


Proposing Usability-UX technologies for the design and evaluation of text-based chatbots


Malu Gabriele Silva Mafra   [Federal University of Maranhão | malu.gabriele@discente.ufma.br]

Kennedy Nunes  [Federal University of Maranhão | kennedy.anderson@discente.ufma.br]


Simara Rocha  [Federal University of Maranhão | simara.rocha@ufma.br]


Geraldo Braz Junior  [Federal University of Maranhão | geraldo.braz@ufma.br]

Aristofanes Silva  [Federal University of Maranhão | ac.silva@ufma.br]

Davi Viana  [Federal University of Maranhão | davi.viana@ufma.br]

Williamson Silva  [Federal University of Pampa | williamsonsilva@unipampa.edu.br]

Luis Rivero  [Federal University of Maranhão | luis.rivero@ufma.br]

 Universidade Federal do Maranhão, Federal University of Maranhão, Av. dos Portugueses, 1966 - Bacanga, São Luís, MA, 65080-805, Brazil.

Received: 17 November 2023 • **Accepted:** 04 March 2024 • **Published:** 18 March 2024

Abstract Chatbots are interactive systems that communicate using natural language with human users, via a textual interface or voice activation. These tools are useful for many spheres of business such as Customer Service, Sales, Education and Learning, Health and Entertainment. Recently, chatbots have become popular, with significant growth in the software industry, especially text-based chatbots. This is encouraging developers to create their own tools, as well as attracting efforts from researchers into this area. Despite this highlight, technologies to guarantee the quality of chatbots and user satisfaction are not keeping up with the growing demand for these tools. Considering this, there is a need to propose technologies capable of supporting developers and development teams in the process of building and evaluating chatbots. Therefore, this research proposes to develop artifacts applicable to the design and evaluation process of chatbots, based on quality attributes identified in systematic literature reviews related to Usability and User Experience (UX), due to the importance and impact that these aspects have on user satisfaction and the perceived quality of the system. The first artifact is the U2Chatbot inspection checklist, developed to assist development teams in the process of identifying defects in text-based chatbots. The second artifact is a set of interface design patterns, DP-U2Chatbot, containing useful examples to support developers in the process of building chatbots. The technologies were subjected to the necessary evaluations. The results of the empirical study regarding the U2Chatbot inspection checklist indicated that participants considered the technology useful for discovering defects in chatbots, however, ease of use could be improved. The participants' experience discreetly influenced the effectiveness and efficiency of the technique, leading us to believe that professionals with a certain level of inspection experience can benefit more from the checklist. Regarding the evaluation of DP-U2Chatbot design patterns, the results generally indicated that the technology is easy to understand and useful in supporting the design of chatbots, helping to build better tools.

Keywords: Inspection Checklist, Chatbots Evaluation, Usability, User Experience, Design Patterns

1 Introduction

Chatbots are computer interactive systems developed to interact with humans in natural language [Sharma *et al.*, 2017]. Some purposes for this type of system include entertainment, education, healthcare, customer service, and business [Georgescu *et al.*, 2018; Sperlí, 2020; Rahman *et al.*, 2017; Rosruen and Samanchuen, 2018]. The term *chatbot* is widely used and covers intelligent *bots*, conversational agents and intelligent personal assistants [Adamopoulou and Moussiades, 2020; Motaung, 2022].

The approach most used by chatbots is text-based, however, these tools can integrate voice and graphic animation that portray humanity [Radziwill and Benton, 2017]. Furthermore, they can be simple just based on pre-established rules and keywords [Gomes *et al.*, 2020] or use complex artificial intelligence concepts. Currently, Intelligent Personal Assistants - IPAs are on the rise, such as Amazon's Alexa tool [Brill *et al.*, 2019], however, text-based chatbots are still the

most widespread on the market [Thorat and Jadhav, 2020; Rapp *et al.*, 2021] and lead users' preference, according to the study by [Ciechanowski *et al.*, 2018].

Recently, the popularity of chatbots has increased, generating a high demand for these tools [Sharma, 2021; Chaves and Gerosa, 2021]. This growth creates the opportunity to propose solutions that help these tools to be developed with quality, satisfying users who are increasingly demanding [Muñoz and Avila, 2019], in addition to helping and facilitating the work of developers. As with other systems, it is important that chatbots reach a certain level of quality to be truly useful for their users [Guerino and Valentim, 2020]. To this end, applying evaluation techniques considering aspects of Usability and User Experience (UX) before the system is launched on the market can contribute to the quality of the software [Madan and Dubey, 2012].

Both Usability and User Experience are considered determining factors for the success of an interactive system [Hasan and Galal-Edeen, 2017; Cruz *et al.*, 2015]. Usability can

be defined according to ISO 9241-11 as the extent to which a product can be used by users to achieve specific objectives of effectiveness, efficiency and satisfaction, in a given context of use. UX, in turn, refers to a user's perceptions and reactions when interacting with a system or product, being influenced by factors such as the context of use, functionality, performance, presentation, interactive behavior of the system, in addition to factors such as personality, skills, experiences, internal and physical state of the user [Mirnig et al., 2015]. However, it is noted that there is still a lack of evaluation and support technologies for chatbot design, which unite these concepts and which have been subjected to appropriate empirical evaluations [Guerino and Valentim, 2020].

Considering the above, this research aims to contribute to the design and evaluation process of chatbots, proposing artifacts focused on Usability and User Experience that improve the quality of these tools from the point of view of end users. The first artifact is an inspection *checklist* to identify defects in text-based chatbots, a verification/evaluation technique that does not require a complete version of the system, and can be used before the *release* of the software on the market. In this way, it can contribute to reducing the costs of correcting identified defects [Frazao et al., 2020; Alsayed et al., 2017]. Furthermore, there is a proposal to develop a set of Design Patterns (guides) aimed at the chatbot design process, providing general solutions to problems that frequently occur in a given context in the software project [Gomes et al., 2021] and achieve assist developers in the design process of these tools.

This study is divided as follows: Section 2 presents a brief contextualization and the work related to this research. Section 3 presents a summary of the stage of identifying the quality attributes that will make it possible to propose the artifacts. Section 4 presents the process of creating the inspection checklist. Section 5 shows the procedures adopted to evaluate the checklist. Section 6 presents the conception of design patterns (UI) for chatbots and Section 7 shows the results of the technology assessment. Section 8 brings some important points to be highlighted. Finally, the conclusions of this study can be found in Section 9.

2 Current Status and Related works

Currently, intelligent chatbots launched in the industry are increasingly sophisticated. With the use of machine learning and artificial intelligence concepts (NLP, for example), current chatbots are complex and can perform multiple tasks [Sharma and Joshi, 2020], [Reichert et al., 2022]. Despite technological developments, chatbots available in the industry often fail to satisfy users. When looking at comments from chatbot users in the *Google Play* App Store, for example, it is possible to find several user complaints. Complaints include poor performance, unnatural interactions, inadequate responses, difficulty understanding complex questions and lack of customization options [Mafra, 2023], which can frustrate users and damage the reputation of the company or business that the chatbot represents.

With the aim of contributing to the understanding and resolution of these problems, researchers have proposed studies

focused on Usability and User Experience for chatbots, as both concepts are related to user satisfaction. One of these works is that of Barbosa et al. [2022], which carried out an exploratory study on User Experience evaluation methods in chatbots. Another work to be highlighted is that of De Souza Monteiro et al. [2023], which investigated the Usability of national and international chatbots. To propose technologies capable of improving Usability and User Experience in chatbots, one approach is to identify important quality attributes for chatbots. The work of Radziwill and Benton [2017] states that quality attributes can be used as *checklists* by development teams in inspections to evaluate whether the system addressed essential requirements. Assessment methods like this help reviewers in the process of discovering defects in software products [Brykczynski, 1999] before the chatbot is launched on the market, contributing to reducing costs with correcting identified defects [Alsayed et al., 2017].

As an example of work that presents an evaluation method based on *checklist* inspection, we can mention Sugisaki and Bleiker [2020], which presents a technology composed of 53 useful verification items for evaluating conversational interfaces based on Nielsen's heuristics. The proposed technology was submitted for evaluation by 15 professionals who analyzed each item and answered a questionnaire about the relevance of each one and how efficient, pleasant, convenient and effective the checklist was. The results indicated that 80% of the checklist items were considered relevant, but some verification items were highlighted by the reviewers as very technical and difficult to understand.

The Borsci et al. [2021] research, in turn, also proposed a *checklist* to evaluate usability in chatbots with artificial intelligence, Bot-Check. The technique with 42 verification items, different from the work of Sugisaki and Bleiker [2020], was not based on Nielsen's Heuristics, but used attributes from the literature review taken from the work of Radziwill and Benton [2017]. The list of quality attributes was validated by chatbot designers and end users to find out which items would be kept or removed from the list. Subsequently, the Bot-Check scale was validated with 141 participants to identify its relevance. The results indicated that the proposed scale can be used by designers as a tool to ensure quality in chatbot design before testing with end users.

The related works presented demonstrate that there are efforts to improve Usability and User Experience in chatbots. Furthermore, the techniques presented are useful and important for evaluating chatbots, however, they mainly focus on evaluating Usability, encompassing few aspects of User Experience (UX). Therefore, considering the results of the studies presented previously and the importance of quality attributes to propose new technologies [Radziwill and Benton, 2017], it is necessary to carry out a study that identifies quality attributes also focused on the user experience for conversational agents in conjunction with attributes of usability, so that evaluation techniques focused on both concepts can be proposed. This process of identifying quality attributes focusing on Usability and User Experience for chatbots can be found in the following section.

3 Identification of Quality Attributes

The technology development process began with a literature review with the aim of identifying works that reported relevant quality attributes for chatbots and that were focused on the concepts of Usability and User Experience. It is important to emphasize that the objective of this article is not to present in detail the process of planning and executing the literature review, but rather the results obtained from it. Details of the systematic literature review can be found in a Technical Report at this [link](#).

When planning the review, research criteria were developed, which included works that addressed attributes related to Usability and UX for text-based chatbots and excluded works that did not meet this criterion, works outside the English or Portuguese language, publications that weren't scientific articles, in addition those that are repeated or unavailable for download/access. The search string defined for this review was constructed using keywords and synonyms taken from the works of Coppola and Ardito [2021]; Suhaili *et al.* [2021]; Guerino and Valentim [2020]; Cabrejos *et al.* [2018]: (*“Attribute” OR “Feature” OR “Characteristic” OR “Aspect” OR “Heuristic” OR “Principle” OR “Requirement” AND (“Chatbot” OR “Conversational User Interface” OR “Conversational Agent”) AND (“Usability” OR “Usable” OR “User Experience” OR “UX”)*). The database selected to search for scientific articles was Scopus, as it is a robust and reliable scientific library [Codina, 2005], which indexes publications that are present in other libraries such as IEEE and ACM Digital Library.

Regarding the selection period, articles published until February 2022 were considered. In total, 185 articles were returned and, after applying all inclusion and exclusion filters, 18 articles in total were selected, which provided 313 quality attributes generics related to Usability and User Experience for text-based chatbots. We chose to focus on generic attributes so that technologies developed with the identified attributes could cover different types of chatbots, which would not occur with the use of specific attributes for a given type of chatbot, for example, anamnesis attributes specific to medical chatbots. Some of these attributes can be found in Table 1.

