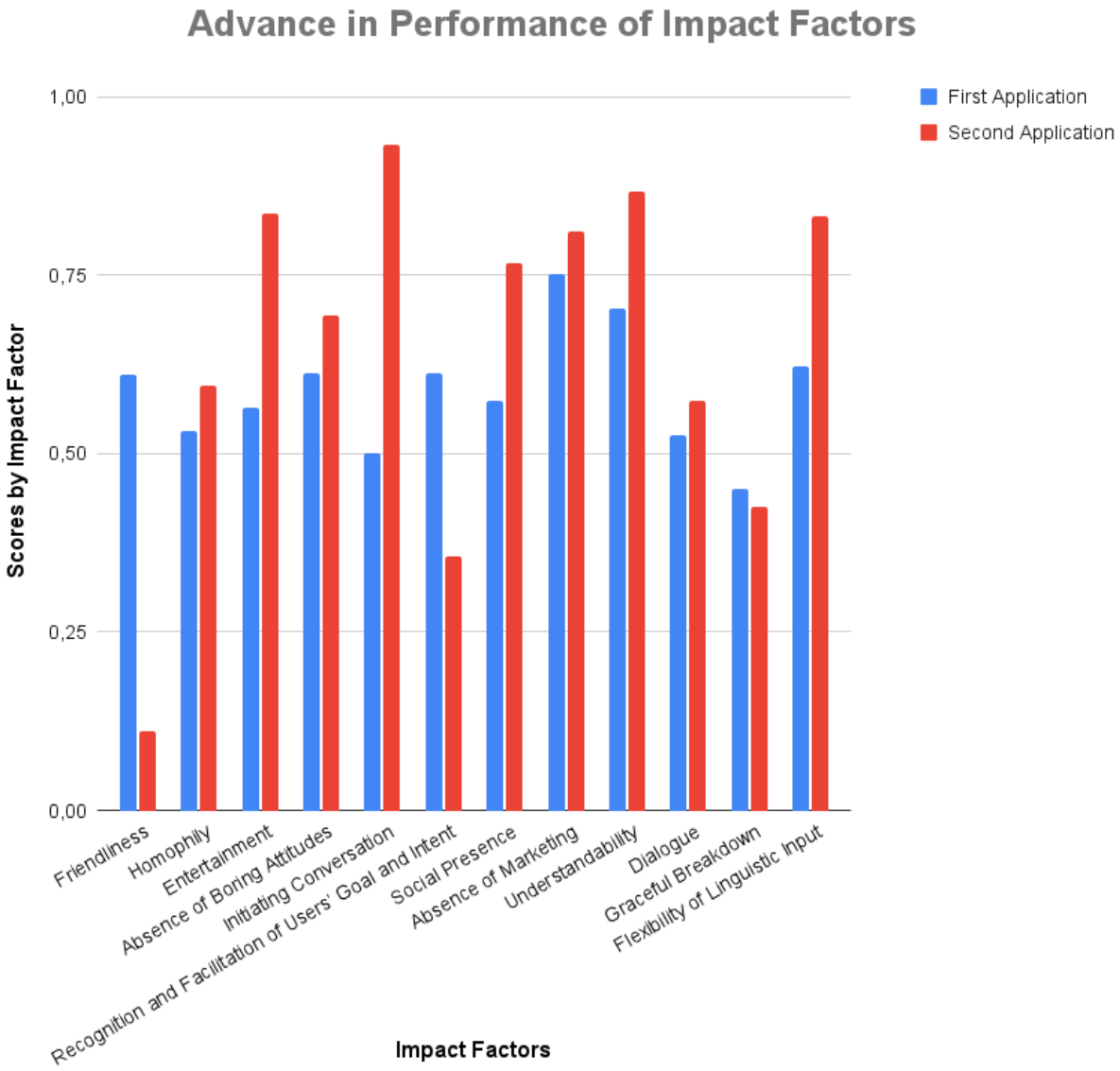

Figure 5.13: Advance Performance of Impact Factors in the High Prototype of Group 04

All humanization assessments of group 04, along with the grades acquired, the pre-processed grades, and the calculations proposed in the methodology, are presented in the following link.

5.4 Discussion

We can discuss the results mainly by looking at the weighting data. In the methodology, we propose that the weighting results are the main differentiator, in which chatbots are described by what is desired as a final result. PCA (Principal Component Analysis), a dimensionality reduction method, was applied to analyze how the weighting data

behaves more deeply, as we can see in image 5.14. We could observe some groupings according to the proposed chatbot theme. For example, circled in red, we have chatbots for recommendations. Meanwhile, circled in blue, we have chatbots for academic-professional services, which are next to each other. Surrounded in orange are the chatbots for the local target audience. In purple, chatbots for mental health are added in lighter blue by chatbots for general assistance. Finally, in pink, we have a specific chatbot developed for the hospital public, which has a very different purpose from the others. Furthermore, as chatbots have results close to weighting according to context, we can prove how much context influences the conception of humanity.

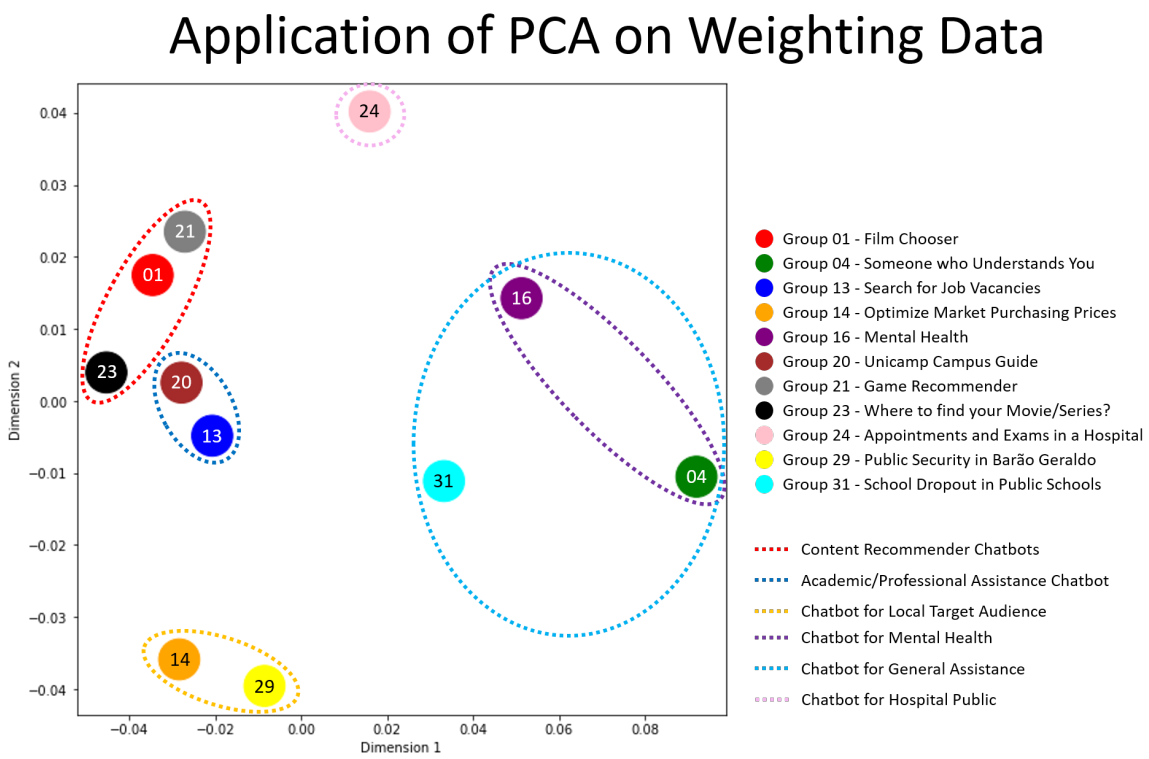

Figure 5.14: Application of PCA on Pre-Processed and Normalized Weighting Data

We observe in Table 5.2 that certain impact factors vary significantly from others, whereas some impact factors vary by medium and others by a little. Here it is suggested that impact factors that vary slightly should always be treated with the same standardized weight. On the other hand, impact factors that vary in a median way can be evaluated only in cases where it is necessary or relevant to the chatbot context. We suggested that the weighting questionnaire is always applied to verify the weights of impact factors with high variance. We also acquired the average of the weights for each of the impact factors. This average weighting can be used in cases where the weighting questionnaire is not desired for an adaptive assessment and the exclusive application of the perception questionnaire is desired.

Table 5.2: Variance and Average by Impact Factor in Weightings

Impact Factor	Variance	Average
Entertainment	8.628×10^{-4}	-0.0132
Friendliness	7.729×10^{-4}	0.0131
Information and Updates	6.421×10^{-4}	0.0235
Homophily	5.801×10^{-4}	0.0065
Social Presence	4.788×10^{-4}	0.0154
Predictability	3.916×10^{-4}	0.0183
Dialogue	3.854×10^{-4}	0.0104
Novelty and Inspiration	3.845×10^{-4}	-0.0029
Human-Likeness	3.582×10^{-4}	0.0008
Expertise	3.291×10^{-4}	0.0301
Credibility	2.978×10^{-4}	0.0356
Absence of Marketing	2.952×10^{-4}	0.0247
Privacy	2.887×10^{-4}	0.0185
Reference to Service	2.561×10^{-4}	0.0265
Help and Assistance	1.877×10^{-4}	0.0249
Ability to Maintain Themed Discussion	1.814×10^{-4}	0.0097
Content Relevance	1.557×10^{-4}	0.0365
Speed	1.381×10^{-4}	0.0269
Graceful Breakdown	1.344×10^{-4}	0.0246
Communication Quality	1.278×10^{-4}	0.0338
Absence of Repetitiveness	1.123×10^{-4}	0.0125
Flexibility of Linguistic Input	1.040×10^{-4}	0.0332
Contingency	1.030×10^{-4}	0.0238
Recognition and Facilitation of Users' Goal and Intent	1.006×10^{-4}	0.0358
Initiating Conversation	9.949×10^{-5}	0.0291
Absence of Boring Attitudes	9.402×10^{-5}	0.0277
Absence of Strange or Rude Responses	8.842×10^{-5}	0.0317
Absence of Interpretation Issues	8.689×10^{-5}	0.0351
Absence of Unwanted Events	8.262×10^{-5}	0.0287
Absence of Inability to Help	8.180×10^{-5}	0.0320
Ease of Use	6.605×10^{-5}	0.0416
Expectation Setting	6.300×10^{-5}	0.0200
Response Quality	5.340×10^{-5}	0.0370
Communication Effort	4.045×10^{-5}	0.0294
Understandability	2.107×10^{-5}	0.0385
Response Clarity	2.100×10^{-5}	0.0390
Total Average	**2.352×10^{-4}**	**0.0239**

5.5 Conclusion

We conclude two main results with this case study. The first is that the weighting results depend a lot on the context in which the chatbot is inserted, that is, what type of service it will perform, which audience it will serve, and, mainly, what is desirable for this chatbot. The weights are strongly dependent on the context in which the chatbot is inserted (cf. Figure 5.14). Another conclusion is that the impact factors have different variations and that, therefore, there are impact factors that vary little from one context to another. There are impact factors that almost always vary according to the context in which it is inserted. Thus, we can also have as another methodology application result, in addition to an experiment proving the advantages of using the presented methodology, the average weighting of the impact factors. Such average weighting can be used in cases where the design team wants to apply only the Perceptions Questionnaire.

Chapter 6

Conclusion

We conclude that our chatbot humanization assessment methodology was an innovative method, especially when it comes to evaluating qualitative factors. Here we are dealing with state of the art in humanization evaluation for chatbots since it is an adaptive approach. The different contexts of chatbots and adaptability include the different types of target audiences served by such chatbots. In addition, a wide variety of impact factors on humanization were investigated, and this analysis of such impact factors, as well as the results in humanization acquired with each new cycle of evolution of a chatbot, help in design decisions, especially in what concerns what needs to be improved in the chatbot being analyzed. The main advantage of this methodology is that it seeks the evolutionary and adaptive development of chatbots according to the context in which they are inserted and the target audiences they will serve. Prior to the method presented here, factors related to humanization were only evaluated qualitatively. Here we are proposing and demonstrating a quantitative way of evaluating which impact factors influence the perception of humanity that real human beings have, which can be quantified in the evolution of chatbots more effectively with each new development iteration, that is, with each new versioning.

6.1 Contributions

This work has as contributions the set of impact factors obtained through the literature, as well as an adaptive methodology for evaluating chatbots, which can be adapted even to new sets of impact factors. An overall assessment of which impact factors are more or less relevant to be investigated in chatbots was also acquired, and what would be the average weighting for each impact factor.

6.2 Limitations

The main limitation of this method is the number of impact factors, which can be better investigated or summarized according to the context of each chatbot. We can also see that the rating depends heavily on the audience, and the audience may often not have a beneficial interest in rating chatbots. In the use case presented by this paper, the participants were

highly interested in the evaluation results and collaborating with colleagues, as this cross-collaboration benefited everyone. However, in all cases, the evaluators are not really motivated and adequately motivated and mobilized. It is essential that the target audience that evaluates chatbots is involved in this and feels benefited by the results acquired through the evaluation. These results make chatbots increasingly adapted and pleasant to the target audience. That is why it is also highly recommended that the target audience of chatbots in reviews is the same audience participating in the reviews.

6.3 Future Work

In future work, we aim to investigate further the impact factors mentioned here. Algorithms that can punctually improve each of the specific impact factors will also be proposed, mainly algorithms for augmenting dialogues that can be used during the training of state-of-the-art chatbots. It is also foreseen the application of this methodology in other contexts, mainly in industrial contexts. Finally, the use of this methodology in other human-computer interfaces will be investigated to analyze whether this methodology is effective in the general evaluation of humanity in software.