We can see in Table 1 that there are a variety of quality attributes identified in the literature review, even though the table is summarized. It is possible to identify attributes of Humanity, such as “Small Talk”, “Maintenance of Context”, of Accessibility, such as “Ease of Use”, of Affection, such as “Empathy” and “Expression of Emotions”, etc.

During the analysis of the 18 publications selected in the systematic literature review, we identified 5 scientific articles that describe evaluation techniques for chatbots and provide most of the quality attributes identified in the study, that is, of the 313 attributes identified, only 5 articles were responsible for provide 173 quality attributes. Considering this, we realized the importance of carrying out a second literature review to identify other chatbot evaluation techniques that could also contribute with more important quality attributes for chatbots focused on Usability and UX, in order to avoid fundamental aspects not being discovered. We chose to keep the two literature reviews separate so that the results could

Table 1. Excerpt from the Quality Attributes SLR

Paper	Quality Attributes
A006	Context Understanding, Help Options, Empathy, Natural Language, Typing Error Understanding, etc.
A013	Correct answers, Context Maintenance, Appropriate vocabulary, Ability to deal with clarifying questions, etc
A014	Humanity, Empathy, Data security, Good performance, Ease of Use, User Satisfaction, Relevant answers, etc.
A015	Persistent menu, Quick Response Buttons, Pleasant personality and Explanation of the type of input expected.
A016	Follow-up questions, Small Talk, Choosing sophisticated words and well-constructed sentences, Expressing emotions, etc.

benefit two audiences, a group that only seeks UX and Usability quality attributes for chatbots and another group that seeks information about chatbot evaluation techniques.

To carry out the second systematic literature review, research criteria were also developed that included works that addressed evaluation techniques for chatbots that focused on the concepts of Usability and UX for text-based chatbots and excluded works that did not meet this criterion, articles in other language other than English or Portuguese, publications that weren't scientific articles and those that were repeated or unavailable for download/access. The search string constructed for this review, in turn, used keywords and synonyms taken from the works of Denecke and Warren [2020]; Coppola and Ardito [2021]; Suhaili *et al.* [2021]; Guerino and Valentim [2020]; Cabrejos *et al.* [2018]: (*“Technique” OR “Instrument” OR “Tool” OR “Checklist” OR “Questionnaire” OR “Approach” OR “Method” OR “System” OR “Scale” OR “Scheme” OR “Framework” OR “Model” AND (“Evaluation” OR “Assessment” OR “Measurement” OR “Testing” OR “Recognition” OR “Measure” OR “Evaluating” OR “Tracking” OR “Assess”) AND (“Usability” OR “Usable” OR “User Experience” OR “UX”) AND (“Chatbot” OR “Conversational User Interface” OR “Conversational Agent”)*). The database selected to search for scientific articles was also Scopus, as in the first literature review.

The selection period also took place at the beginning of 2022 and we considered articles published until April 2022. In total, 272 articles returned and, after applying the inclusion and exclusion filters, 14 articles were selected that presented evaluation techniques related to Usability and User Experience for text-based chatbots. These publications provided 273 generic quality attributes. An excerpt with part of these attributes can be found in Table 2 below.

In Table 2 it is possible to find different types of attributes, which include Subject Maintenance, Simple language, Conversation Tips, Linguistic Accuracy, Responses in a reasonable time, Ease of Use, and others. Comparing this set of attributes with those discovered in the first review, a much smaller number of quality attributes referring to factors such as Humanity and Affection aimed at chatbots were observed.

Table 2. Excerpt from the Results of the second SLR

Paper	Evaluation Techniques and Attributes
B004	Subject maintenance, Appropriate linguistic register, Conversation Tips, Ability to deal with inappropriate utterances and damage control, etc.
B011	Realistic and engaging personality, Avoid appearing too robotic, Ease of Navigation, Ability to give useful, appropriate and informative responses, etc.
B016	Responsiveness, Simple language, Linguistic Accuracy, Use of bright colors for fonts, Provision of reliable information, etc.
B018	Ease of Use, Well-integrated functions, Easy to learn how to use, Avoid being unnecessarily complex, Avoid inconsistencies, etc.
B019	Response in reasonable time, Operation in parallel with other software without losing performance, Use of resources efficiently, Satisfactory graphical interface, etc.

Despite this, this complementary review allowed us to discover attributes such as Use of bright colors for fonts, Responsiveness, Operation in parallel with other software without losing performance, Well-integrated functions.

With the two lists of quality attributes identified in the first and second Systematic Literature Review, it becomes possible to propose useful technologies to evaluate and assist the design of intelligent chatbots. More information about the systematic literature reviews presented is in the [link](#). The following section presents the process of creating the inspection checklist to identify defects in chatbots.

4 U2CHATBOT Checklist Inspection

To develop the inspection *checklist* useful for evaluating and identifying defects in chatbots, the following steps were considered, according to the methodology used in the work of Frazão [2021]: (1) identification of quality attributes to chatbots related to Usability and User Experience in systematic literature reviews. (2) analysis of the identified quality attributes, grouping similar ones and organizing them in an inspection *checklist*.

This first step is described in Section 3. The second stage was the analysis of lists of attributes obtained from systematic literature reviews. The first list discovered contained 313 quality attributes and the second list contained 273 quality attributes. When analyzing the two lists, we observed the presence of items that were the same or that established a relationship, generating the need to group these similar and/or related attributes to reduce redundancies. As an example of similar items capable of being grouped, there are the attributes: (A006) Permanent menu and help options; (A015) Use of persistent menu (Help, Menu, I'm lost); (A060) Chatbot offers permanent menu and help option; (A031) The help page is helpful. All these attributes refer to help options and access to documentation.

As a grouping of related items, we have as an example of

suitable attributes: (A006) Ability to carry out small talk to remedy the problem of artificial conversations; (A016) The chatbot knows how to deal with when small talk is not the user's style; (A013) Does the chatbot handle generic and off-topic requests (e.g. small talk) appropriately?; (A025) Can the chatbot maintain focus during the conversation? Although these items are not the same, they were grouped together because they establish a relationship linked to the focus aspect, as some refer to the use of the small talk resource (diversion from focus) and others to the establishment of focus. After this grouping procedure, the first list was reduced to 162 attributes and the second list was reduced to 109 attributes. However, so that the inspection *checklist* could be created, there was still a need to group the two lists of discovered quality attributes (with 162 attributes and with 109 attributes). Therefore, we carried out the grouping process of similar attributes again, to avoid redundancies in the inspection *checklist* items. After the treatment carried out on the two lists of quality attributes, the result was a list with 107 useful items used in the creation of the evaluation technology. An excerpt from the U2CHATBOT inspection *checklist* can be found in the Table 3 below.

When analyzing the Table 3, it is observed that some items in the *checklist* U2CHATBOT may cause difficulties in understanding or even erroneous understanding by some users of the tool. To facilitate understanding, these items were described with brief clarifications and examples in parentheses. An example is **item P-5: “Does the chatbot perform effective task allocation, providing appropriate escalation channels for humans?”**. An inspector new to the area might not immediately understand what the item means, so, to avoid confusion and wasted time when having to go and research what it means, the item was rewritten adding a brief clarification **“P- 5: Does the chatbot perform effective task allocation (decide whether a certain function will be performed by the system or will be escalated to a human attendant), providing appropriate escalation channels for humans?”**.

The *checklist* U2CHATBOT categories were taken from the articles that provided the quality attributes. Some attributes already had categorization, others did not. Thus, for standardization purposes, the items were analyzed one by one and were then placed in categories from the articles that contributed most with Usability and UX attributes, such as the paper by Sugisaki and Bleiker [2020], which categorizes the items according to Nielsen's Heuristics and the paper by Anshu *et al.* [2021], with general categories for UX and usability.

To facilitate the use of U2CHATBOT *checklist*, we have developed tool support in spreadsheet format compatible with the Excel (*Microsoft Office*) and Calc (*LibreOffice, OpenOffice*) utilities. The tool support developed contains the following tabs: (1) brief summary of the U2CHATBOT checklist, with instructions for use; (2) the inspection checklist with the verification items, the field for responding to the items, fields for inserting the inspector's data, automatically updating graphics, a problem description field, a field to indicate the location of the problem and suggestions; and (3) field to report other problems that occurred during the inspection. To carry out the inspection with the *checklist*, the pos-

sible answers are *Yes*, *No* and *Not Applicable*. **Yes** must be chosen if the chatbot presents/answers the verification item. **No** should be chosen if the chatbot does not present/does not respond to the item and, if this occurs, it constitutes a possible defect that must be described in the Problem Description column and indicated in the Problem Location column. **Not Applicable** should be chosen if the verification item is not related to the type of chatbot evaluated, for example, an item related to handoff to a human expert would apply to a customer service chatbot but not to random conversations or entertainment from a chatbot. The complete U2CHATBOT *checklist* can be found **here**.

Table 3. Excerpt from the inspection checklist U2CHATBOT

Category	Checklist Items
VS-1	The chatbot gives immediate feedback to the user about their actions/transactions, in addition to information in a reasonable time about the system status throughout the interaction (<i>when it is processing a response, whose turn it is in the conversation and who said what in the conversation history</i>)?
PE-1	The chatbot requests confirmation and shows a summary before any action or transaction in order to prevent errors, dialogue failures and irreversible actions (<i>e.g. a permanent deletion of data</i>)?
AD-2	Does the chatbot provide access to help options, documentation, navigation options, and permanent menu to support the user?
F-20	Can the chatbot adjust to both a larger screen (<i>tablet or laptop</i>) and a smaller (<i>mobile phone</i>)?
FE-2	Chatbot allows experienced users to access advanced functions (<i>e.g. shortcuts, abbreviations, etc.</i>) to interact and correct errors more quickly and efficiently, following established conversational principles from others chatbots (<i>same shortcut keys, for example</i>)?
P-5	Does the chatbot perform effective task allocation (<i>decide whether a certain function will be performed by the system or will be escalated to a human attendant</i>), providing appropriate escalation channels for humans?
EC-3	Does the chatbot avoid requesting personal data (<i>Full name and CPF, for example</i>) of the user or any other unnecessary information?
ACE-1	Is the chatbot easy to use (<i>has intuitive navigation</i>), easy to start a conversation, easy to learn how to use it?

In the following subsection, we present a proof of concept carried out to verify the feasibility of the technology.

4.1 U2CHATBOT Checklist Proof of Concept

In order to verify whether *checklist* U2CHATBOT is truly capable of detecting defects in chatbots, we carried out a proof

of concept. We applied *checklist* to an intelligent text-based chatbot chosen from the Google Play Store, *Bing Chat with AI*. The criteria for selecting the application were that the tool is free, does not require a subscription to use and has a large number of downloads, comments and user ratings.