Bibliography

[1] Martin Adam, Michael Wessel, Alexander Benlian, et al. Ai-based chatbots in customer service and their effects on user compliance. *Electronic Markets*, 9(2):204, 2020.

[2] Eleni Adamopoulou and Lefteris Moussiades. An overview of chatbot technology. In *IFIP International Conference on Artificial Intelligence Applications and Innovations*, pages 373–383. Springer, 2020.

[3] Shaghaf Alzorba, Christian Günther, Nicolae Popovici, and Christiane Tammer. A new algorithm for solving planar multiobjective location problems involving the manhattan norm. *European Journal of Operational Research*, 258(1):35–46, 2017.

[4] Divyaa Balaji. Assessing user satisfaction with information chatbots: a preliminary investigation. Master's thesis, University of Twente, 2019.

[5] Himanshu Bansal and Rizwan Khan. A review paper on human computer interaction. *International Journals of Advanced Research in Computer Science and Software Engineering*, 8:53–56, 2018.

[6] Tom Bocklisch, Joey Faulkner, Nick Pawlowski, and Alan Nichol. Rasa: Open source language understanding and dialogue management. *arXiv preprint arXiv:1712.05181*, 2017.

[7] Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Inigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. Multiwoz–a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. *arXiv preprint arXiv:1810.00278*, 2018.

[8] James T Croasmun and Lee Ostrom. Using likert-type scales in the social sciences. *Journal of Adult Education*, 40(1):19–22, 2011.

[9] Jader MC De Sa. Zécarioca: A framework for creating end-to-end chatbots. Master's thesis, State University of Campinas, 2021.

[10] Amar Kumar Dey, Manisha Sharma, and MR Meshram. An analysis of leaf chlorophyll measurement method using chlorophyll meter and image processing technique. *Procedia Computer Science*, 85:286–292, 2016.

[11] Diane Edmondson. Likert scales: A history. In *Proceedings of the Conference on Historical Analysis and Research in Marketing*, volume 12, pages 127–133, 2005.

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

[13] Edward W Forgy. Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *biometrics*, 21:768–769, 1965.

[14] Kiev Gama, Breno Alencar, Filipe Calegario, André Neves, and Pedro Alessio. A hackathon methodology for undergraduate course projects. In *2018 IEEE Frontiers in Education Conference (FIE)*, pages 1–9. IEEE, 2018.

[15] Azad-Asnafor Gelici. Change in digital business: Chatbots and their significance on online purchasing. B.S. thesis, University of Twente, 2020.

[16] Ulrich Gnewuch, Stefan Morana, Marc Adam, and Alexander Maedche. Faster is not always better: understanding the effect of dynamic response delays in human-chatbot interaction. *AIS Electronic Library (AISeL)*, 2018.

[17] Eun Go and S Shyam Sundar. Humanizing chatbots: The effects of visual, identity and conversational cues on humanness perceptions. *Computers in Human Behavior*, 97:304–316, 2019.

[18] Jahnvi Gupta, Vinay Singh, and Ish Kumar. Florence-a health care chatbot. In *2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS)*, volume 1, pages 504–508. IEEE, 2021.

[19] Nick Haslam, Steve Loughnan, and Elise Holland. The psychology of humanness. *Objectification and (de) humanization*, pages 25–51, 2013.

[20] Geert Hofstede. Culture and organizations. *International studies of management & organization*, 10(4):15–41, 1980.

[21] Samuel Holmes, Anne Moorhead, Raymond Bond, Huiru Zheng, Vivien Coates, and Michael McTear. Usability testing of a healthcare chatbot: Can we use conventional methods to assess conversational user interfaces? In *Proceedings of the 31st European Conference on Cognitive Ergonomics*, pages 207–214, 2019.

[22] Ankur Joshi, Saket Kale, Satish Chandel, and D Kumar Pal. Likert scale: Explored and explained. *British Journal of Applied Science & Technology*, 7(4):396, 2015.

[23] Rohan Kar and Rishin Haldar. Applying chatbots to the internet of things: Opportunities and architectural elements. *arXiv preprint arXiv:1611.03799*, 2016.

[24] Anirudh Khanna, Bishwajeet Pandey, Kushagra Vashishta, Kartik Kalia, Bhale Pradeepkumar, and Teerath Das. A study of today's ai through chatbots and rediscovery of machine intelligence. *International Journal of u-and e-Service, Science and Technology*, 8(7):277–284, 2015.

[25] Diksha Khurana, Aditya Koli, Kiran Khatter, and Sukhdev Singh. Natural language processing: State of the art. *Current Trends and Challenges*, 8, 2017.

[26] Haridimos Kondylakis, Dimitrios Tsirigotakis, Giorgos Fragkiadakis, Emmanouela Panteri, Alexandros Papadakis, Alexandros Fragkakis, Eleytherios Tzagkarakis, Ioannis Rallis, Zacharias Saridakis, Apostolos Trampas, et al. R2d2: A dbpedia chatbot using triple-pattern like queries. *Algorithms*, 13(9):217, 2020.

[27] J Kříž. Chatbot for laundry and dry cleaning service. *Masaryk University*, 2017.

[28] Carlos Alberto Joia Lazarin and Leonelo Dell Anhol Almeida. Distributed participatory design web-based groupware: gathering requirements through braindraw. In *Proceedings of the 15th Brazilian Symposium on Human Factors in Computing Systems*, pages 1–10, 2016.

[29] Rohit Binu Mathew, Sandra Varghese, Sera Elsa Joy, and Swanthana Susan Alex. Chatbot for disease prediction and treatment recommendation using machine learning. In *2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI)*, pages 851–856. IEEE, 2019.

[30] S McLeod. Likert scale. simplypsychology. org, 2014.

[31] Geoffrey E Mills and Lorraine R Gay. *Educational research: Competencies for analysis and applications*. ERIC, 2019.

[32] Gregory Mone. The edge of the uncanny. *Communications of the ACM*, 59(9):17–19, 2016.

[33] Prakash M Nadkarni, Lucila Ohno-Machado, and Wendy W Chapman. Natural language processing: an introduction. *Journal of the American Medical Informatics Association*, 18(5):544–551, 2011.

[34] Nishad Nawaz and Anjali Mary Gomes. Artificial intelligence chatbots are new recruiters. *IJACSA) International Journal of Advanced Computer Science and Applications*, 10(9), 2019.

[35] Andreea Niculescu, Betsy van Dijk, Anton Nijholt, Haizhou Li, and Swee Lan See. Making social robots more attractive: the effects of voice pitch, humor and empathy. *International journal of social robotics*, 5(2):171–191, 2013.

[36] Cecilie Bertinussen Nordheim, Asbjørn Følstad, and Cato Alexander Bjørkli. An initial model of trust in chatbots for customer service—findings from a questionnaire study. *Interacting with Computers*, 31(3):317–335, 2019.

[37] Kumud Park. Preventive and social medicine, 2013.

[38] Leo Natan Paschoal, Lucas Lagoa Nogueira, and Patricia Mariotto Mozzaquatro Chicon. Agentes conversacionais pedagógicos: Uma discussão inicial sobre conceitos, estratégias de desenvolvimento e oportunidades de pesquisa. *DIGITALIZAÇÃO DA EDUCAÇÃO: DESAFIOS E ESTRATÉGIAS PARA A EDUCAÇÃO DA GERAÇÃO CONECTADA*, page 23, 2020.

[39] Luiz Pasquali. Psychometrics. *Revista da Escola de Enfermagem da USP*, 43:992–999, 2009.

[40] Dijana Peras. Chatbot evaluation metrics. *Economic and Social Development: Book of Proceedings*, pages 89–97, 2018.

[41] Kiran Ramesh, Surya Ravishankaran, Abhishek Joshi, and K Chandrasekaran. A survey of design techniques for conversational agents. In *International conference on information, communication and computing technology*, pages 336–350. Springer, 2017.

[42] Er B Ravinder and Dr AB Saraswathi. Literature review of cronbachalphacoefficient (a) and mcdonald's omega coefficient (ω). *European Journal of Molecular & Clinical Medicine*, 7(6):2943–2949, 2020.

[43] Byron Reeves and Clifford Nass. *The media equation: How people treat computers, television, and new media like real people*. Cambridge university press Cambridge, United Kingdom, 1996.

[44] William Revelle. Hierarchical cluster analysis and the internal structure of tests. *Multivariate Behavioral Research*, 14(1):57–74, 1979.

[45] Jungwook Rhim, Minji Kwak, Yeaeun Gong, and Gahgene Gweon. Application of humanization to survey chatbots: Change in chatbot perception, interaction experience, and survey data quality. *Computers in Human Behavior*, 126:107034, 2022.

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

[47] Heung-Yeung Shum, Xiao-dong He, and Di Li. From eliza to xiaoice: challenges and opportunities with social chatbots. *Frontiers of Information Technology & Electronic Engineering*, 19(1):10–26, 2018.

[48] Yogesh Kumar Singh. *Fundamental of research methodology and statistics*. New Age International, 2006.

[49] Khai N Truong, Gillian R Hayes, and Gregory D Abowd. Storyboarding: an empirical determination of best practices and effective guidelines. In *Proceedings of the 6th conference on Designing Interactive systems*, pages 12–21, 2006.

[50] Tibert Verhagen, Jaap Van Nes, Frans Feldberg, and Willemijn Van Dolen. Virtual customer service agents: Using social presence and personalization to shape online service encounters. *Journal of Computer-Mediated Communication*, 19(3):529–545, 2014.

[51] David Westerman, Aaron C Cross, and Peter G Lindmark. I believe in a thing called bot: Perceptions of the humanness of "chatbots". *Communication Studies*, 70(3):295–312, 2019.

[52] Jingjun Xu, Izak Benbasat, and Ronald T Cenfetelli. Research note—the influences of online service technologies and task complexity on efficiency and personalization. *Information Systems Research*, 25(2):420–436, 2014.

[53] Jingjun David Xu. Retaining customers by utilizing technology-facilitated chat: Mitigating website anxiety and task complexity. *Information & Management*, 53(5):554–569, 2016.

[54] Yingzi Xu, 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.

[55] Yunyi Yang, Yunhao Li, and Xiaojun Quan. Ubar: Towards fully end-to-end task-oriented dialog systems with gpt-2. *arXiv preprint arXiv:2012.03539*, 2020.

[56] ZENVIA. Conheça os tipos de chatbots e como eles podem ajudar sua empresa. *[Website] Abruf [01. 04. 2021] unter: https://www.zenvia.com/blog/conheca-os-tipos-de-chatbots-atendimento/*, 2019.

[57] Melissa Zhang. *Speeding Up the Prototyping of Low-Fidelity User Interface Wireframes*. PhD thesis, Texas AM University, 2022.

[58] Richard E Zinbarg, William Revelle, Iftah Yovel, and Wen Li. Cronbach's α, revelle's β, and mcdonald's ω h: Their relations with each other and two alternative conceptualizations of reliability. *psychometrika*, 70(1):123–133, 2005.

Appendix A

Detailed Results in the MC750A Course

A.1 Group 01 - Film Chooser

- **Description:** *Chatbot that helps you choose a movie according to your taste.*
- **Final Value of the Metric in Humanization:** 0,816
- **Complete Results of the Humanization Assessment:** link

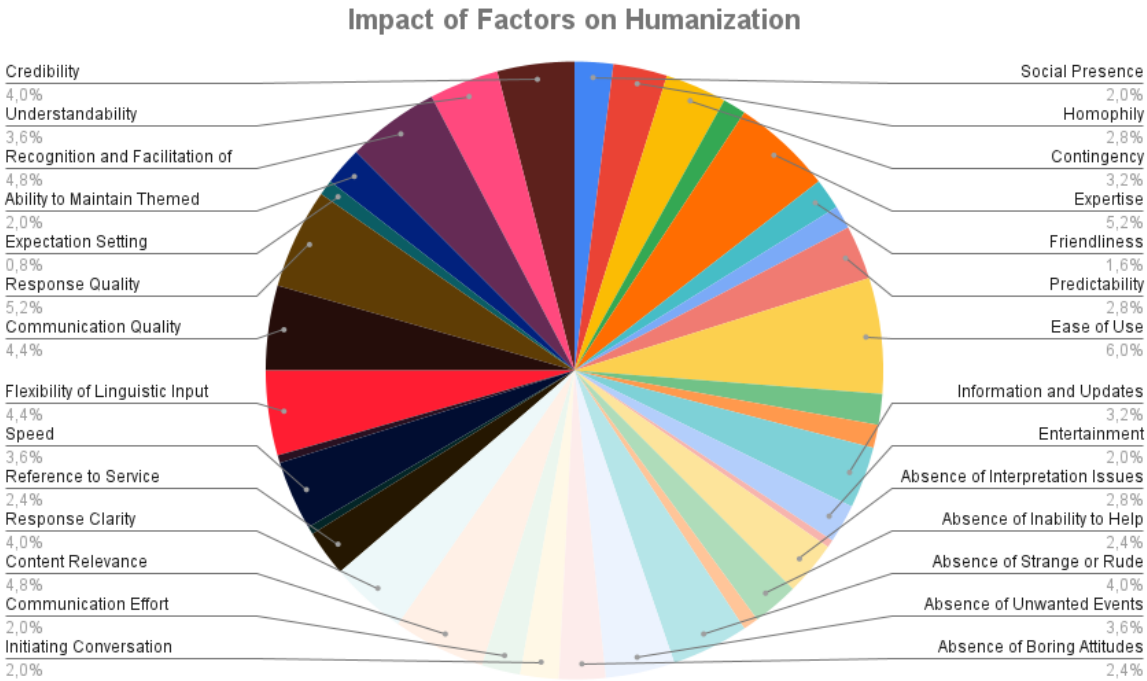

Figure A.1: Impact of Impact Factors on Humanization according to the Weighting Questionnaire data for Group 01 - Film Chooser

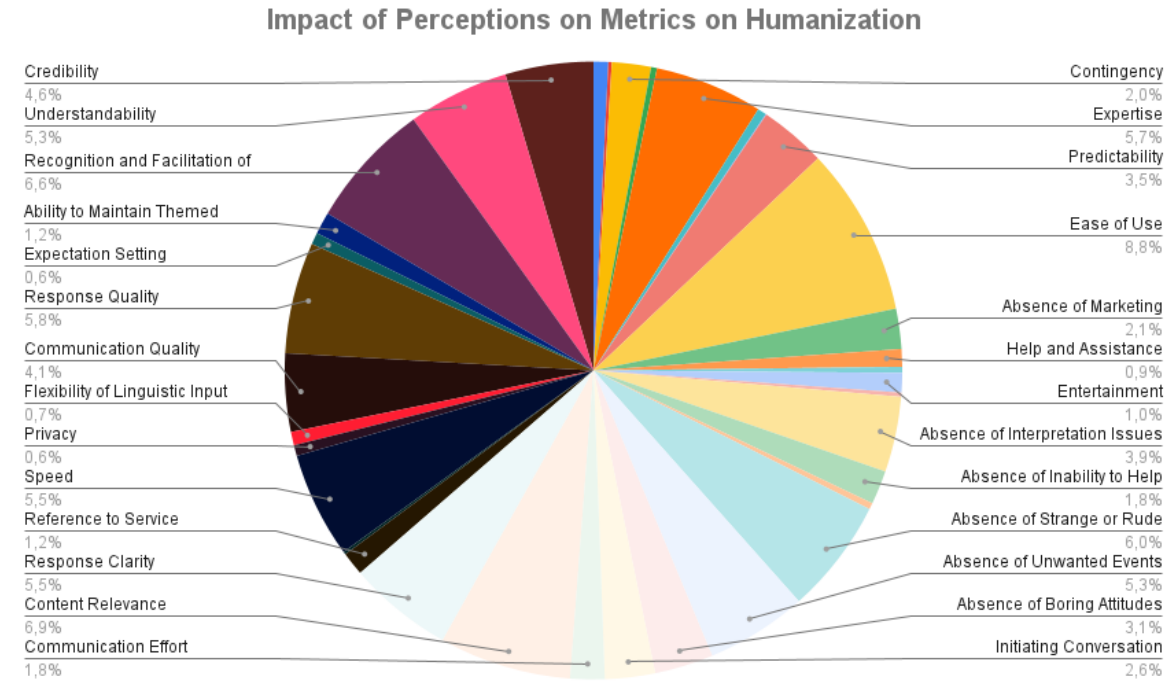

Figure A.2: Impact of Individual Scores for each Impact Factor in the calculation of the Final Humanization Metric of Group 01

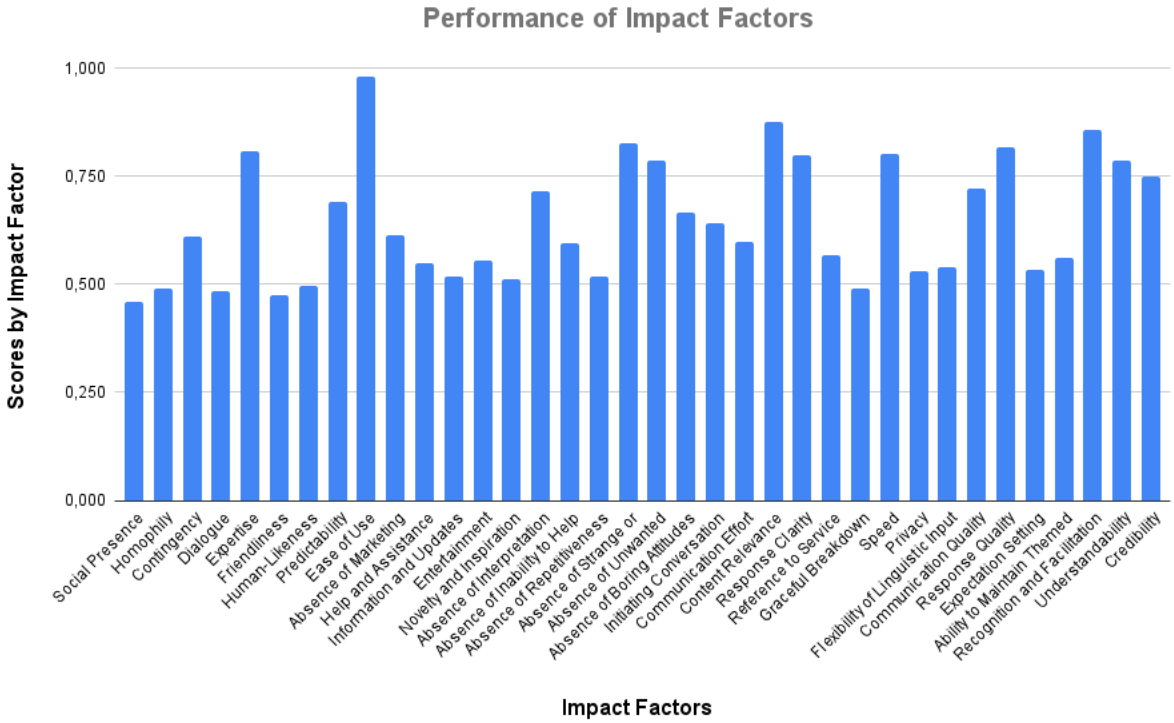

Figure A.3: Final Performance of each Impact Factor in the High Prototype of Group 01

A.2 Group 13 - Search for Job Vacancies

- **Description:** *Searching for an open job vacancy compatible with a university student's education, salary expectations, and place of residence can be time-consuming.*
- **Final Value of the Metric in Humanization:** 0,749
- **Complete Results of the Humanization Assessment:** link

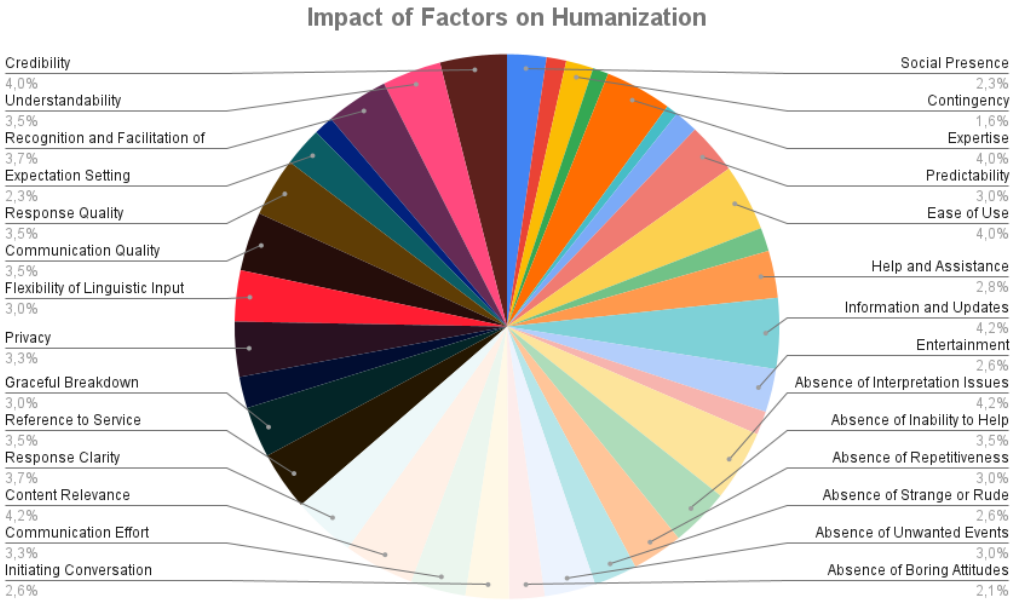

Figure A.4: Impact of Impact Factors on Humanization according to the Weighting Questionnaire data for Group 13 - Search for Job Vacancies

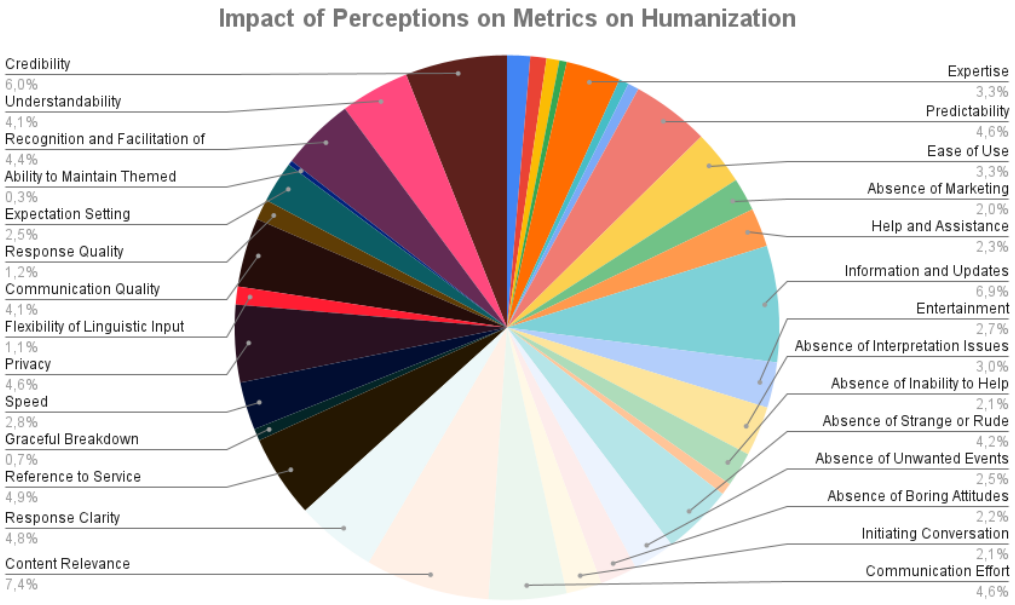

Figure A.5: Impact of Individual Scores for each Impact Factor in the calculation of the Final Humanization Metric of Group 13

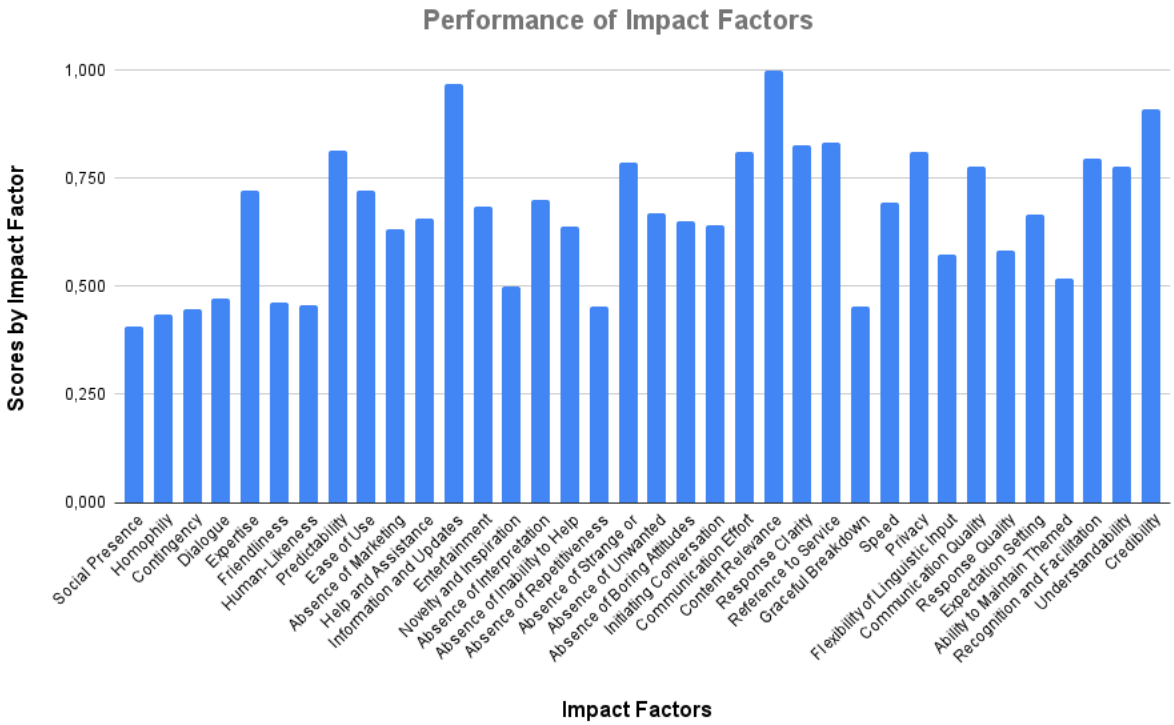

Figure A.6: Final Performance of each Impact Factor in the High Prototype of Group 13

A.3 Group 14 - Optimize Market Purchasing Prices

- **Description:** *Barão Geraldo has several markets, such as Pão de Açúcar, Pague Menos and Dalben. Any rational consumer would like, based on their shopping list, to be able to go to the market that minimizes their cost, but prices and promotions fluctuate a lot. The brute force solution would be to visit all markets each shopping day, but this is not feasible.*
- **Final Value of the Metric in Humanization:** 0,841
- **Complete Results of the Humanization Assessment:** link