When applying *checklist* U2CHATBOT to the *Bing Chat With AI* chatbot, we identified that the tool did not meet 55 verification items (51%), indicating that possible problems were found. Of the items not met, those referring to the Humanity category stand out, in which 11 items were injured (H-1, H-2, H-3, H-4, H-9, H-10, H-12, H-13, H-15, H-16 and H-18). In fact, we observed that the chatbot is quite robotic, it does not maintain or understand the context and intentions. All items related to helping users recover from errors that may occur in the chatbot (ARE-1, ARE-2, ARE-3 and ARE-4) were also not met, as were items dealing with help and documentation (AD -1, AD -2, AD-3, AD-4, AD-5 and AD-6). The chatbot also violates important aspects of Accessibility (ACE-1, ACE-2, ACE-4, ACE-6 and ACE-7) and Flexibility/Efficiency (FE-1, FE-2, FE-3, FE-4 and FE-7). Additionally, items VS-5, CSR-2, CLU-1, CLU-2, CP-2, CP-6, CP-7, PE-1, PE-2, PE-3, RL-2, DEM -1, DEM-2, F-4, F-7, F-8, F-13, F-14, F-17, A-2, A-5, P-2, P-3 and P-4 were violated, indicating that the analyzed chatbot needs many improvements.

The following section will present the procedures and results of carrying out an experimental study to evaluate the U2CHATBOT *checklist* in terms of acceptance, effectiveness and efficiency.

5 U2CHATBOT Checklist Inspection Evaluation and Improvements

With the aim of examining the feasibility of the U2CHATBOT inspection checklist, we carried out an experimental study aiming to measure indicators of efficiency, effectiveness, perceived usefulness, perceived ease of use and intention for future use. For this, software engineers inspected a chatbot available in the *Google Play Store* application store using the *checklist* developed in this research to identify possible defects in the tool. With the data collected we carried out three analyses: (1) Quantitative analysis; (2) Analysis according to the TAM Model and (3) Qualitative analysis.

The experimental study was carried out between June and July 2023 with 29 participants; of which, 24 are undergraduate students and 6 professionals with experience already in the technology market. All participants agreed to the Free and Informed Consent Form (ICF) and filled out the first form with questions about the participants' level of education, experience with software development in general, chatbot development and software inspection. Through responses to these questions, participants were categorized as having Low, Medium or High experience.

All 29 study participants used the U2CHATBOT inspection checklist to inspect the chatbot *Bing: Chat with AI and GPT4*¹ chosen from the online store *Google Play Store* ac-

¹Bing Chat with AI link - <https://tinyurl.com/3vev24jv>

ording to the criteria: (1) that the tool was free; (2) that the chatbot was textual; (3) that the tool had many *downloads* and user comments, suggesting that it was already widely used. *Bing: Chat with IA and GPT4* had around 433 thousand reviews at the date of the experimental study, with 131,300 reviews below 5 stars, indicating the possibility of containing defects; thus making it eligible for use in this experiment. Participants received a file with instructions and descriptions on how to download and install the indicated chatbot on the inspector's device, as well as a script of tasks that the inspector would need to do with the chatbot (request simple and complex calculations, write wrong words, ask open-ended questions and specific tasks, etc.), however, the inspector was free to carry out other tasks if they were necessary to support the inspection process.

After participants carried out the inspections with *checklist* U2CHATBOT, we asked them to respond to a second post-inspection evaluation form containing objective questions that could be completed with four answer options: (1) Totally Agree; (2) Partially Agree; (3) Partially Disagree; and (4) Totally Disagree, according to the *Likert* [Dalmoro and Vieira, 2014] scale. Furthermore, the form also contained subjective questions about changes to the checklist and suggestions for improvements. The form was based on the TAM technology acceptance model, adapted from the work of [Frazão, 2021; Davis et al., 1989]. The inspector could answer the questions according to his level of agreement on the items Perceived Ease of Use (FU); Perceived Usefulness (U) and perceived future Use Intention (IU). To prevent inspectors from giving their opinions on one of these objective questions, the option "Neutral" or "Neither Disagree nor Agree" was not available. The first 7 questions are about Perceived Ease of Use, the next 7 questions are about Perceived Usefulness; and the last 3 questions are about the perceived Future Use Intention. All objective questions of the TAM model can be found [here](#).

Participation in the study took place remotely. Each inspector had an average of 12 days to complete their tasks, including downloading the checklist, installing the indicated chatbot on their mobile devices and inspecting the indicated chatbot for defects. At the end of the participation period, we considered 29 *checklists* completed by inspecting the chatbot *Bing Chat with AI*. An analysis of the discrepancies reported by inspectors was carried out. In total 82 discrepancies were reported, but upon analysis only 78 were in fact real defects and 4 were false positives. To reduce bias from biased opinions, the authors did not classify any of the reported discrepancies; The classification was carried out by another researcher with extensive experience (10 years) in software inspections. In the following subsections, the results of the analyses are presented.

5.1 Quantitative analysis

To carry out the quantitative assessment, we used the results of the discrepancy analysis to measure the performance of the U2CHATBOT inspection checklist. From the data collected, the number of real defects identified, the false positives and the time spent reported by the inspectors were counted. The objective of the quantitative analysis was to verify whether the collected data established any relationship through sta-

tistical tests carried out using the statistical tool SPSS (Statistical Package for the Social Sciences) from the company IBM (International Business Machines Corporation). The collected data can be found [here](#).

For the statistical test, we defined the independent variable called "**Experience**", of the ordinal qualitative type, to inform the level of experience of the study participants, with three possible response options: *Low*, *Medium* and *High*, however, we assign numbers to qualitative variables to make them quantitative in order to enable statistical analysis [Moya, 2021]. The experience dimension used was Inspection Experience. The other dimensions Education Level, Professional Experience, Development Experience and Experience with Chatbots weren't considered for the statistical test as they weren't balanced in the data.

The dependent variables "**Effectiveness**" and "**Efficiency**" are continuous quantitative variables, having been defined to measure the performance of study participants. The effectiveness and efficiency metrics applied to this research are the same ones used by the authors Fernandez et al. [2013] and Frazão [2021]. The Equations 1 and 2 present the effectiveness and efficiency formulas, respectively. In the equations, "**d**" refers to the number of defects identified by a participant in their inspection; "**D**" is the total number of defects identified after analyzing all discrepancies; and "**t**" refers to the individual time spent by the inspector to carry out his inspection.

$$Effectiveness = (d / D) * 100 \quad (1)$$

$$Efficiency = d / (t / 60) \quad (2)$$

The data collected is of the unpaired type, as the groups of participants do not have dependencies between them, and there is no repetition of participants in the data collected. To check the normality of the data, we performed the Shapiro-Wilk test (recommended for samples smaller than 50 [Mishra et al., 2019]) in SPSS, to identify whether the data distribution is normal or not. It is important to highlight that the number of participants (n=29) in this study indicates that the sample is small, however, according to Indrayan and Mishra [2021] small samples do not necessarily produce invalid information, this depends on the research context. Second, smaller samples, provide quick results and are easy to obtain ethics committee approval [Indrayan and Mishra, 2021]. The normality test results were 0.010 for Effectiveness and 0.006 for Efficiency. Considering the level of significance ($p > 0.05$) for the data to be considered normal, we conclude that the distribution of the data is not normal, that is, the data are not parametric.

Depending on the defined factors and the type of data collected for this quantitative evaluation, the statistical test chosen was Kruskal-Wallis, ideal for analyzing non-parametric data with 3 groups in the independent variable [Almeida et al., 2022]. The hypotheses established for the statistical test in relation to **Effectiveness** are:

- H_0 : There are no significant differences in effectiveness between the groups with low, medium and high inspection experience.

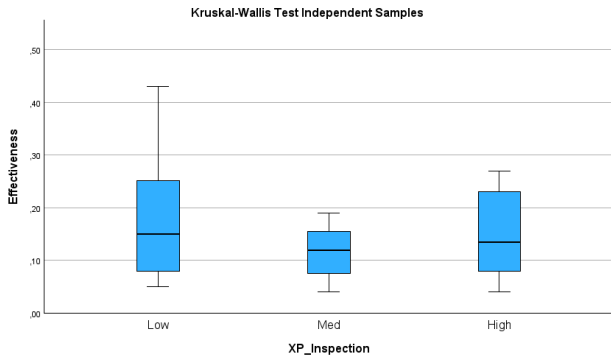


Figure 1. Boxplot Effectiveness

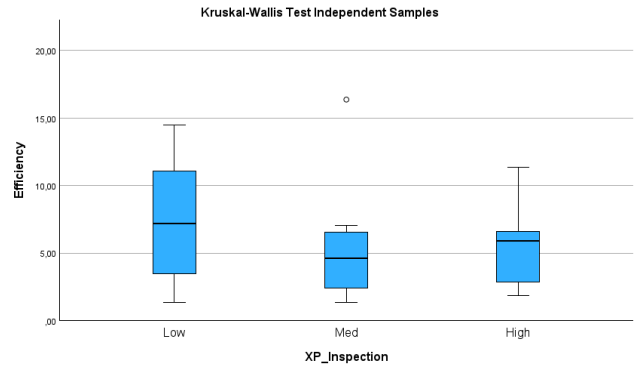


Figure 2. Boxplot Efficiency

- H_1 : There is at least one significant difference in the effectiveness of the low, medium and high inspection experience groups.

Performing the Kruskal-Wallis Test in SPSS, for the **effectiveness** data, we obtained the result ($p = 0.622$). Considering that the stipulated significance level is $\alpha > 0.05$, this means that the null hypothesis H_0 was accepted and there is no statistically significant difference in effectiveness between the groups. For the H_1 hypothesis to be accepted, the p -value should be less than or equal to 0.05 [Moya, 2021].

Observing the graph *Box plot* (generated by SPSS) that compares the effectiveness indicator, presented in Figure 1, it is noted that the quartiles of the groups with Medium and High Experience in Inspection are less distributed, leading believe that they may have performed slightly better than the group with Low Inspection Experience. This descriptive result suggests that the U2CHATBOT checklist inspection had promising effectiveness with inspectors with a certain degree of experience, even if the differences between the groups weren't statistically significant.

In turn, the hypotheses established for the statistical test in relation to **Efficiency** are:

- H_0 : There are no significant differences in efficiency between the groups with low, medium and high inspection experience.
- H_1 : There is at least one significant difference in the efficiency of the groups with low, medium and high inspection experience.

When performing the same Kruskal-Wallis statistical test, also in the SPSS tool, for **efficiency** data, the result obtained was ($p = 0.584$). This also means that the null hypothesis H_0 was accepted and there is no statistically significant difference in efficiency between the groups analyzed. The *Box plot* graph (generated by SPSS) that compares the efficiency indicator is shown in Figure 2.

In relation to Figure 2, the graph shows us that the quartiles of the groups with High and Medium Inspection Experience have a smaller distribution in relation to the group with Low Inspection Experience. This result is similar to that in the Efficacy analysis, suggesting that the inspection with the U2CHATBOT checklist had promising efficiency with inspectors with some level of experience, even if the differences between the groups were not statistically significant.