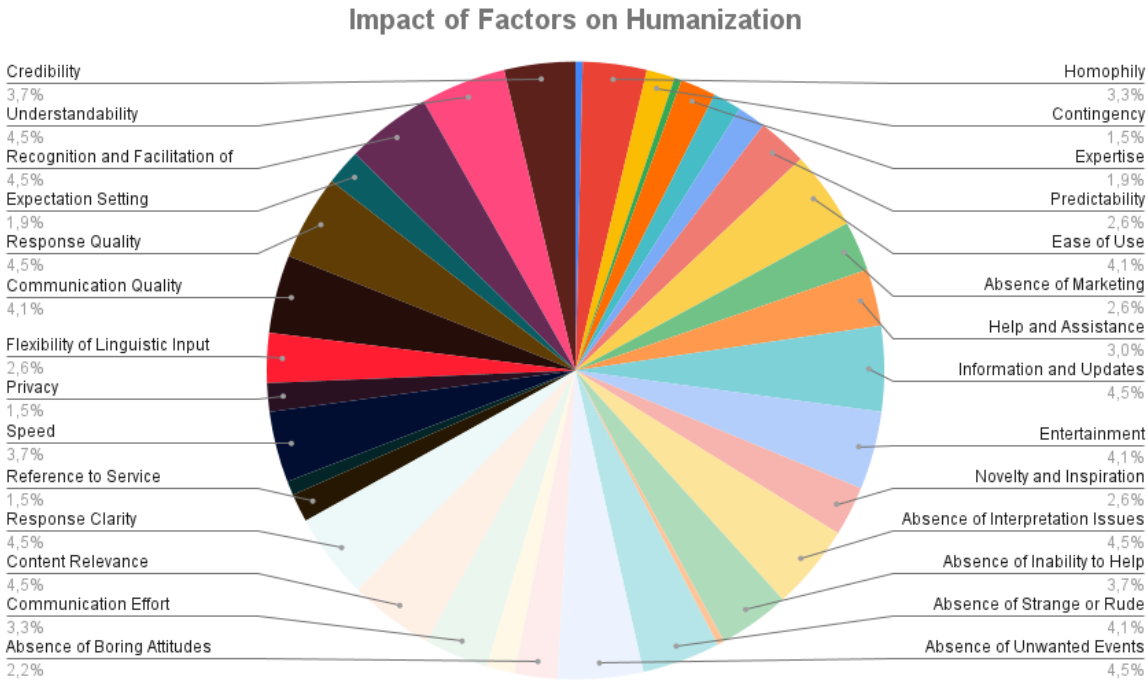

Figure A.7: Impact of Impact Factors on Humanization according to the Weighting Questionnaire data for Group 14 - Optimize Market Purchasing Prices

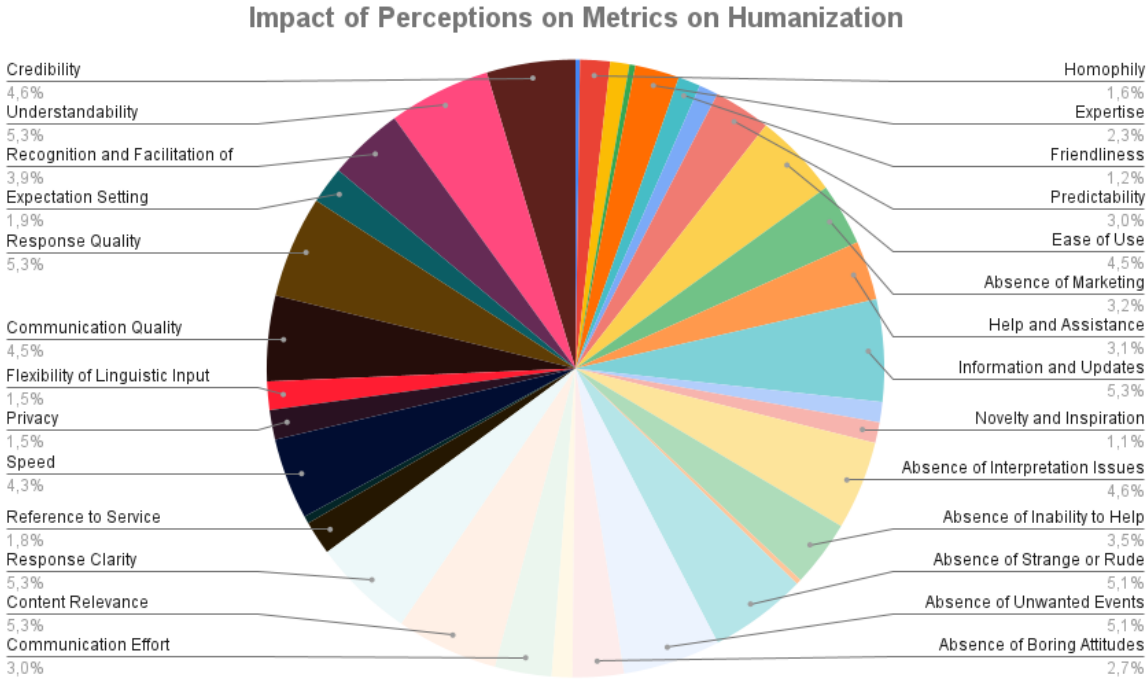

Figure A.8: Impact of Individual Scores for each Impact Factor in the calculation of the Final Humanization Metric of Group 14

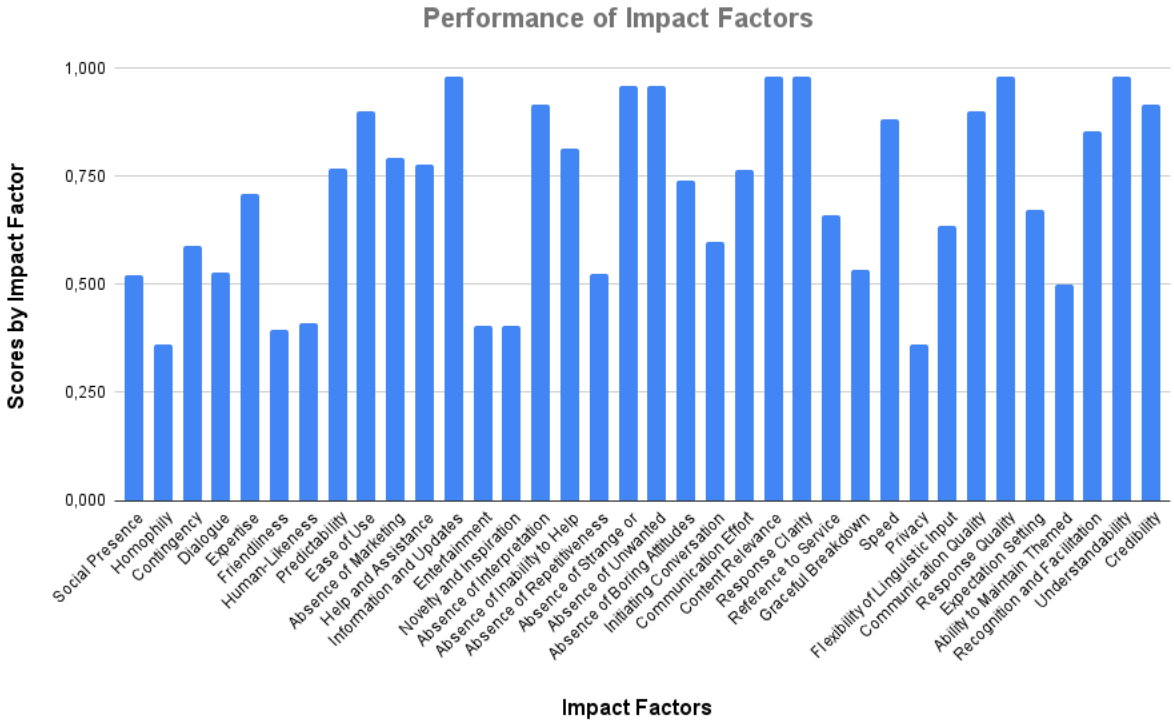

Figure A.9: Final Performance of each Impact Factor in the High Prototype of Group 14

A.4 Group 16 - Mental Health

- **Description:** *Many people cannot attend a psychologist or psychiatrist, or sometimes they are too resistant and proud to accept that they need professional help. Other times they do not even know they need help, not to mention those who already have more serious psychological problems, with depression itself making it difficult to leave the house or talk to someone to get help. The suggestion would be to create a chatbot that somehow helps these people. The idea came from the book "The Diary of Anne Frank", in which Anne uses the diary to talk to her imaginary friend Kitty, sharing some of the emotional weight resulting from the various stressful situations experienced in the circumstances of the war and also from the secret annex in which relations between the villagers was chaotic and toxic, especially for Anne. In this context, the writer always looked for Kitty when she was on the edge of stress, anger, sadness, depression, and loneliness. "Paper is more patient than men", writes Anne at various times. Imagine if, instead of having to go to Kitty when she felt the need, Kitty went to Anne to ask if everything was okay before things got to the point. I have also had periods of deep depression and have seen people around me suffer from it (common in engineering courses, unfortunately). Observing myself and other people, I came to believe that the ideal would be for psychological treatment to go to the people and not for people to go to the professionals because many emotional problems, especially depression, in addition to being silent, have symptoms that make it difficult to perform basic activities such as getting out of bed or eating correctly, imagine how difficult it can be to take the attitude of seeking help from an unknown*

professional. Depression attacks the tools we have to recover, like an emotional autoimmune disease. As a result, people rarely get better spontaneously, like the flu. Given all this, the idea would be to create a chatbot that would build an environment in which the user would feel comfortable and encouraged to share their emotions, guaranteeing that no one would have access to the conversation. Thus, the main objectives would be:

1. *Prevention and relief of depression and other emotional crises, such as anxiety, through conversations with the bot, in which the person would report what he felt he needed. An essential part of this point is that the bot would call the user, encouraging them to share their feelings at least once a week. The conversation would be almost a monologue as if the bot were a friendly ear that would already take a great deal of weight off our back when listening to the outburst.*
2. *To serve as a bridge, an intermediary, between the moment when the user does not have any professional support, either because he has some resistance to the idea of going to the psychologist/psychiatrist or because he does not know what type of professional to seek, or because he does not they are aware that they need help, either because they do not have enough courage or hope to do anything about it, among other reasons, it is the moment when the user starts to have professional help. For this, the bot would suggest, work and develop throughout the conversation the idea of seeking professional help, informing about what kind of specialist to look for, and indicating professionals and places where the person can find care near their location. Thus decreasing the "activation energy" of the users' reaction to seek help, in other words, facilitating and enabling the process.*
3. *Prevent suicides. Based on conversations and data collected from the user's device (it would have access to the cell phone's microphone as well), the bot, when it noticed signs that indicate a tendency to self-harm or suicide, would act to prevent the worst from happening, increasing the frequency of contact. with the user, changing the type of questions and messages he sends in order to develop a conversation aimed at valuing life, also signaling the CVV to get in touch (and not the user to contact the CVV).*

Note: The bot's name would be Kitty!

- **Final Value of the Metric in Humanization:** 0,886
- **Complete Results of the Humanization Assessment:** link

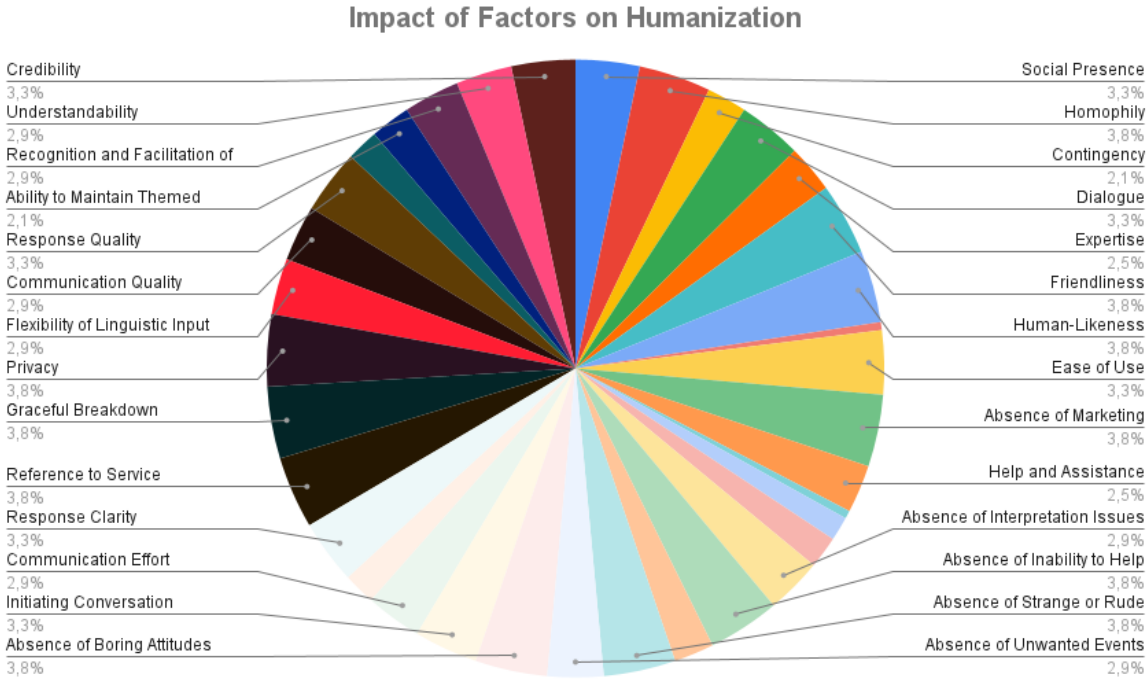

Figure A.10: Impact of Impact Factors on Humanization according to the Weighting Questionnaire data for Group 16 - Mental Health

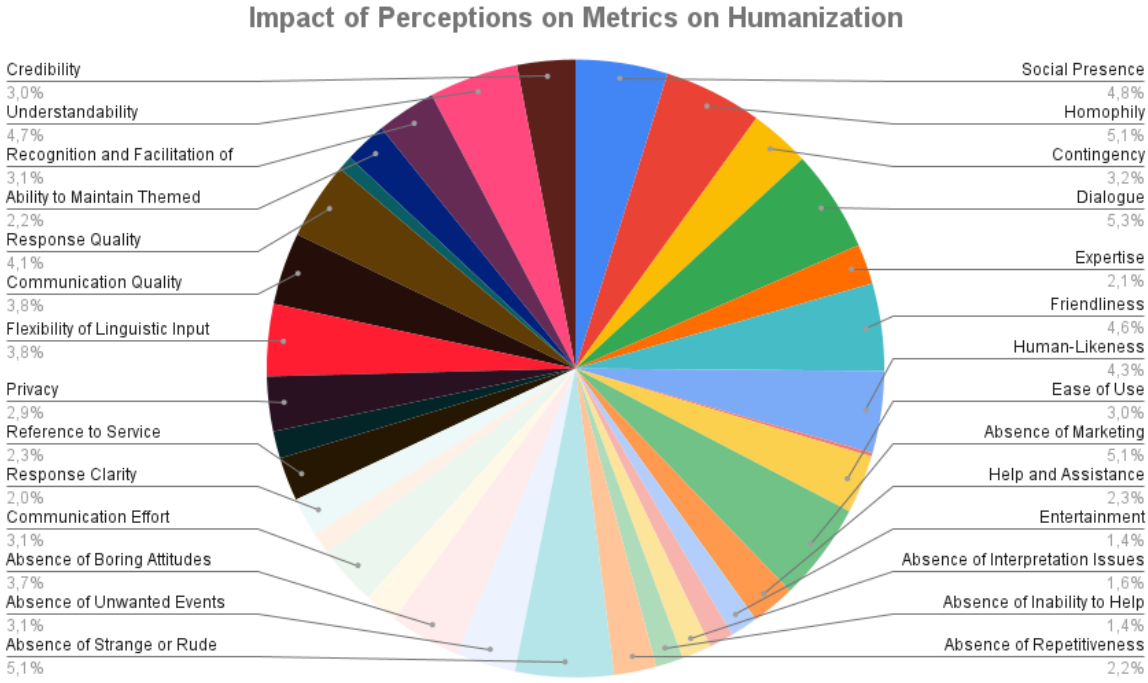

Figure A.11: Impact of Individual Scores for each Impact Factor in the calculation of the Final Humanization Metric of Group 16

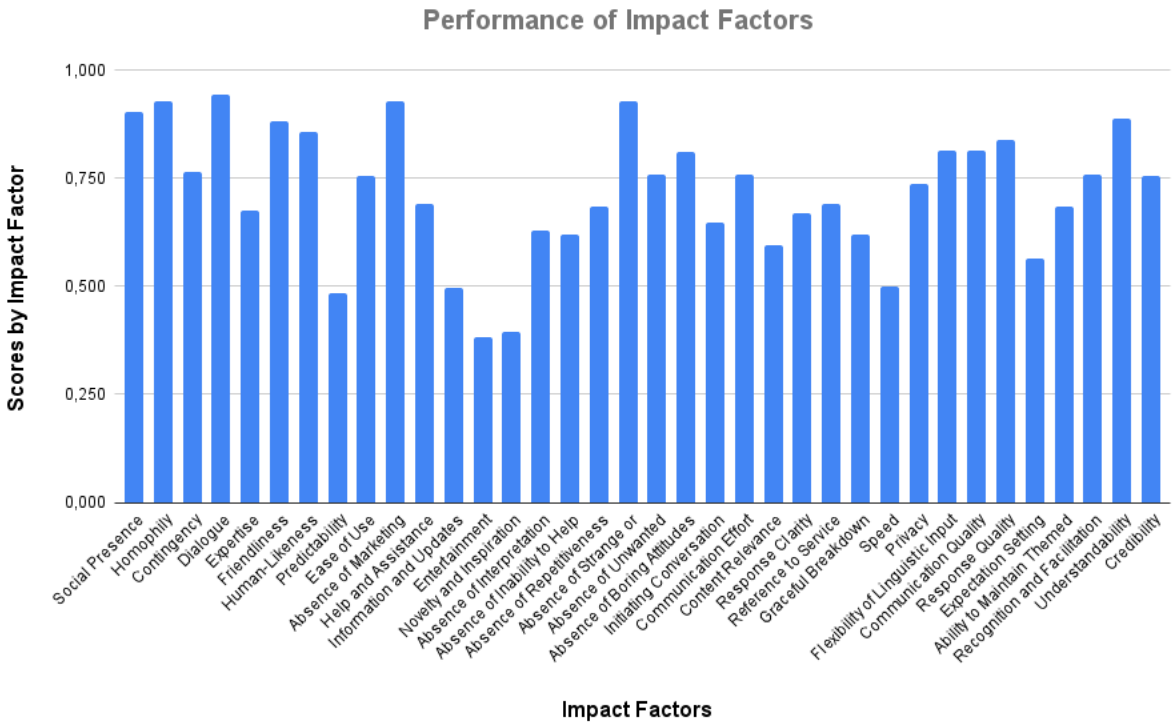

Figure A.12: Final Performance of each Impact Factor in the High Prototype of Group 16

A.5 Group 20 - Unicamp Campus Guide

- **Description:** *Joãozinho is a student at Unicamp, and this is his first year in person at the university, so he is lost. He does not know where his classrooms are, or institute hours, whether there are other restaurants outside the UK or where there are suitable places to sit and study- Anyway, he is lost with the Campus scheme. How to solve your problem?*
- **Final Value of the Metric in Humanization:** 0,844
- **Complete Results of the Humanization Assessment:** link

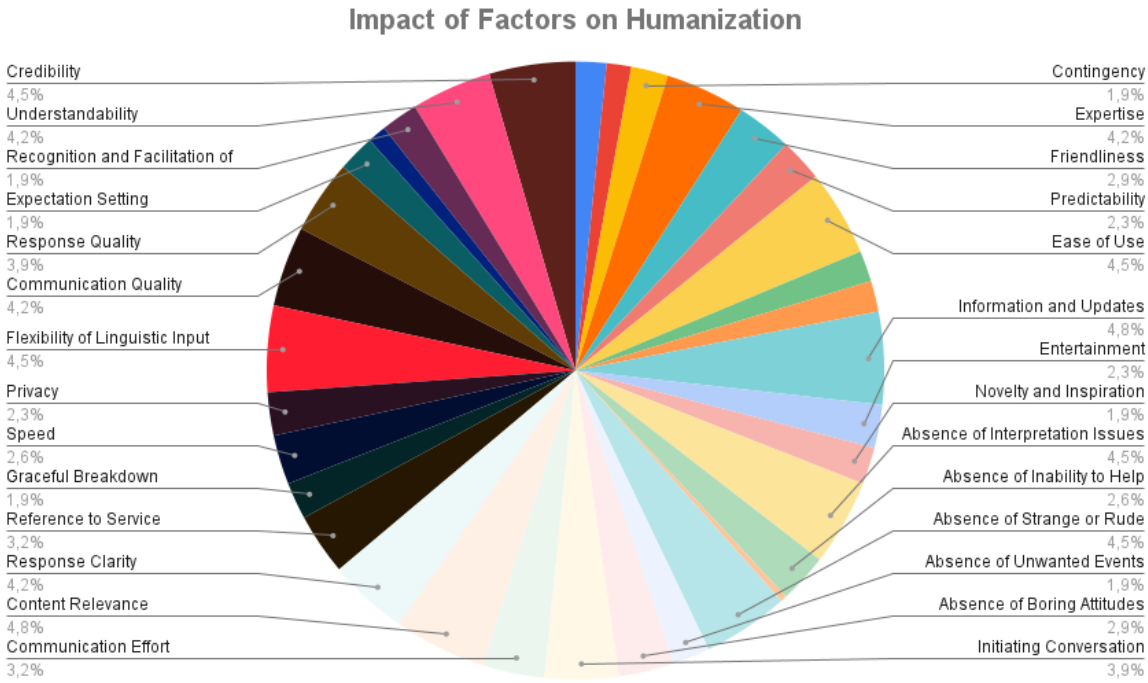

Figure A.13: Impact of Impact Factors on Humanization according to the Weighting Questionnaire data for Group 20 - Unicamp Campus Guide

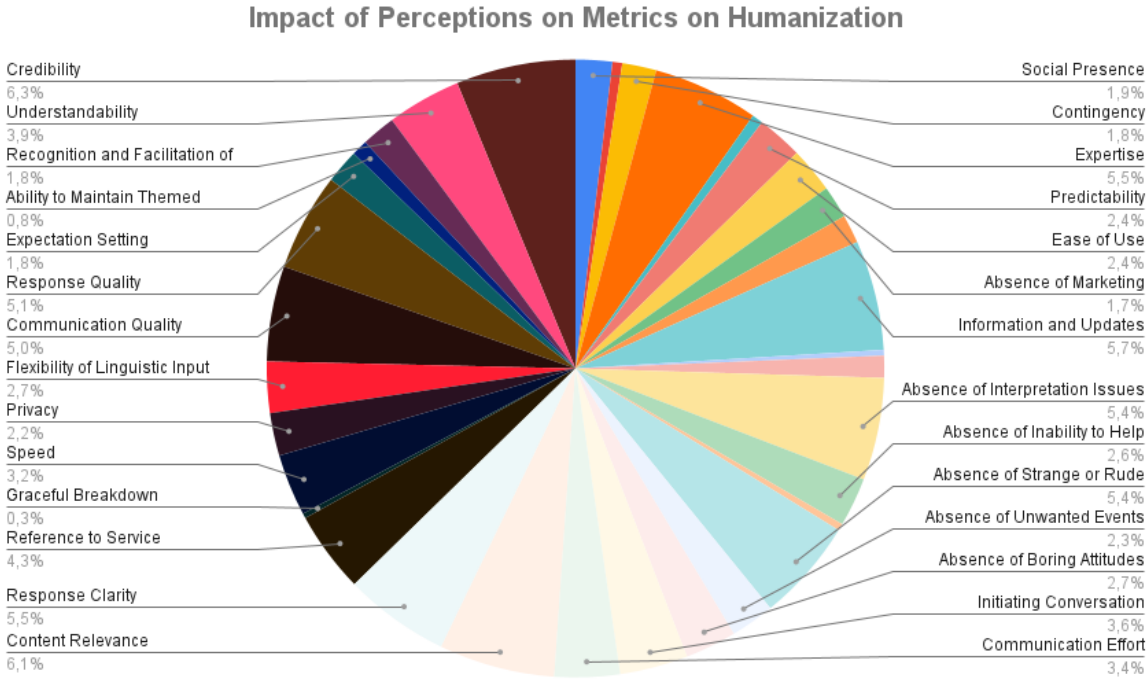

Figure A.14: Impact of Individual Scores for each Impact Factor in the calculation of the Final Humanization Metric of Group 20

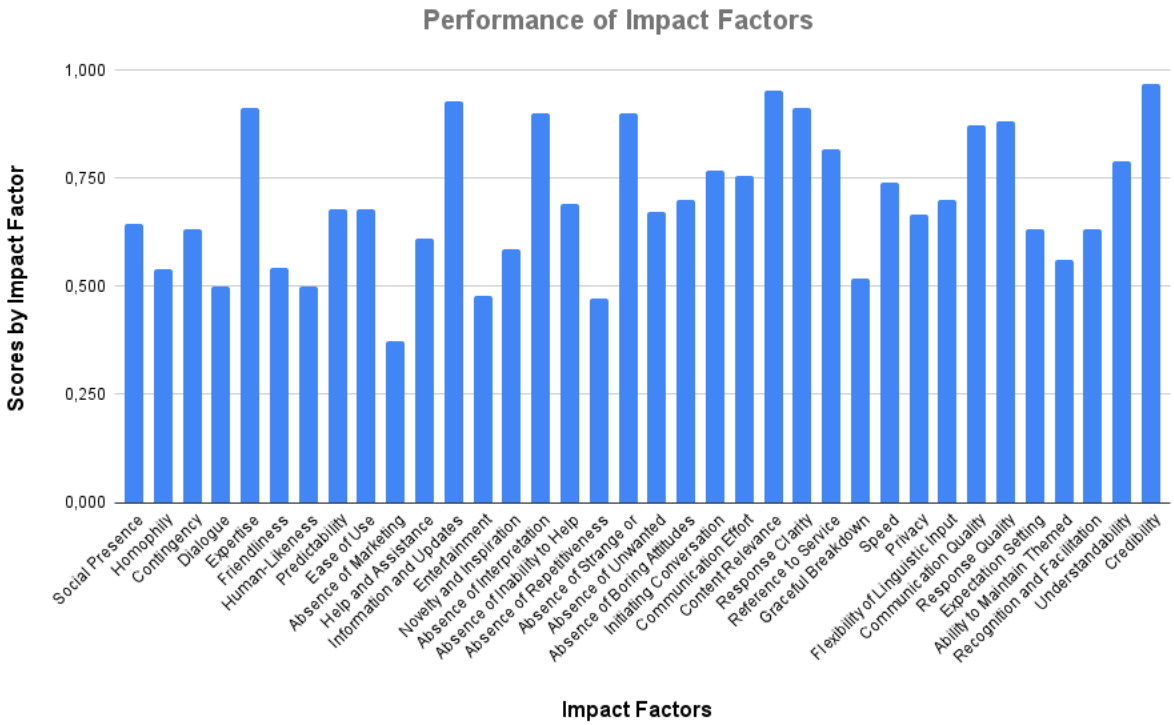

Figure A.15: Final Performance of each Impact Factor in the High Prototype of Group 20