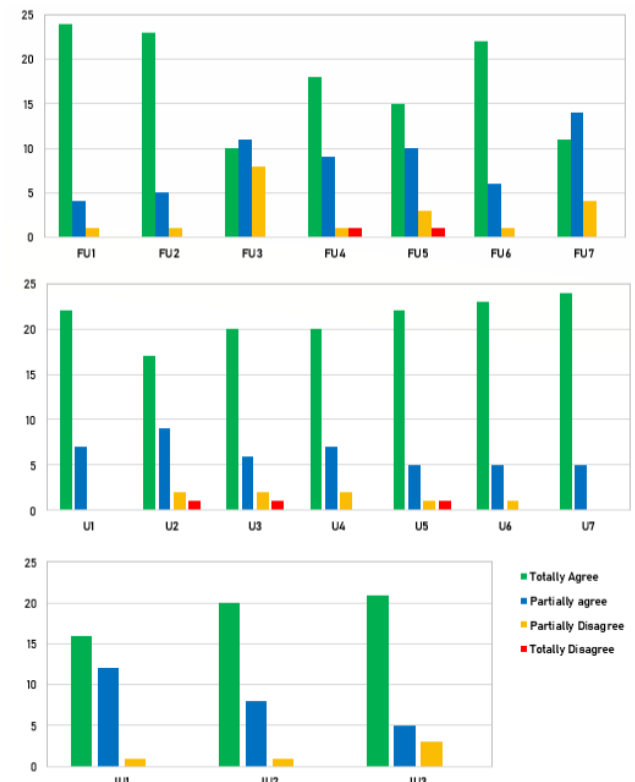


Figure 3. Results on Ease of use (FU), Usefulness (U) and Future Use Intention (UI) on the U2CHATBOT checklist

5.2 TAM Model Analysis

For the evaluation with the TAM model, we considered the objective responses from the post-inspection evaluation form. In this subsection we present the results of the evaluation of the U2CHATBOT checklist with the technology acceptance model (TAM). The Figure 3 presents the data obtained on Perceived Ease of Use (FU), Perceived Utility (U) and Perceived Future Use Intention (IU) according to the judgment of the inspectors who used the U2CHATBOT checklist. All objective questions of the TAM model can be found [here](#).

When analyzing Figure 3, we noticed that in relation to Perceived Ease of Use, the U2CHATBOT checklist was generally well evaluated by the participants, however it is possible to identify some weaknesses, mainly in relation to the aspects Easy Application (FU7), Effort Mental (FU3). Eight participants disagreed that the tool is easy to use to find de-

fects in chatbots and four disagreed that it requires little mental effort.

The Uncomfortable (FU4) and Understandable (FU5) aspects also had discordant evaluations; These results may have been caused by the high number of U2CHATBOT check items, which demands more time and mental effort from the inspector, which can generate confusion and frustration. Regarding Perceived Usefulness, most of the marked items are positive (in green), indicating good results. In fact, few participants disagreed regarding the usefulness of the U2CHATBOT checklist. Regarding the Future Use Intention aspect, many participants agreed, totally or partially, with the intention of using the tool on a future occasion, if there is an opportunity. In the following subsection we present the qualitative analysis.

5.3 Qualitative Analysis

For this qualitative assessment we considered the subjective responses from the post-inspection assessment form. The subjective questions aimed to better understand the biggest difficulties and possible changes or improvements to *checklist* U2CHATBOT. In general, participants' responses indicated that *checklist* has perceived usefulness for evaluating a chatbot.

We have some examples of comments: *"The checklist as a whole was easy to use, the way it was organized and the instructions that came with it made it a lot easier. [I28]"*; and *"The questions, in general, are clear and very objective. The answer options are precise and there is the possibility of detailing the defect, if necessary. [I10]"*. However, some inspectors reported problems such as:

- The problem related to the large number of verification items; some inspectors found the process tiring. For example: *"The time needed to complete the task, the number of questions, some of which were a bit repetitive, made the task a bit tedious. [I28]"* and *"Lots of questions to answer. [I15]"*
- Difficulty understanding some checklist items: *"Certain questions weren't so clear when identifying in the chatbot. [I11]"* and *"Some questions were difficult to answer. [I26]"*

Still on the interpretation of verification items; some inspectors reported that the items needed to be more detailed, as they found it difficult to understand what was being requested by the check item:

- *"I found it a little difficult to identify what certain questions really meant. It would be interesting to include a table with additional, more specific information about each question on the checklist. [I18]"*

This inspector's comment indicates that the verification items in the U2CHATBOT checklist may not be sufficient for some professionals to understand what a chatbot project focused on quality should have. The need for additional information reported by the inspector suggests the need to create an artifact with a visual approach that provides real examples

to developers, facilitating the understanding of important aspects for chatbots.

Inspectors were also asked about the degree of completeness of the *checklist*, that is, whether the *checklist* is complete and covers all important aspects. Most inspectors considered that the verification items addressed the most recurring problems in chatbots, such as: *"The checklist has several specific items, which left nothing out. [I19]"* and *"The checklist is very complete. I believe it covers all aspects of performing a chatbot inspection. [E10]"*. However, inspectors indicated the need to add more response options beyond No, Yes and Not Applicable:

- *"... I think more answer options should be included... [I02]"* and *"I would add an answer option "Partially Applies" because on some occasions it is not possible to say "Yes" or "No" very precisely [I09]"*

Regarding possible suggestions for changes and improvements to *checklist* U2CHATBOT, it is possible to highlight:

- Changes in the answer options available in the checklist. Added more answer options besides No, Yes and Not Applicable.
- Division of the checklist into categories instead of a single, large list.
- Improve the explanation of *checklist* items to be easier to understand.
- Reduction in the number of checklist items to speed up inspection.
- Include examples of the tasks to be performed for each of the assessment items.

In the following subsection, some of these changes and improvements to *checklist* U2CHATBOT are applied.

5.4 Checklist U2CHATBOT Improvements

The experimental study carried out to evaluate *checklist* U2CHATBOT, presented evidence of the viability of the technology. Participants evaluated *checklist* as useful for identifying defects in chatbots, but there are aspects to be improved. According to participants, some items in the *checklist* U2CHATBOT are difficult to understand and too complex to analyze in an inspection, hindering the evaluation process. Furthermore, *checklist* was considered tiresome because it had a large number of check items.

This subsection describes the improvements made in the *checklist* U2CHATBOT refinement, in order to make the technology less tiring and difficult to use, in addition to correcting items that could lead future inspectors to report false positives during an inspection.

We carried out an analysis to identify which U2CHATBOT check items could lead inspectors to mistakenly identify defects; that is, we identified which verification items led inspectors to point out false positives (discrepancies that, in reality, are not real defects) [Kalinowski and Spínola, 2008].

The false positives found in the discrimination meeting on the items on the U2CHATBOT checklist were:

The first item in Table 4, CP-4, aims to understand whether the evaluated chatbot was modeled considering the domain

Table 4. False Positives Found

Checklist Item	U2CHATBOT	Item Description by Inspector
CP-4:	Does the chatbot use domain model from the user's perspective, i.e., is the abstract representation (use case diagrams, classes, etc.) of the chatbot focused on users' needs and behaviors?	Does it avoid responding in a abstract.
FE-6:	When applicable, does the chatbot allow the user to transfer the conversation from the chatbot to a human agent at any time?	Transfer to human agents is not supported.
Q-5:	Does the chatbot perform effective task allocation (decide whether a function will be performed by the system or escalated to a human agent) by providing appropriate escalation channels for humans?	The chatbot does not allow the user to speak to a human agent, all conversations are done by the chatbot only.
F-9:	Does the chatbot encourage users to use the correct syntax to mention others in a team chat, or does it have the ability to understand this automatically when applicable?	Does the chat understand despite incorrect writing and respond as if not there were no spelling errors.

model from the user's perspective, that is, whether the user is the main figure in the abstract representation (class diagrams, use cases, etc.) of the chatbot, however, the inspector interpreted the item incorrectly, considering the answers given by the chatbot. This item, in fact, is complex to understand, especially for beginner inspectors. Item FE-6 deals with the chatbot's ability to transfer the conversation to a human agent. In fact, the chatbot did not have this capability, but it did not constitute a defect, as the item just did not apply to the type of chatbot evaluated. In this case, the inspector should have selected the answer "Not applicable" in *checklist* U2CHATBOT. The same occurs with items P-5 and F-9.

In addition to the false positives found, other items in the *checklist* U2CHATBOT were misinterpreted by some inspectors, that is, although these items were useful for detecting real defects by inspectors in the evaluated chatbot, some inspectors did not understand them correctly. The checklist items with the potential to generate false positives were CP-1, F-20, H-17, DEM-2, PE-1, ARE-4, EC-5 and H-12. In the Table 5 we present some of these problematic items.

When analyzing Table 5, in item F-20 the inspector did not understand that the issue was about the chatbot's ability to be responsive and adapt to various screen sizes. In turn, item H-17 deals with the chatbot's ability to communicate on various subjects, understanding that the user is the one who decides, but the inspector did not understand that "at the same time" is a conjunctive phrase that is equivalent to "while". This lack of understanding by inspectors also occurred with H-12 and the others already indicated. Therefore, we consider it necessary to improve the text of these items that were misin-

Table 5. Potential False Positives

Checklist Item	U2CHATBOT	Item Description by Inspector
F-20:	Can the chatbot adjust to both a larger screen (tablet or laptop) and a smaller one (mobile phone)?	I don't know how to make the screen smaller and have two windows on my phone.
H-17:	Does the chatbot communicate with the user on various topics, at the same time which understanding which active conversation (inputs) the user belongs to?	At the same time, no. When it starts generating the response, it does not let the user send any more messages until it stops generating the response.
H-12:	Does the chatbot negotiate conversational topics discussed with the user?	Very general question, it could be more specific and detailed. The purpose of the question is not clear.

terpreted by some inspectors, to make them clearer for understanding. and reduce the chances of possible false positives.

To reduce the chances of false positives, reduce the complexity of interpreting checklist items for inexperienced inspectors and avoid very restricted items related to a specific type of chatbot, we chose to remove items CP-4, FE-6, P-5 and F-9 that generated false positives when using *checklist* U2CHATBOT, as shown in Table 4. We also removed the item **H-18: "Does the chatbot include errors to increase realism?"**, in which Inspector 16 commented "...I don't consider this a problem, I prefer it to always be right". In a way, making mistakes actually refers to the human characteristic of making mistakes, however, this characteristic in chatbots can cause the user to lack confidence in the quality and veracity of the tool's responses. With this justification, aiming to improve the evaluation technique developed, items CP-4, FE-6, P-5, F-9 and H-18 were removed from *checklist* U2CHATBOT. The complete *checklist* U2CHATBOT with the improvements made can be found [here](#).

Finalizing the *checklist* proposal, we analyzed the possibility of taking advantage of the identified attributes for the development of another artifact that enriches the chatbots quality improvement process. A search was carried out in the scientific literature aiming to identify another technology that could also be designed using these important aspects. The search results pointed to Design Patterns, designed for the chatbot design stage. This technology is presented in the following section.

6 Design Patterns DP-U2CHATBOT

Design Patterns are solutions to trivial problems encountered in the process of developing a technology. The idea consists of capturing the essence of the problems and proposing the appropriate solutions in a compact way [Van Duyne et al., 2007]. According to Vora [2009], design patterns focus on the context of technology use and guide designers in *how*,

when and where a solution can be applied. Furthermore, they serve to describe good design practices and incorporate quality principles into the system.

With the use of design patterns, interface designers have access to real solutions, not just theoretical and abstract concepts; They see an increase in productivity as they reduce the time spent searching for solutions in other references. Patterns are even useful to guide inexperienced designers, with clear textual instructions and visual examples [Gomes et al., 2021].

In the scientific literature it is possible to find works proposing interface design patterns for software applications, such as the study by Nilsson [2009] and the work by Gomes et al. [2021]. However, regarding specific interface design patterns for conversational agents, there are gaps to be filled, as the contributions are still embryonic. In this way, aiming to contribute with guidelines capable of supporting development teams in the process of building chatbots that satisfy their users, this section presents a set of design patterns for chatbots called DP-U2CHATBOT.

The quality attributes used to create design patterns were taken from the systematic reviews in Section 3. The attributes used to develop *checklist* U2CHATBOT were analyzed to verify the feasibility of being transformed into patterns. We identified that of the 107 quality attributes present in the U2CHATBOT checklist, not all of them could be molded into interface design patterns, as some attributes weren't present in the example applications found or made reference to something that could not be described visually. The Table 6 presents some of these attributes used. The complete list of quality attributes that could be transformed into design patterns is [here](#).

Table 6. Design Patterns Attributes

ID	Attributes
DP_FE 1	Selection and management of preferences in relation to the system interface, communication style, degree of human similarity, font size, among others.
DP_FE 2	Option to transfer to human agents, if the chatbot performs any type of customer service.
DP_ACE 1	Conversational tips to improve interaction and ask clarifying questions if the chatbot does not understand any user input.
DP_ACE 2	The chatbot's position on the web page must be easy to locate and access.
DP_EC 2	The answers provided must be accompanied by the sources of information.
DP_F 1	Quick reply buttons allow you to speed up interaction.
DP_F 2	The chatbot web page should be responsive and work well in the format of most devices.
DP_F 3	The font colors should be bright, contrasting with the application background.
DP_F 4	Providing hyperlinks on the topic discussed to complement the information generated by the chatbot.

After completing the analysis of viable quality attributes, the next step was to identify examples of how these attributes were implemented in real chatbots, to make it possible to offer practical implementation examples that could be added to the proposed interface design patterns. In this way, the search for chatbots in mobile or website format began. The applications and websites were chosen following criteria such as free availability and use, as well as content in Portuguese or English.

To identify mobile chatbots, we carried out an automatic search in the Play Store application store, for Android devices, prioritizing those with an extensive number of downloads and an evaluation score of 4 or higher. Websites, in turn, are derived from an automatic search, carried out through the Google search engine. It was necessary to search for several real chatbots in operation, to then evaluate each one and identify whether they presented quality attributes. Chatbots that did not present any of the attributes were discarded. The chatbots that implemented the quality attributes were selected to be used as an example.

At the end of the process, the following applications and websites were discovered and selected: (APP01) Bing AI, (WEB01) Kuki AI, (APP02) Goat Chat, (WEB02) C&A, (WEB03) Octa AI, (WEB04) Sofia Botfriend, (WEB05) Zurich Seguros and (WEB06) Evie Bot. After choosing the applications, an analysis was carried out with the purpose of identifying which quality attributes were present in each of them. Below, we present which quality attributes are present in the mobile chatbots and web applications analyzed:

- **Bing AI:** DP_ACE 1, DP_EC 1, DP_EC 2, DP_F 3, DP_H 2, DP_H 3, DP_H 5, DP_VS 1, DP_CP 1 and DP_P 1, totaling 10 quality attributes.
- **Kuki AI:** DP_FE 1, DP_F 2, DP_F 3, DP_H 4, DP_H 5 and DP_AD 1, totaling 6 quality attributes.
- **Goat Chat:** DP_F 3, DP_F 4, DP_AD 1 and DP_CP 1, totaling 4 quality attributes.
- **C&A:** DP_FE 2, DP_ACE 2 and DP_F 3, totaling 3 quality attributes.
- **Octa AI:** DP_F 1, DP_F 3 and DP_RL 1, totaling 3 quality attributes.
- **Sofia Botfriend:** DP_H 1 and DP_PE 1, totaling 2 quality attributes.
- **Zurich Seguros:** DP_ACE 1 and DP_F 3, totaling 2 quality attributes.
- **Evie Bot:** DP_H 1 and DP_H 4, totaling 2 quality attributes.

In the following subsection, we present important elements of the proposed DP-U2CHATBOT design patterns.

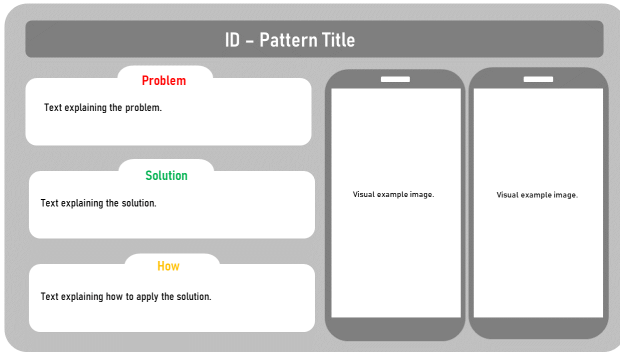


Figure 4. Design Pattern Template

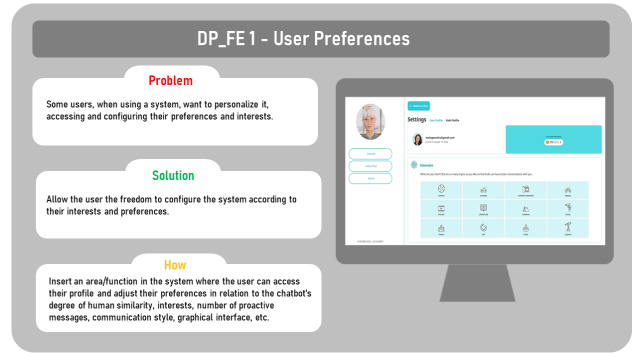


Figure 5. Design Pattern DP_FE 1

6.1 Design Patterns Proposal

So that design patterns can clearly represent what they are, how they work and how they should be applied to solve a design problem, it is important to document them. To develop the documentation of design patterns, the basic elements suggested by Vora [2009] were used, also tested in the work of Gomes *et al.* [2021]. In this way, each pattern developed includes:

- (1) **Code:** Pattern identifier according to its category.
- (2) **Pattern name:** Brief title that expresses what the pattern is about.
- (3) **Problem:** Brief summary of the problem to be solved by the pattern.
- (4) **Solution:** The proposed solution to the presented problem.
- (5) **Implementation:** Instructions and tips on how to apply the solution.
- (6) **Example:** Image containing an example of implementing the pattern based on real chatbots.

With these elements suggested by Vora [2009], it was possible to adapt and structure the set of patterns. Considering this, Figure 4 presents the design pattern model developed for this research. The following subsection presents the set of design patterns developed in this study.

6.2 The set of Design Patterns DP-U2CHATBOT

After carrying out the necessary procedures to create design patterns for chatbots, a set of 21 patterns was obtained, covering problems of humanity, accessibility, functionality, performance, among others. All design patterns developed followed the model in Figure 4. The Figures 5 and 6 presents some patterns from the DP-U2CHATBOT set of design patterns developed.

The DP_FE 1 design pattern in Figure 5 was created containing guidelines on the need to allow the user to personalize the interaction with the chatbot, whether with options relating to changes in the interface, conversational style, similarity human, among other types of personalization. In turn, the DP_EC 2 design pattern in Figure 6 was developed containing solutions on the importance of a chatbot always indicating the sources of information from which it takes its answers. The complete list of all 21 design patterns developed in this

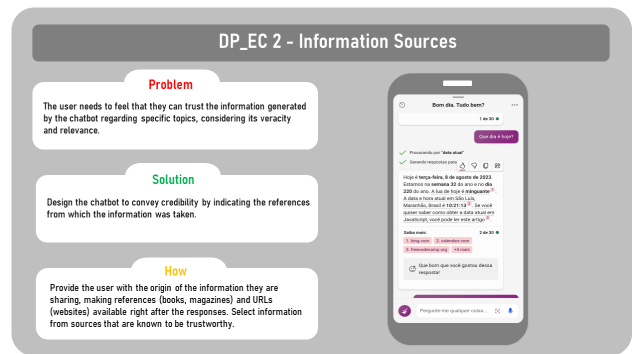


Figure 6. Design Pattern DP_EC 2

study is **here**. The evaluation process and applications for technology improvements are in the following subsection.

7 Design Patterns DP-U2CHATBOT Evaluation and Improvements

Previously, the procedures adopted so that design patterns for DP-U2CHATBOT chatbots could be developed were presented. Taking into account the importance of validating the documentation of DP-U2CHATBOT design patterns, this section presents a study that aimed to analyze whether the design patterns for DP-U2CHATBOT chatbots met the comprehensibility and usefulness indicators defined based on the work of Gomes *et al.* [2021].

The study was carried out in September 2023 with the participation of expert software developers with experience in chatbot projects. All study participants agreed to the Free and Informed Consent Form – ICF. To check the viability of the proposed patterns, we created a questionnaire in the Google Forms tool with three sections: The first section was designed to collect the demographic data of the experts; the second section focused on individually evaluating the comprehensibility and usefulness aspects of each proposed design pattern and the third section focused on an overall assessment of the design patterns.

Regarding the **Comprehensibility** aspect, the question asked for each pattern individually was: “How do you evaluate the comprehensibility (ease of understanding) of each pattern?” For the **Utility** aspect, the question was: “How do you evaluate the usefulness of each pattern?”

The invitation to participate in the evaluation was sent

to professionals with the link to the developed evaluation questionnaire and the set of 21 DP-U2CHATBOT design patterns to be evaluated. After the deadline for participation (1 month) had passed, we received 4 (four) responses to the form. The experts who responded were professors, engineers, and researchers with different levels of experience in chatbot design/development. The average experience of participants with chatbot development is 3.5 years.

During the individual evaluation of each pattern, experts provided a score to assess understandability and usefulness of each of them, with the corresponding scores being: 4 – Very Understandable/Very Useful, 3 – Understandable/Useful, 2 – Not very Understandable/Not very Useful and 1 – Not at all Understandable/Not at all Useful. To make it impossible for participants not to give their opinions on any of the questions applied, the “Neutral” option was not available. The evaluation results are presented in the following section.

7.1 Assessment Results

The designed design patterns were subjected to evaluation regarding comprehensibility and usefulness indicators. All grades assigned to the patterns were counted and, according to the questionnaire developed, they ranged from the highest grade of 4 (Very Understandable/Very Useful) to the lowest grade of 1 (Not at all Understandable/Not at all Useful). The measure adopted to analyze the scores was the calculation of the median, which has the statistical purpose of informing the central position or “the midpoint” of the analyzed values. It is important to highlight that the median calculation was adopted because in the data obtained it is possible to identify extreme values that would affect the arithmetic mean [Pachani, 2006]. The quantitative data of this individual assessment can be found in tables 7 and 8.

The table 7 refers to participants’ scores on the Comprehensibility aspect, that is, how easy the patterns are to understand. According to the median of the scores given by the participants, it is observed that the design patterns were well evaluated in relation to this aspect. The lowest median identified is 3.00, leading us to believe that participants generally found the design patterns understandable.

The design patterns evaluated with the maximum score of 4.00 (Very Understandable) by all participants in the Comprehensibility aspect were: DP_EC 2 and DP_H 2. The design patterns that received at least a score of 1.00 (Not at all understandable) were: DP_EC 1, DP_H 1, DP_H 4, DP_CP 1 and DP_P 1. However, even receiving a score of 1.00 from one participant, the other participants evaluated these standards with scores of 3.00 or 4.00. Therefore, it is possible that the participant was more careful in the evaluation than the others or may not have understood the technology proposal or even this pattern should be refined to improve comprehensibility.