A.6 Group 21 - Game Recommender

- **Description:** *A big problem for game lovers is finding games that match your profile. The choice involves several factors, such as platform, theme, graphics, and gameplay. A bot that could recommend the best games according to the gamer's characteristics would make finding new games more efficient and reduce the risk of losing money on games that do not satisfy the player's taste.*
- **Final Value of the Metric in Humanization:** 0,782
- **Complete Results of the Humanization Assessment:** link

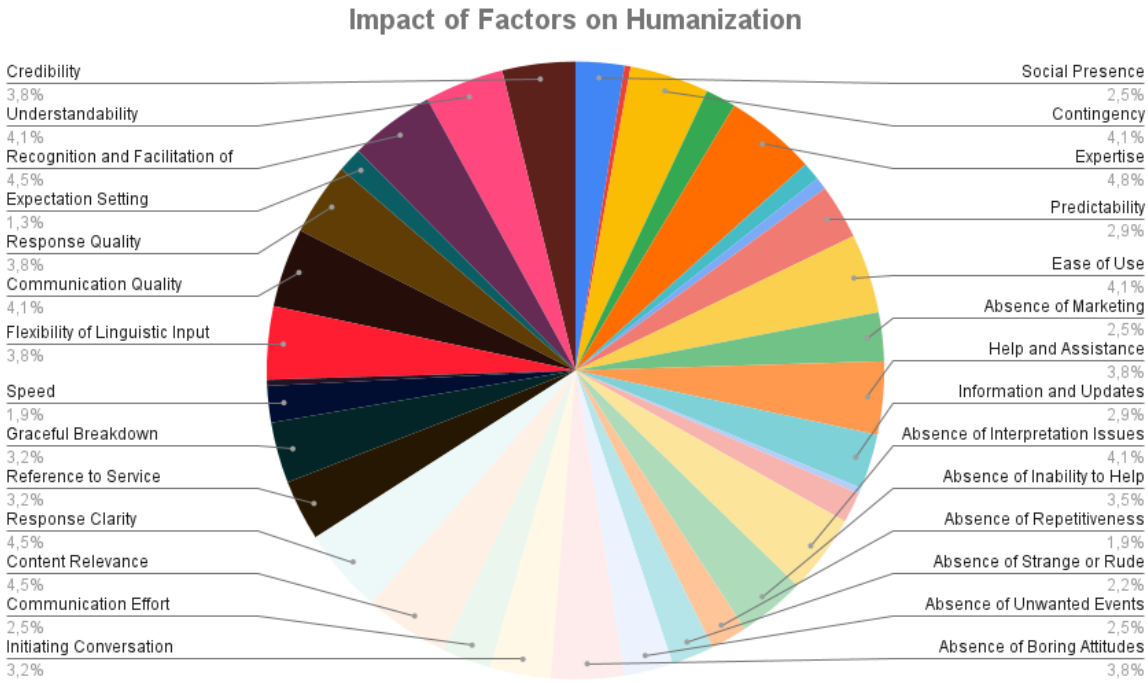

Figure A.16: Impact of Impact Factors on Humanization according to the Weighting Questionnaire data for Group 21 - Game Recommender

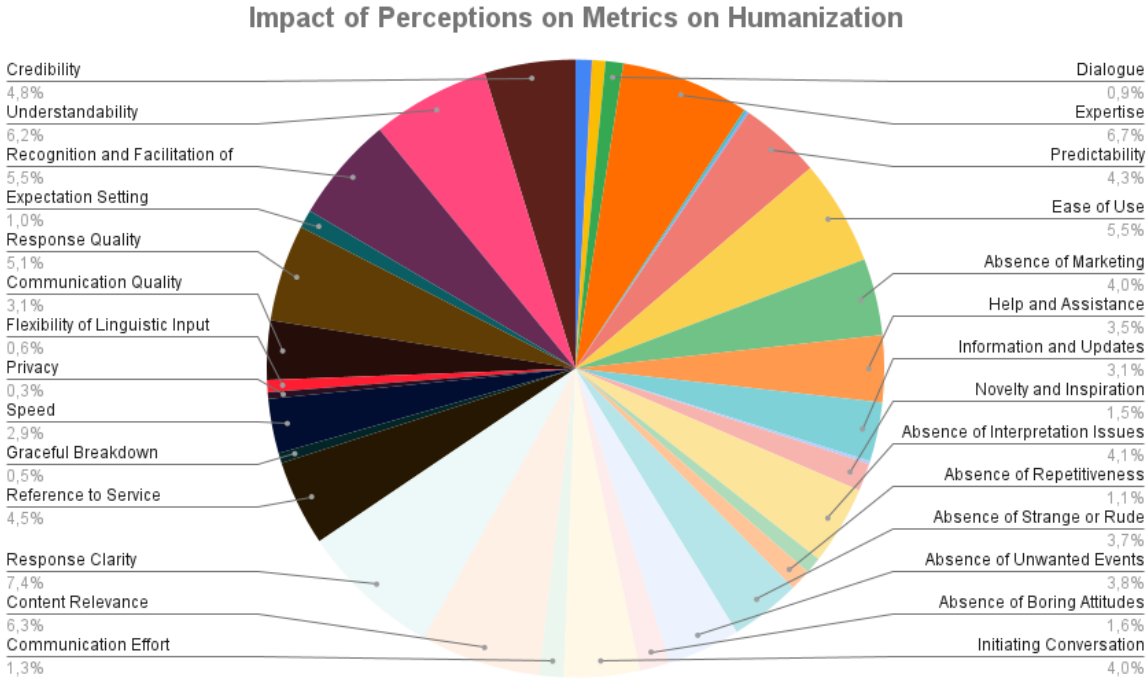

Figure A.17: Impact of Individual Scores for each Impact Factor in the calculation of the Final Humanization Metric of Group 21

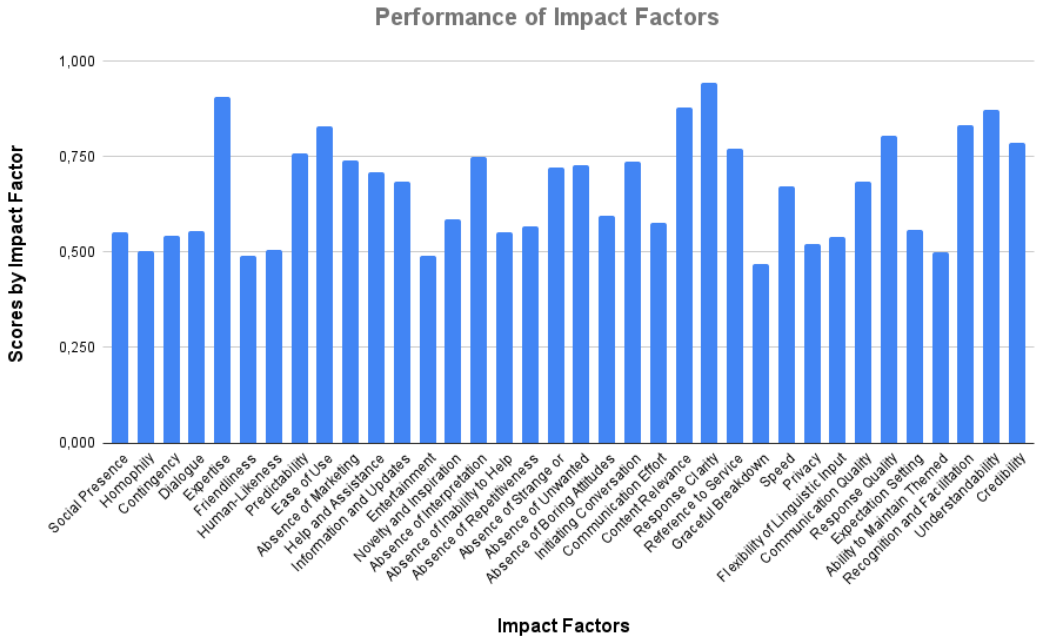

Figure A.18: Final Performance of each Impact Factor in the High Prototype of Group 21

A.7 Group 23 - Where to find your Movie/Series?

- **Description:** *A chatbot that informs you on which streaming platform or purchases a particular movie or series is available, facilitating access to content.*
- **Final Value of the Metric in Humanization:** 0,858
- **Complete Results of the Humanization Assessment:** link

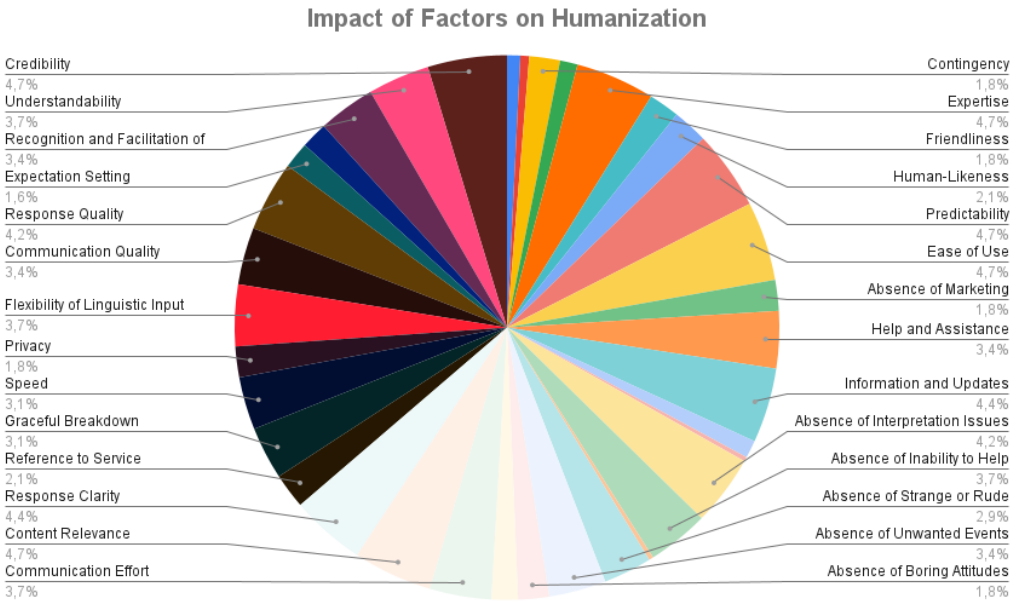

Figure A.19: Impact of Impact Factors on Humanization according to the Weighting Questionnaire data for Group 23 - Where to find your Movie/Series?

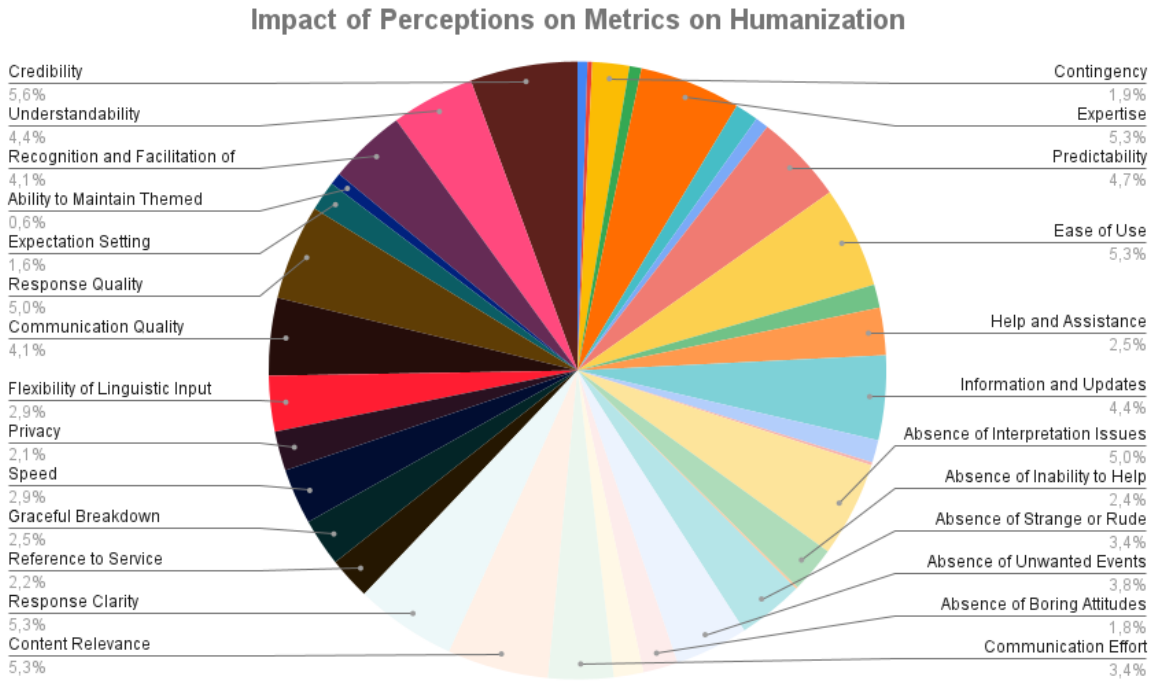

Figure A.20: Impact of Individual Scores for each Impact Factor in the calculation of the Final Humanization Metric of Group 23

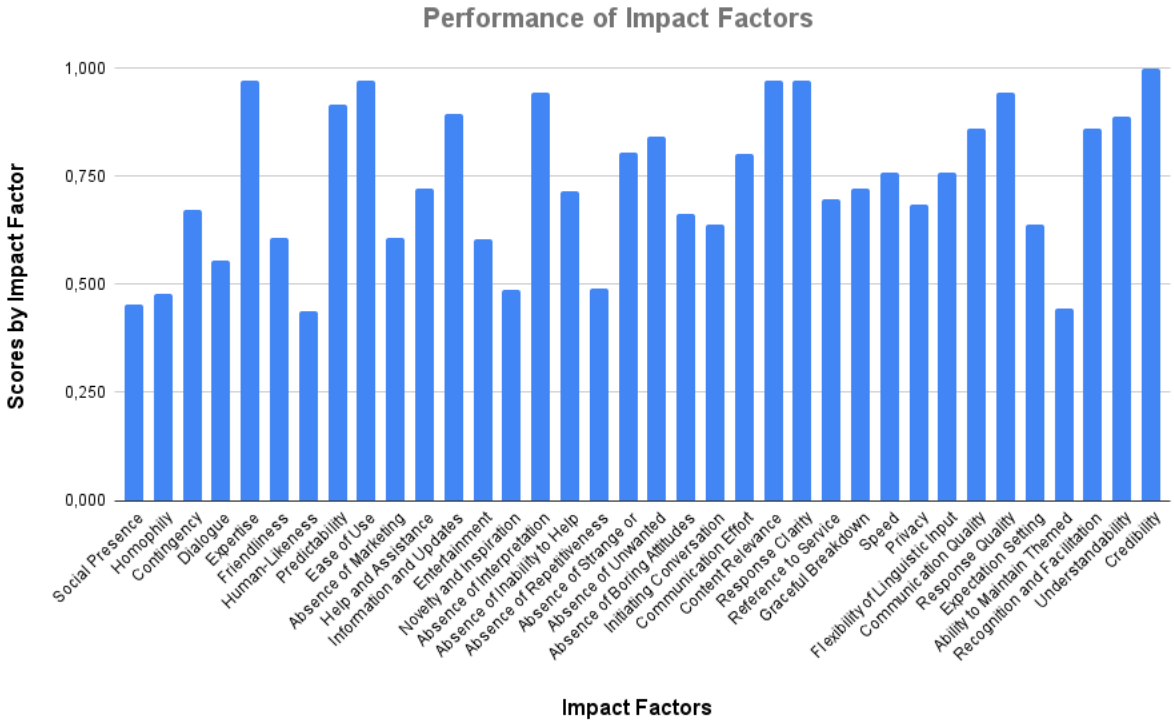

Figure A.21: Final Performance of each Impact Factor in the High Prototype of Group 23

A.8 Group 24 - Lack of Practicality in Scheduling Appointments and Exams in a Hospital

- **Description:** *It is proposed to create a Chatbot to facilitate this experience for the patient, avoiding queues or busy phones.*
- **Final Value of the Metric in Humanization:** 0,742
- **Complete Results of the Humanization Assessment:** link

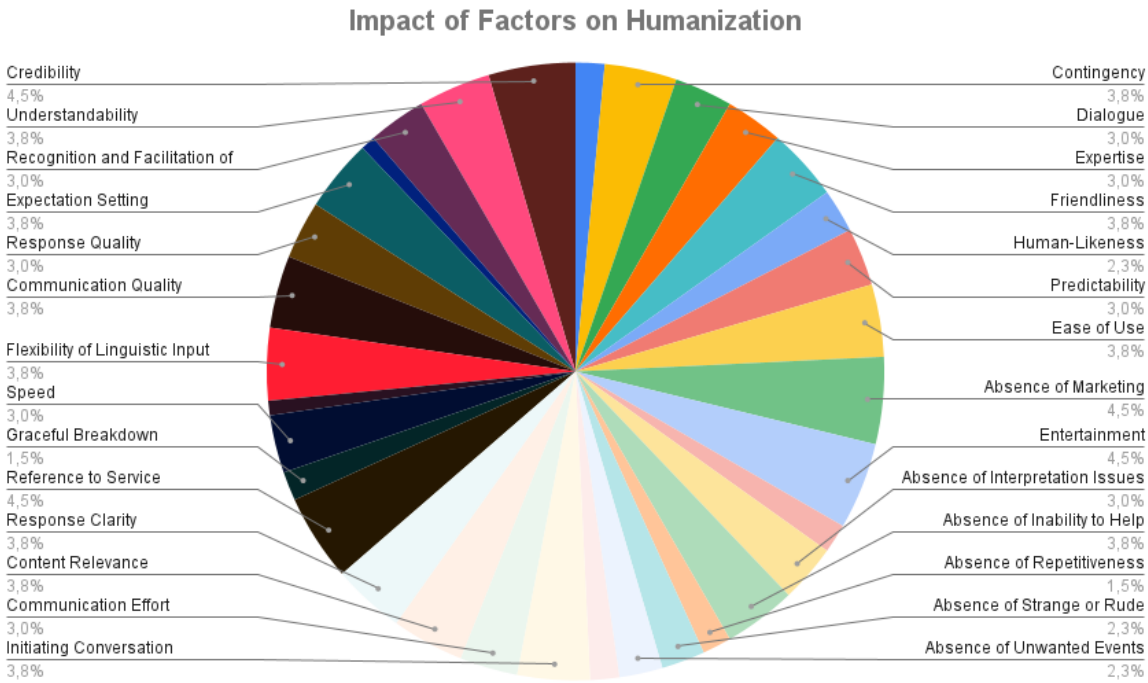

Figure A.22: Impact of Impact Factors on Humanization according to the Weighting Questionnaire data for Group 24 - Lack of Practicality in Scheduling Appointments and Exams in a Hospital

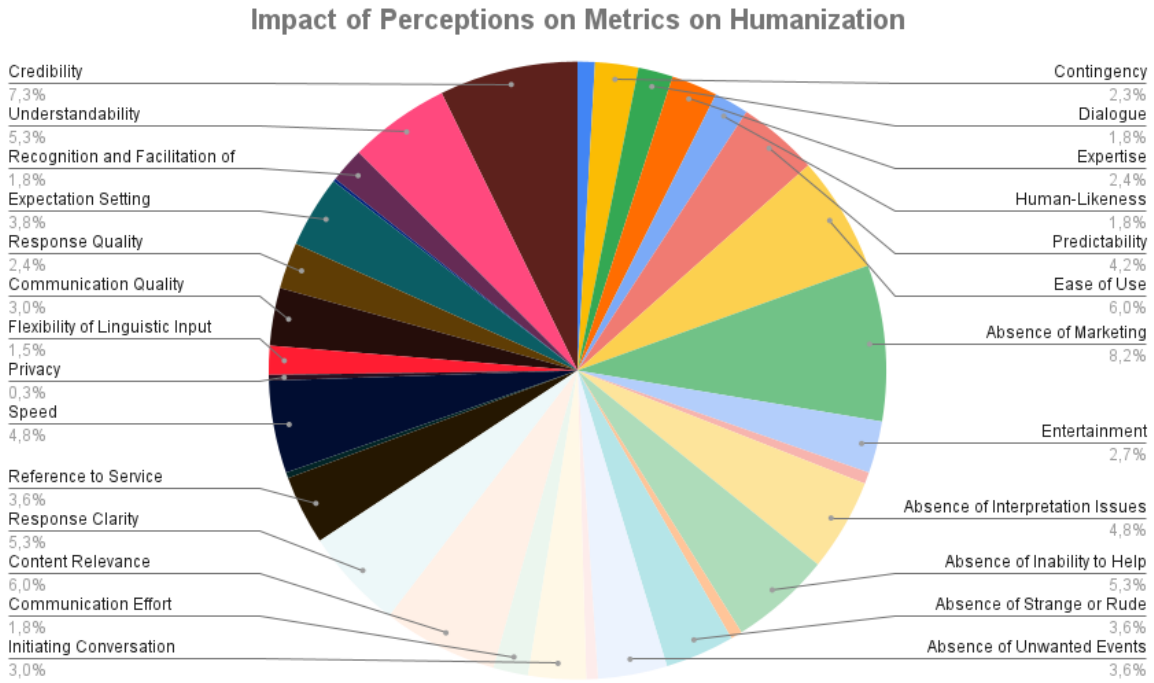

Figure A.23: Impact of Individual Scores for each Impact Factor in the calculation of the Final Humanization Metric of Group 24

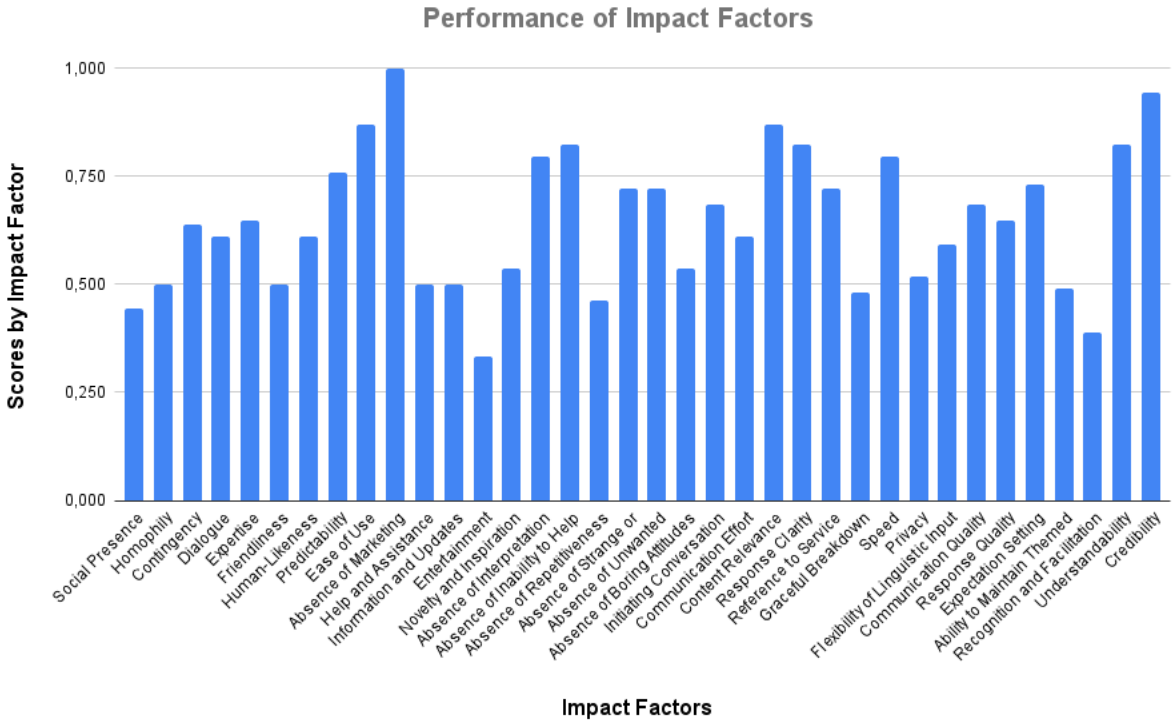

Figure A.24: Final Performance of each Impact Factor in the High Prototype of Group 24

A.9 Group 29 - Public Security in Barão Geraldo

- **Description:** *Assaults and robberies in Barão Geraldo increased, and with that, the feeling of insecurity became constant among those who lived in the region. Unicamp students articulated, and the media began to report the situation. Even with the increase in policing, cases remain a concern.*
- **Final Value of the Metric in Humanization:** 0,843
- **Complete Results of the Humanization Assessment:** link

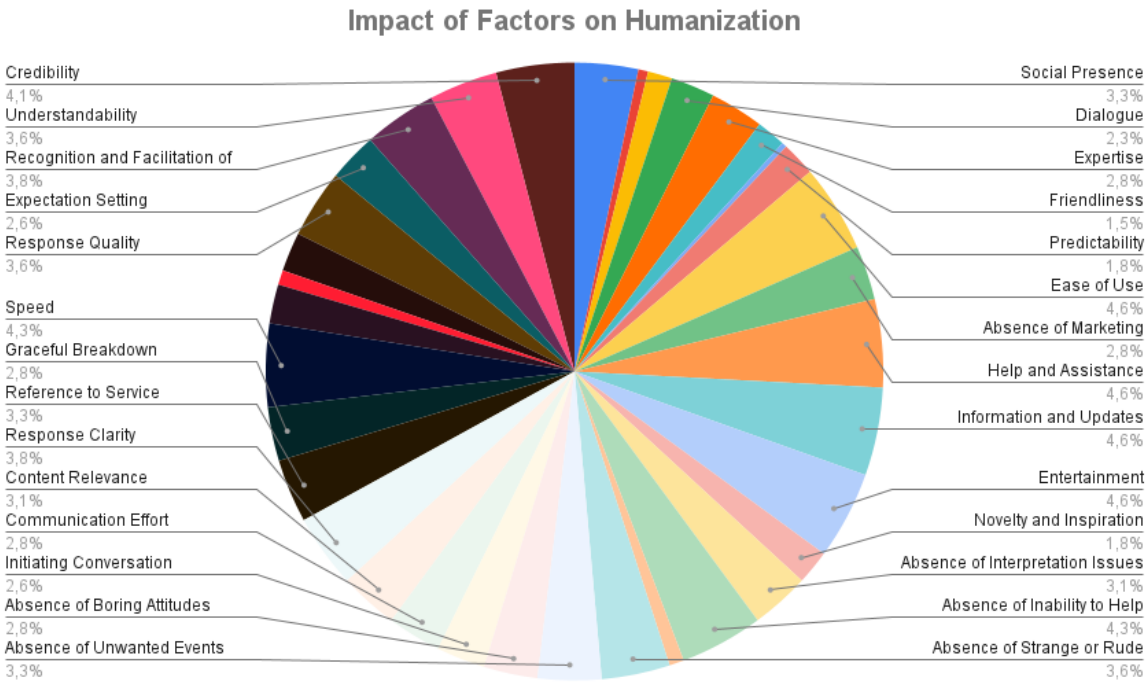

Figure A.25: Impact of Impact Factors on Humanization according to the Weighting Questionnaire data for Group 29 - Public Security in Barão Geraldo