The table 8, in turn, refers to the subjects related to the Utility aspect. According to the median of the scores given by the participants, it can be seen that the patterns were also well evaluated in relation to their appearance, leading us to believe that the participants in general considered the design patterns useful to support the design of chatbots. However, the DP_AD 1 pattern, on *Documentation and Help*, received

Table 7. Subjects given by experts to the patterns considering the Comprehensibility aspect

Pattern	Scores on Comprehensibility				
	Subj.1	Subj.2	Subj.3	Subj.4	Median
DP_FE-1	4,0	3,0	3,0	4,0	3,50
DP_FE-2	4,0	4,0	3,0	4,0	4,00
DP_ACE-1	4,0	4,0	2,0	4,0	4,00
DP_ACE-2	4,0	4,0	2,0	4,0	4,00
DP_EC-1	4,0	4,0	1,0	4,0	4,00
DP_EC-2	4,0	4,0	4,0	4,0	4,00
DP_F-1	4,0	4,0	3,0	4,0	4,00
DP_F-2	4,0	3,0	4,0	4,0	4,00
DP_F-3	4,0	4,0	4,0	3,0	4,00
DP_F-4	3,0	3,0	4,0	3,0	3,00
DP_H-1	4,0	3,0	1,0	4,0	3,50
DP_H-2	4,0	4,0	4,0	4,0	4,00
DP_H-3	4,0	4,0	4,0	3,0	4,00
DP_H-4	4,0	4,0	1,0	4,0	4,00
DP_H-5	4,0	4,0	3,0	4,0	4,00
DP_RL-1	3,0	4,0	2,0	4,0	3,50
DP_VS-1	3,0	4,0	4,0	3,0	3,50
DP_PE-1	3,0	3,0	4,0	2,0	3,00
DP_AD-1	4,0	4,0	3,0	4,0	4,00
DP_CP-1	4,0	4,0	1,0	4,0	4,00
DP_P-1	4,0	4,0	1,0	4,0	4,00

Table 8. Subjects given by experts to the patterns considering the Utility aspect

Pattern	Scores on Utility				
	Subj.1	Subj.2	Subj.3	Subj.4	Median
DP_FE-1	4,0	3,0	3,0	4,0	3,50
DP_FE-2	3,0	4,0	3,0	4,0	3,50
DP_ACE-1	4,0	4,0	2,0	4,0	4,00
DP_ACE-2	4,0	4,0	4,0	4,0	4,00
DP_EC-1	4,0	4,0	1,0	3,0	3,50
DP_EC-2	3,0	4,0	2,0	4,0	3,50
DP_F-1	4,0	4,0	3,0	4,0	4,00
DP_F-2	4,0	3,0	4,0	3,0	3,50
DP_F-3	4,0	4,0	3,0	3,0	3,50
DP_F-4	3,0	3,0	3,0	4,0	3,00
DP_H-1	3,0	3,0	3,0	4,0	3,00
DP_H-2	4,0	4,0	4,0	4,0	4,00
DP_H-3	3,0	4,0	2,0	4,0	3,50
DP_H-4	4,0	4,0	2,0	4,0	4,00
DP_H-5	4,0	4,0	4,0	4,0	4,00
DP_RL-1	4,0	4,0	4,0	4,0	4,00
DP_VS-1	4,0	4,0	4,0	4,0	4,00
DP_PE-1	4,0	3,0	3,0	4,0	3,50
DP_AD-1	3,0	3,0	2,0	2,0	2,50
DP_CP-1	4,0	4,0	4,0	2,0	4,00
DP_P-1	4,0	4,0	4,0	4,0	4,00

Table 9. Subjects given by experts for the established criteria

Criteria	General Utility				
	Subj.1	Subj.2	Subj.3	Subj.4	Average
Criteria 1	4,0	4,0	4,0	4,0	4,0
Criteria 2	4,0	4,0	4,0	4,0	4,0
Criteria 3	4,0	4,0	4,0	2,0	3,5

the lowest median of 2.50, tending to be of little use, according to the participants’ assessment. The design patterns evaluated with the maximum score of 4.00 (Very Useful) by all participants in the Utility aspect were: DP_ ACE 2, DP_H 2, DP_H 5, DP_RL 1, DP_VS 1 and DP_P 1, all with a maximum score of 4.00. The design pattern that received at least a score of 1.00 (Not at all Understandable) was only DP_EC 1, however, the other participants rated this pattern as Useful (3.0) or Very Useful (4.0). Considering the data regarding the Understandability and Utility aspects, the worst grades were given by participants who had less experience in the development/design of chatbots compared to the others. Therefore, the level of experience in this regard may have influenced the analysis of what is useful and understandable in relation to the proposed design patterns.

Participants were also asked how they generally evaluated the usefulness of the patterns according to the criteria: (1) Identifying problems to be avoided in the development of chatbots; (2) Search for solutions to problems identified during the chatbot development process and (3) Understand interface design proposals for a chatbot focused on Usability and User Experience. The subjects relating to these criteria can be found in table 9.

In table 9 we present the general evaluation of the patterns with the grades given by the participants and the average of the grades. In this case, the calculation of the arithmetic mean was adopted, since the data obtained do not present extreme values [Pachani, 2006]. It is observed that the first two criteria defined scored an average score of 4.0 (maximum score), meaning that they were well evaluated by professionals and that they are very useful in these aspects. The third criterion was assessed by one of the participants with a score of 2.0 (Not very useful), however the other participants assessed it with the maximum score of 4.0 (Very Useful), totaling a positive average score of 3.50. Considering this, the participant’s little experience in the design/development of chatbots may also have influenced the analysis of the criterion’s usefulness.

In addition to the evaluation, we asked professionals to indicate up to (5) five design patterns that should be prioritized in the chatbot design/development process. The table 10 presents this prioritization.

According to Table 10, we observed that some participants selected the same patterns to be prioritized, but most of the patterns indicated were different. The most cited design patterns to be prioritized in the chatbot design/development process were:

- DP_FE 2 – Transfer to Human Attendant
- DP_F 2 – Responsive Design

Table 10. Design Patterns that should be prioritized according to evaluators

Participant Evaluator	Design Patterns to be prioritized
1	DP_FE 2 - Transfer to Human Attendant, DP_F 2 - Responsive Design, DP_FE-1 User Preferences, DP_H 5 - Proactivity in Chatbot Communication and DP_P 1 - Robustness to Unexpected Inputs
2	DP_F 1 - Quick Reply Buttons, DP_F 3 - Bright colors for Fonts, DP_H 2 - Understanding Grammatical and Typing Errors, DP_H 4 - Avatar, DP_RL 1 - Descriptive Visual Elements and Explicit Instructions
3	DP_F 2 - Responsive Design, DP_H 2 - Understanding Grammatical and Typing Errors, DP_VS 1 - Feedback and System Status, DP_CP 1 - Spelling and Linguistic Registration, DP_P 1 - Robustness to Inputs Unexpected
4	DP_FE 2 - Transfer to Human Attendant, DP_F 1 - Quick Reply Buttons, DP_EC 2 - Information Sources, DP_RL 1 - Descriptive Visual Elements and Explicit Instructions, DP_VS 1 - Feedback and Status System

- DP_P 1 – Robustness to Unexpected Inputs
- DP_F 1 – Quick Response Buttons
- DP_H 2 – Understanding Grammatical and Typing Errors
- DP_RL 1 – Descriptive Visual Elements and Explicit Instructions
- DP_VS 1 – Feedback and System Status

Still in relation to the design patterns considered priorities, the subjective nature of the evaluators may have influenced the prioritization process, since the concept of what is a priority or not in this context may vary from professional to professional, however, it is something that cannot be said with certainty, as the evaluation sample is very low.

In addition to the objective questions, the Design Patterns evaluation form also contained a field for subjective responses that provided some insights for a qualitative analysis. In general, the experts considered that the patterns developed support the construction of chatbots focused on meeting users’ needs, as highlighted by one of them “...cover the main points in the development of chatbots”. Another expert highlighted: “...they establish a set of details that one would not normally think of for a chatbot”. These comments indicate that the set of patterns developed are capable of assisting in the design of these tools. We also asked whether, in addition to the elements used in defining the technology (code, title, problem, solution, implementation, example), any other information would be added to the definition of the patterns presented. One of the experts commented “I would add a counter example or anti-pattern”, which means that the participant would add examples of what should not be done when designing a chatbot.

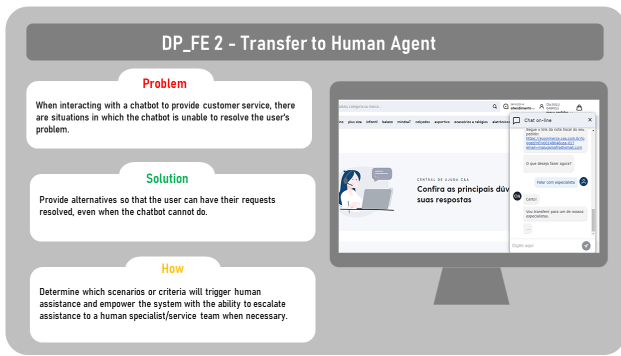


Figure 7. Design Pattern DP_FE 2 after improvement

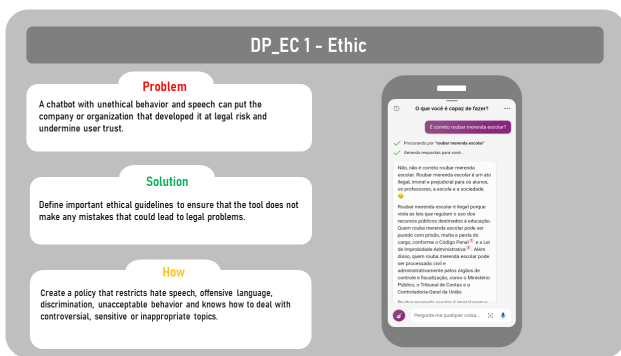


Figure 8. Design Pattern DP_EC 1 after improvement

7.2 Improvements U2CHATBOT Design Patterns

The design patterns evaluation form also contained questions about suggestions for changes and improvements to the design patterns presented in this work. One evaluator pointed out that in some patterns the visual example chosen may not have been the most appropriate *“In the images of each design pattern, I am not sure whether the example shows adherence to the pattern or a lack of adherence to the pattern”*, indicating the item **DP_H 1 - Context Preservation** for improvement in this regard.

Another participant highlighted in relation to the item **DP_EC 1 - Ethics and Cultural Knowledge**, *“...I found ethics to be confused with culture. I think the two things are different and even in some cases, exclusive. Imagine a “tribe” that only speaks slang and swear words. This is the culture of the tribe, but for your chat it would be barred by ethics”*. Furthermore, there was a pertinent guidance of caution when using visual examples containing the logo of the companies responsible for the selected chatbots *“You need to be careful with some images. For example, in DP_FE 2 there is a logo for the company C&A. The same goes for the others. Is it allowed to use?”*.

Considering the subjects and suggestions for improvements indicated by the professionals who participated in the evaluation of the set of U2CHATBOT design patterns developed in this research, changes were made to some design patterns to improve their comprehensibility and also avoid possible future legal problems. The DP_FE 2 design pattern, presented in Figure 7 had its visual example modified so as not to display the logo of the company that owns the chatbot.

This measure was adopted to avoid legal problems. Likewise, other patterns such as **DP_ACE 1** and **DP_F 3** were also refined for the same reason.

In turn, the DP_EC 1 - Ethics and Cultural Knowledge design pattern was fragmented into two, the DP_EC 1 - Ethics pattern and the DP_EC 3 - Cultural Knowledge pattern. This is because these design patterns separately make more sense than together, as they deal with different subjects. In Figure 8 it is possible to find the DP_EC 1 design pattern, already updated.

Other design patterns were improved, for example, the **DP_H1 - Context Preservation** pattern, which the chosen visual example did not have the best grip on the chatbot’s ability to maintain context between conversation sessions, however, There is not enough space in this publication to present all the changes. The complete list of DP-U2CHATBOT design patterns with the improvements applied can be found [here](#).

8 Discussions and limitations

The motivation of our study was to contribute with technologies to facilitate the work of developers in the stages of designing and evaluating text-based chatbots. The study by Guerino and Valentim [2020] identified one of the gaps that motivated our work. The authors conducted research on evaluation methods for chatbots focused on Usability and User Experience. As a result, they identified a lack of evaluation techniques focused on Usability and User Experience that have been submitted and tested according to empirical studies. Additionally, the study by Mafra [2023] analyzed user comments of a chatbot on the Google Play Store and found user complaints, confirming dissatisfaction with the technologies available in the market. Regarding Design Patterns for chatbots, our exploratory research identified that the subject is still in its infancy. In the scientific literature, there are already design patterns to assist in the construction of web applications and to assist in the construction of technologies for the autistic community. However, there were no design patterns with useful recommendations for building better chatbots. Below we will discuss some points and limitations of the work.

Discussions

In the first stage of our research methodology, we conducted systematic literature reviews to build upon previous work and identify quality attributes for chatbots related to Usability and User Experience. Two systematic reviews were necessary to ensure that important quality attributes were not overlooked. From the list of attributes found, we proposed the U2CHATBOT inspection checklist with 107 verification items to identify defects in chatbots. The U2CHATBOT inspection checklist includes items from various techniques and attributes of previous works. This makes it a more comprehensive option than others for analyzing various aspects impacting Usability and User Experience in a chatbot. Additionally, as a distinctive feature, we developed an automated spreadsheet to facilitate the inspection process with the checklist.

The evaluation of the U2CHATBOT checklist was conducted to verify if the artifact indeed aids in discovering defects in chatbots. Software Engineering professionals were invited to inspect a chatbot using the developed checklist. Following the inspection, participants responded to a questionnaire with both objective and subjective questions, providing their opinion on the technology. The experimental study was conducted using the evaluation results, and findings indicate that the U2CHATBOT checklist is capable of assisting developers in finding defects in chatbots, including defects less known to inspectors. However, we must highlight the tool's weaknesses regarding ease of use. Due to its many verification items, the tool tends to require more time and mental effort from inspectors. This issue can be minimized if the inspection is conducted in parts, focusing on the categories to be evaluated.

The findings from evaluations conducted on the U2CHATBOT checklist indicated that for some professionals, verification items alone do not provide a clear understanding of what a chatbot truly requires. For this professional profile, there is a need for visual support to facilitate comprehension. Therefore, we developed the DP-U2CHATBOT Design Patterns set using attributes identified in systematic reviews and examples drawn from functioning real-world chatbots. These patterns were crafted to serve as a guide of design recommendations for chatbots and to address developers' uncertainties. Evaluation of the Design Patterns was carried out with experts in chatbots and design patterns to assess the comprehensibility and usefulness of each pattern. Experts analyzed each pattern and responded to a questionnaire containing both objective and subjective questions. From this, we conducted a quantitative analysis. Qualitative analysis was not extensively pursued as not all participants responded to the subjective questions. Results from the evaluation of DP-U2CHATBOT Design Patterns indicated that the technology was well-received, considered useful, and comprehensible in supporting chatbot design.

Limitations

Regarding the ethical aspects of this research, all study participants read and agreed to the Free and Informed Consent Form, designed to make them aware of the objective of the research and the confidentiality of the information requested. It is worth emphasizing that this research was not submitted to an Ethics Committee. According to Amorim *et al.* [2019] research, opinion polls with unidentified participants or those based on educational and/or professional practices and that do not reveal data that identifies the participant, should not be registered or evaluated by the CEP/CONEP system, being the case of this present study. Although this study was not submitted to the Ethics Committee, the identification of the participants was preserved and any negative impacts of their participation are non-existent. Participants were only invited to give their opinion on the two technologies developed in this work. The entire study was conducted remotely (both the Checklist and Design Patterns assessment) and participants had no contact with each other.

The following section presents the conclusions of this study.

9 Conclusions

Although there are studies focused on proposing technologies to evaluate and enhance the quality of chatbots, we noticed that users still express dissatisfaction with these tools. For instance, in app stores, it's common to find numerous complaints and negative reviews from users. To address this issue, the development of an inspection checklist was proposed to identify defects in text-based chatbots, containing quality attributes focused on Usability and User Experience extracted from systematic literature reviews. This is because both Usability and User Experience are crucial factors to consider when aiming to ensure the success of a system and the satisfaction of its users.

After developing U2CHATBOT, we conducted an empirical study involving 29 participants capable of inspecting a chatbot with the checklist to assess its usability. The evaluation results suggested that the U2CHATBOT checklist is useful for identifying defects in chatbots. However, the number of verification items constitutes a weakness, as it affects the perceived ease of use of the technique. On the other hand, by including more verification items, U2CHATBOT becomes a better option for cases requiring thorough and comprehensive inspections. Statistical tests conducted indicated that there were not many statistically significant differences between inspectors with high or low inspection experience; however, the effectiveness and efficiency in using the checklist may be subtly influenced by the level of inspection experience. In other words, experienced inspectors may achieve slightly better results than inspectors with less experience.

In addition to the U2CHATBOT checklist, we observed a lack of contributions in the chatbot design stage. Studies on design patterns for these tools are still in their infancy. Thus, we utilized the discovered quality attributes to develop 21 practical design patterns to aid developers in constructing these chatbots. The technology was also evaluated by four software developers experienced in chatbots. The professionals assessed each pattern individually and responded to a questionnaire designed to gauge the comprehensibility and usefulness of the technique. The results indicated that participants generally found the design patterns to be helpful and understandable. Some patterns underwent modifications to refine their proposal. These changes primarily involved visual examples to ensure greater adherence to the pattern, removal of logos, and fragmentation of a pattern into two for enhanced clarity. Following these refinements, 22 design patterns were derived.

It is essential to highlight that the small number of participants in the evaluation of the U2CHATBOT checklist may have influenced the results not showing statistically significant differences between the groups of inspectors, as we considered only 29 participants who completed all stages of the study. Similarly, the small number of experts who evaluated the DP-U2CHATBOT design patterns may have influenced the results of comprehensibility and usefulness of the technology. Considering this, attracting more participants in future work can contribute to improving the quality of results and providing greater credibility.

Overall, the evaluation results demonstrated that the tech-

nologies developed in this research were well-received by the participants, proving useful in aiding software engineers and developers, regardless of their experience, in evaluating and enhancing their chatbots. These results also open up space for new perspectives that can be explored in future work, such as updating usability and UX quality attributes to include voice-activated chatbots in the checklist inspection and creating anti-patterns to highlight real examples of what should be avoided in chatbot design.

Acknowledgment

We would like to thank everyone involved so that this research could be developed. Thank you to the study participants. Thank you to Journal On Interactive System and its renowned editors for the opportunity. This work was carried out with the support of the Coordination for the Improvement of Higher Education Personnel – Brazil (CAPES) – Financing Code 001. The authors would like to thank the support of the Maranhão Research and Scientific Development Support Foundation (FAPEMA) and the National Council of Scientific and Technological Development (CNPq).

Authors' Contributions

Malu Mafra contributed to the Conceptualization, Methodology, Writing – original draft, Writing – review & editing, Data curation and Investigation of this study. Mafra is the main contributor and writer of this manuscript.

Kennedy Nunes contributed to the Data curation and Investigation. Simara Rocha, Geraldo Braz Jr, Aristófanés Silva, Davi Viana and Williamson Silva contributed to the Validation, Investigation and Writing – review & editing.

Luis Rivero contributed to the Supervision, Conceptualization and Methodology of this study.

All authors read and approved the final manuscript.

References

- Adamopoulou, E. and Moussiades, L. (2020). An overview of chatbot technology. In *IFIP international conference on artificial intelligence applications and innovations*, pages 373–383. Springer. DOI: https://doi.org/10.1007/978-3-030-49186-4_31.
- Almeida, D. C., Pitanga, H. N., Silva, T. O. d., Silva, N. A. B., and Avelar, M. G. d. (2022). Utilização dos testes estatísticos kruskal-wallis e mann-whitney para avaliação de sistemas de solos reforçados com geotêxteis. *Matéria (Rio de Janeiro)*, 27. DOI: <https://doi.org/10.1590/1517-7076-RMAT-2021-45351>.
- Alsayed, A. O., Bilgrami, A. L., and Foster, W. (2017). Improving software quality management: testing, review, inspection and walkthrough. *International Journal of Latest Research in Science and Technology*, 6(1):7–12.
- Amorim, P. F., Sacramento, C., Capra, E. P., Tavares, P. Z., and Ferreira, S. B. L. (2019). Submit or not my hci research project to the ethics committee, that is the question. In *Proceedings of the 18th Brazilian Symposium on Human Factors in Computing Systems*, pages 1–11. DOI: <https://doi.org/10.1145/3357155.3358473>.
- Anshu, K., Gaur, L., and Solanki, A. (2021). Impact of chatbot in transforming the face of retailing-an empirical model of antecedents and outcomes. *Recent Advances in Computer Science and Communications (Formerly: Recent Patents on Computer Science)*, 14(3):774–787. DOI: <https://doi.org/10.2174/2213275912666190809110804>.
- Barbosa, M., Valle, P., Nakamura, W., Guerino, G., Finger, A., Lunardi, G., and Silva, W. (2022). Um estudo exploratório sobre métodos de avaliação de user experience em chatbots. In *Anais da VI Escola Regional de Engenharia de Software*, pages 21–30. SBC. DOI: <https://doi.org/10.5753/eres.2022.227723>.
- Borsci, S., Malizia, A., Schmettow, M., Van Der Velde, F., Tariverdiyeva, G., Balaji, D., and Chamberlain, A. (2021). The chatbot usability scale: the design and pilot of a usability scale for interaction with ai-based conversational agents. *Personal and Ubiquitous Computing*, pages 1–25. DOI: <https://doi.org/10.1007/s00779-021-01582-9>.
- Brill, T. M., Munoz, L., and Miller, R. J. (2019). Siri, alexa and other digital assistants: a study of customer satisfaction with artificial intelligence applications. *Journal of Marketing Management*, 35(15-16):1401–1436. DOI: <https://doi.org/10.1080/0267257X.2019.1687571>.
- Bryczynski, B. (1999). A survey of software inspection checklists. *ACM SIGSOFT Software Engineering Notes*, 24(1):82. DOI: <https://doi.org/10.1145/308769.308798>.
- Cabrejos, L. J. E. R., Viana, D., and dos Santos, R. P. (2018). Planejamento e execução de estudos secundários em informática na educação: Um guia prático baseado em experiências. *Jornada de Atualização em Informática na Educação*, 7(1):21–52.
- Chaves, A. P. and Gerosa, M. A. (2021). How should my chatbot interact? a survey on social characteristics in human–chatbot interaction design. *International Journal of Human–Computer Interaction*, 37(8):729–758. DOI: <https://doi.org/10.1080/10447318.2020.1841438>.
- Ciechanowski, L., Przegalinska, A., and Wegner, K. (2018). The necessity of new paradigms in measuring human–chatbot interaction. In *Advances in Cross-Cultural Decision Making: Proceedings of the AHFE 2017 International Conference on Cross-Cultural Decision Making, July 17-21, 2017, The Westin Bonaventure Hotel, Los Angeles, California, USA 8*, pages 205–214. Springer. DOI: https://doi.org/10.1007/978-3-319-60747-4_19.
- Codina, L. (2005). Scopus: el mayor navegador científico de la web. *El profesional de la información*, 14(1):44–49. Available in: <https://www.epn.edu.ec/wp-content/uploads/2017/03/Scopus-el-mayor-navegador.pdf>.
- Coppola, R. and Ardito, L. (2021). Quality assessment methods for textual conversational interfaces: A multivocal literature review. *Information*, 12(11):437. DOI: <https://doi.org/10.3390/info12110437>.
- Cruz, Y. P., Collazos, C. A., and Granollers, T. (2015). The thin red line between usability and user experiences. In *Proceedings of the xvi international conference on human computer interaction*, pages 1–2. DOI:

- <https://doi.org/10.1145/2829875.2829915>.
- Dalmoro, M. and Vieira, K. M. (2014). Dilemas na construção de escalas tipo likert: o número de itens e a disposição influenciam nos resultados? *Revista gestão organizacional*, 6(3). DOI: <https://doi.org/10.22277/rgo.v6i3.1386>.
- Davis, F. D., Bagozzi, R. P., and Warshaw, P. R. (1989). User acceptance of computer technology: A comparison of two theoretical models. *Management science*, 35(8):982–1003. DOI: <https://doi.org/10.1287/mnsc.35.8.982>.
- De Souza Monteiro, M., da Silva Batista, G. O., and de Castro Salgado, L. C. (2023). Investigating usability pitfalls in brazilian and foreign governmental chatbots. *Journal on Interactive Systems*, 14(1):331–340. DOI: <https://doi.org/10.5753/jis.2023.3104>.
- Denecke, K. and Warren, J. (2020). How to evaluate health applications with conversational user interface? *Studies in health technology and informatics*, 270:976–980. DOI: <https://doi.org/10.3233/SHTI200307>.
- Fernandez, A., Abrahão, S., and Insfran, E. (2013). Empirical validation of a usability inspection method for model-driven web development. *Journal of Systems and Software*, 86(1):161–186. DOI: <https://doi.org/10.1016/j.jss.2012.07.043>.
- Frazaó, K. et al. (2020). Analyzing app store comments and quality attributes for defining an inspection checklist for mobile educational games. In *Proceedings of the 34th Brazilian Symposium on Software Engineering*, pages 854–859. DOI: <https://doi.org/10.1145/3422392.3422477>.
- Frazaó, K. A. (2021). Ic-meg: Um checklist específico para avaliação de jogos educacionais digitais em plataformas móveis. Master's thesis, Universidade Federal do Maranhão. Available in: <https://tedebc.ufma.br/jspui/handle/tede/4437>.
- Georgescu, A.-A. et al. (2018). Chatbots for education—trends, benefits and challenges. In *Conference proceedings of eLearning and Software for Education «(eLSE)»*, volume 2, pages 195–200. “Carol I” National Defence University Publishing House. DOI: <https://doi.org/10.12753/2066-026X-18-097>.
- Gomes, B. R., Jacob Jr, A. F. L., Pinto, I. d. J. P., and Colcher, S. (2020). Ágata: um chatbot para difusão de práticas para educação ambiental. *Anais Estendidos do XXVI Simpósio Brasileiro de Sistemas Multimídia e Web*, pages 85–89. DOI: https://doi.org/10.5753/webmedia_estendido.2020.13068.
- Gomes, D., Pinto, N., Melo, A., Maia, I., Paiva, A., Barreto, R., Viana, D., and Rivero, L. (2021). Developing a set of design patterns specific for the design of user interfaces for autistic users. In *Proceedings of the XX Brazilian Symposium on Human Factors in Computing Systems*, pages 1–7. DOI: <https://doi.org/10.1145/3472301.3484347>.
- Guerino, G. C. and Valentim, N. M. C. (2020). Usability and user experience evaluation of conversational systems: A systematic mapping study. In *Proceedings of the 34th Brazilian Symposium on Software Engineering*, pages 427–436. DOI: <https://doi.org/10.1145/3422392.3422421>.
- Hassan, H. M. and Galal-Edeen, G. H. (2017). From usability to user experience. In *2017 International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS)*, pages 216–222. IEEE. DOI: <https://doi.org/10.1109/ICIIBMS.2017.8279761>.
- Indrayan, A. and Mishra, A. (2021). The importance of small samples in medical research. *Journal of Postgraduate Medicine*, 67(4):219. DOI: https://doi.org/10.4103/jpgm.JPGM_230_21.
- Kalinowski, M. and Spínola, R. O. (2008). Introdução à inspeção de software. *Revista Engenharia de Software: Qualidade de software*, 1:68–74. Available in: <https://www-di.inf.puc-rio.br/~kalinowski/publications/KalinowskiS07.pdf>.
- Madan, A. and Dubey, S. K. (2012). Usability evaluation methods: a literature review. *International Journal of Engineering Science and Technology*, 4(2):590–599.
- Mafrá, M. G. S. (2023). Desenvolvimento de artefatos para apoiar o design e a avaliação de chatbots focando em usabilidade e user experience. Master's thesis, Universidade Federal do Maranhão. Available in: <http://www.tedebc.ufma.br:8080/jspui/handle/tede/5135>.
- Mirrig, A. G., Meschtscherjakov, A., Wurhofer, D., Meweweger, T., and Tscheligi, M. (2015). A formal analysis of the iso 9241-210 definition of user experience. In *Proceedings of the 33rd annual ACM conference extended abstracts on human factors in computing systems*, pages 437–450. DOI: <https://doi.org/10.1145/2702613.2732511>.
- Mishra, P., Pandey, C. M., Singh, U., Gupta, A., Sahu, C., and Keshri, A. (2019). Descriptive statistics and normality tests for statistical data. *Annals of cardiac anaesthesia*, 22(1):67. DOI: https://doi.org/10.4103/aca.ACA_157_18.
- Motaung, T. (2022). *Design attributes for a successful Online Retail Chatbot Information System*. PhD thesis, University of Johannesburg. Available in: <https://ujcontent.uj.ac.za/esploro/outputs/9921405707691>.
- Moya, C. R. (2021). Como escolher o teste estatístico: um guia para o pesquisador iniciante. Master's thesis, Universidade Cruzeiro do Sul. Available in: <https://www.sbquadriil.org.br/app/uploads/2021/10/Como-escolher-o-teste-estati%CC%81stico-Um-guia-para-o-pesquisador-iniciante.pdf>.
- Muñoz, L. and Avila, O. (2019). A model to assess customer alignment through customer experience concepts. In *International Conference on Business Information Systems*, pages 339–351. Springer. DOI: https://doi.org/10.1007/978-3-030-36691-9_29.
- Nilsson, E. G. (2009). Design patterns for user interface for mobile applications. *Advances in engineering software*, 40(12):1318–1328. DOI: <https://doi.org/10.1016/j.advengsoft.2009.01.017>.
- Pachani, R. A. (2006). Cálculo e uso de mediana. *Exacta*, 4(2):417–423. Available in: <https://www.redalyc.org/pdf/810/81040222.pdf>.
- Radziwill, N. M. and Benton, M. C. (2017). Evaluating quality of chatbots and intelligent conversational agents. *arXiv preprint arXiv:1704.04579*. DOI:

- <https://doi.org/10.48550/arXiv.1704.04579>.
- Rahman, A., Al Mamun, A., and Islam, A. (2017). Programming challenges of chatbot: Current and future prospective. In *2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC)*, pages 75–78. IEEE. DOI: <https://doi.org/10.1109/R10-HTC.2017.8288910>.
- Rapp, A., Curti, L., and Boldi, A. (2021). The human side of human-chatbot interaction: A systematic literature review of ten years of research on text-based chatbots. *International Journal of Human-Computer Studies*, 151:102630. DOI: <https://doi.org/10.1016/j.ijhcs.2021.102630>.
- Reichert, L., Park, G. W., and Rogers, Y. (2022). Extending chatbots to probe users: Enhancing complex decision-making through probing conversations. In *Proceedings of the 4th Conference on Conversational User Interfaces*, pages 1–10. DOI: <https://doi.org/10.1145/3543829.3543832>.
- Rosruen, N. and Samanchuen, T. (2018). Chatbot utilization for medical consultant system. In *2018 3rd technology innovation management and engineering science international conference (TIMES-iCON)*, pages 1–5. IEEE. DOI: <https://doi.org/10.1109/TIMESiCON.2018.8621678>.
- Sharma, P. (2021). Review paper on contextual chatbot for covid-19 updates. *IITM Journal of Management and IT*, 12(1):36–37. Available in: <https://www.indianjournals.com/ijor.aspx?target=ijor:iitmjmit&volume=12&issue=1&article=008>.
- Sharma, R. K. and Joshi, M. (2020). An analytical study and review of open source chatbot framework, rasa. *Int. J. Eng. Res*, 9(06):1011–1014. DOI: <https://doi.org/10.17577/ijertv9IS060723>.
- Sharma, V., Goyal, M., and Malik, D. (2017). An intelligent behaviour shown by chatbot system. *International Journal of New Technology and Research*, 3(4):263312.
- Sperli, G. (2020). A deep learning based chatbot for cultural heritage. In *Proceedings of the 35th Annual ACM Symposium on Applied Computing*, pages 935–937. DOI: <https://doi.org/10.1145/3341105.3374129>.
- Sugisaki, K. and Bleiker, A. (2020). Usability guidelines and evaluation criteria for conversational user interfaces: a heuristic and linguistic approach. In *Proceedings of the Conference on Mensch und Computer*, pages 309–319. DOI: <https://doi.org/10.1145/3404983.3405505>.
- Suhaili, S. M., Salim, N., and Jambli, M. N. (2021). Service chatbots: A systematic review. *Expert Systems with Applications*, 184:115461. DOI: <https://doi.org/10.1016/j.eswa.2021.115461>.
- Thorat, S. A. and Jadhav, V. (2020). A review on implementation issues of rule-based chatbot systems. In *Proceedings of the international conference on innovative computing & communications (ICICC)*. DOI: <http://dx.doi.org/10.2139/ssrn.3567047>.
- Van Duyne, D. K., Landay, J. A., and Hong, J. I. (2007). *The design of sites: Patterns for creating winning web sites*. Prentice Hall Professional.
- Vora, P. (2009). *Web application design patterns*. Morgan Kaufmann. DOI: <https://doi.org/10.1016/B978-0-12-374265-0.X0001-1>.