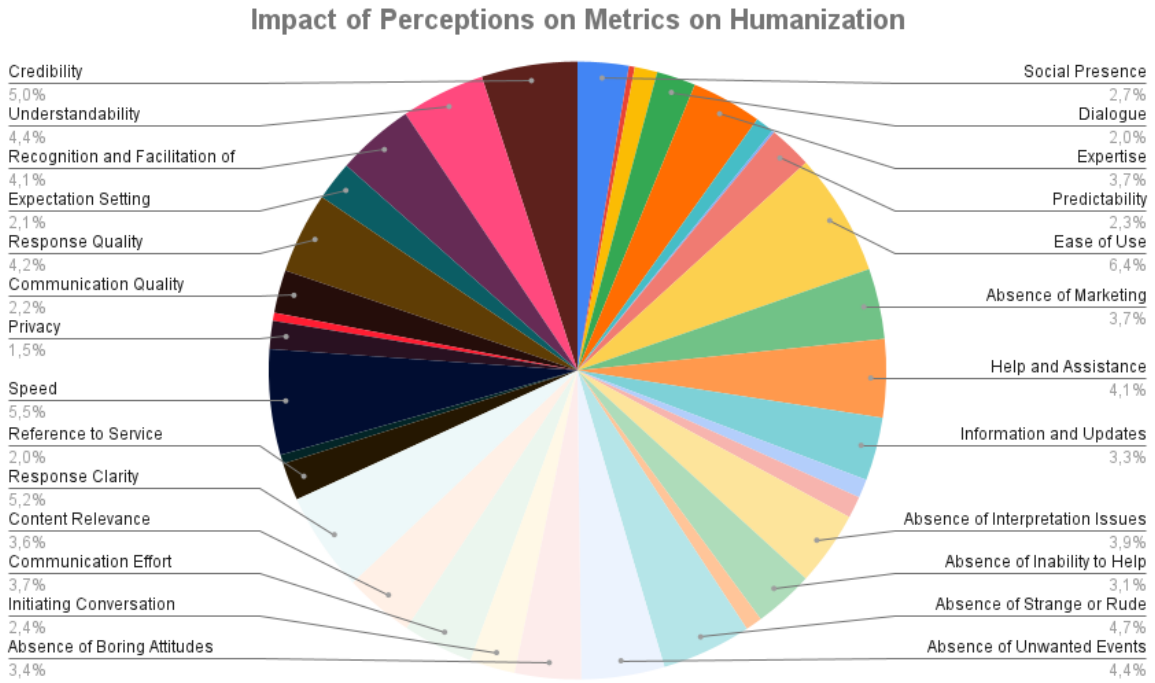

Figure A.26: Impact of Individual Scores for each Impact Factor in the calculation of the Final Humanization Metric of Group 29

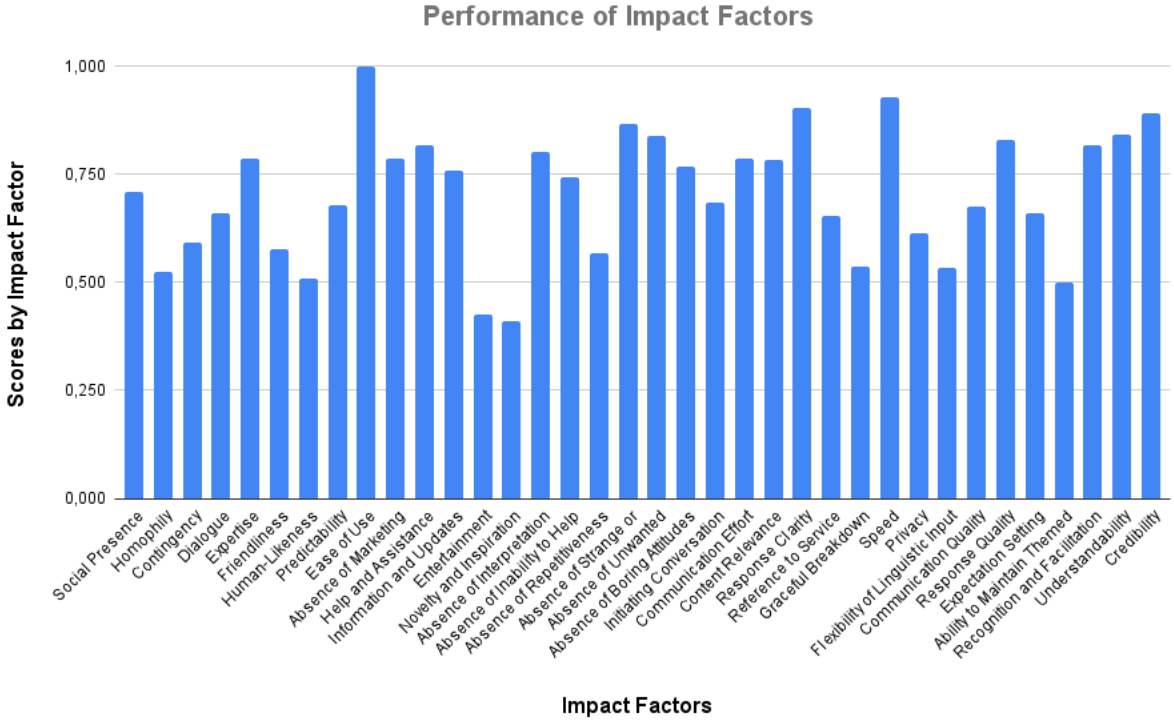

Figure A.27: Final Performance of each Impact Factor in the High Prototype of Group 29

A.10 Group 31 - School Dropout in Public Schools

- **Description:** *Public school students face many problems in their personal lives that affect their education directly or indirectly. In addition, schools responsible for the education of their students, overloaded by their activities, cannot map the action on these problems faced by their students, generating barriers that imply difficulty in studies and sometimes school dropout.*
- **Final Value of the Metric in Humanization:** 0,846
- **Complete Results of the Humanization Assessment:** link

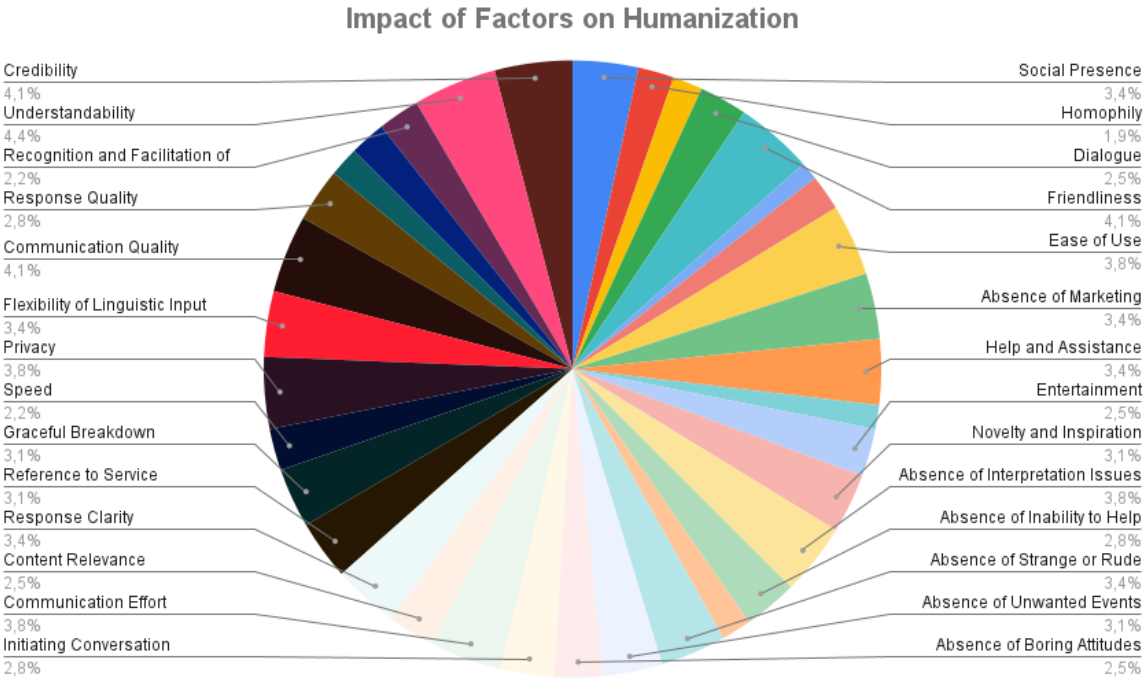

Figure A.28: Impact of Impact Factors on Humanization according to the Weighting Questionnaire data for Group 31 - School Dropout in Public Schools

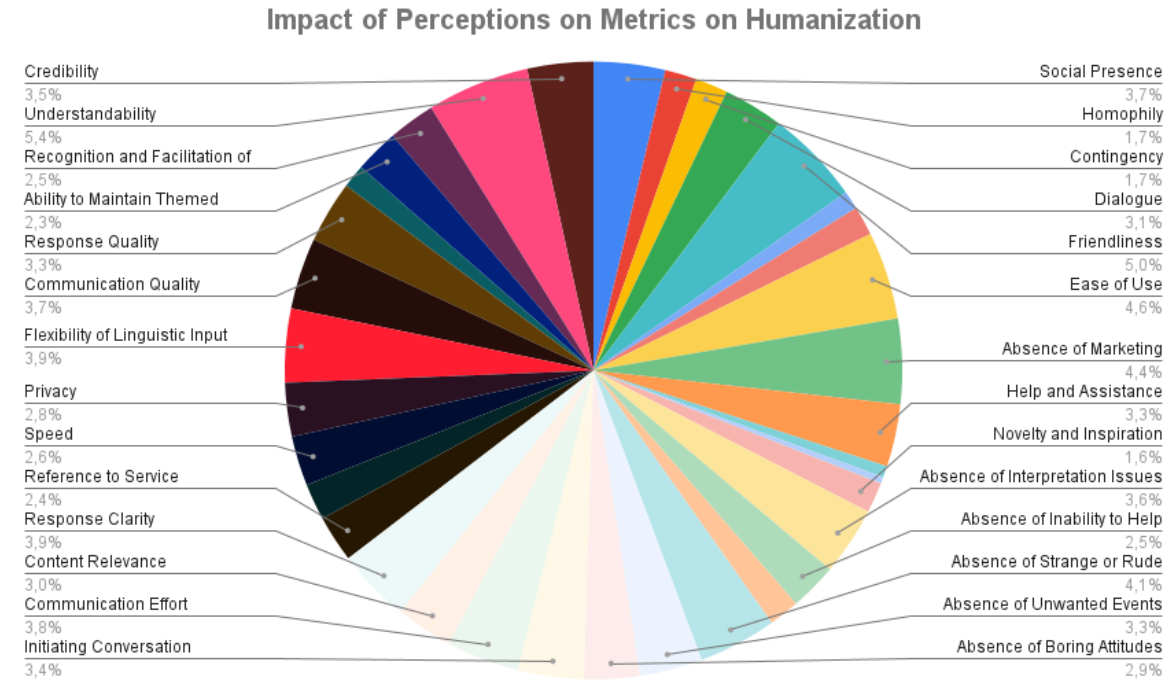

Figure A.29: Impact of Individual Scores for each Impact Factor in the calculation of the Final Humanization Metric of Group 31

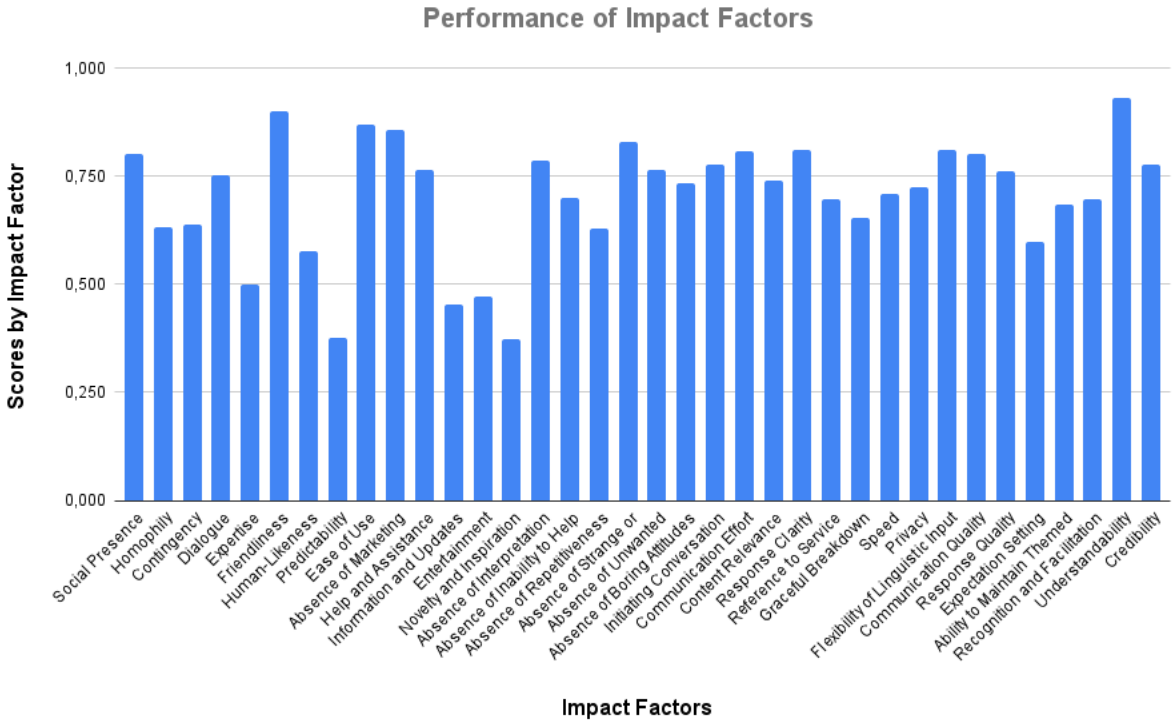

Figure A.30: Final Performance of each Impact Factor in the High Prototype of Group 31