

Designing and Evaluating In-Situ Assistive Features to Anticipate Text-Based Responses of Conversational Agents

Michelle Xiao Lin Foo

LMU Munich
Munich, Germany
memichelleful@gmail.com

Ilhan Aslan

Aalborg University
Aalborg, Denmark
ilas@cs.aau.dk

Sven Mayer

LMU Munich
Munich, Germany
TU Dortmund University
Dortmund, Germany
info@sven-mayer.com

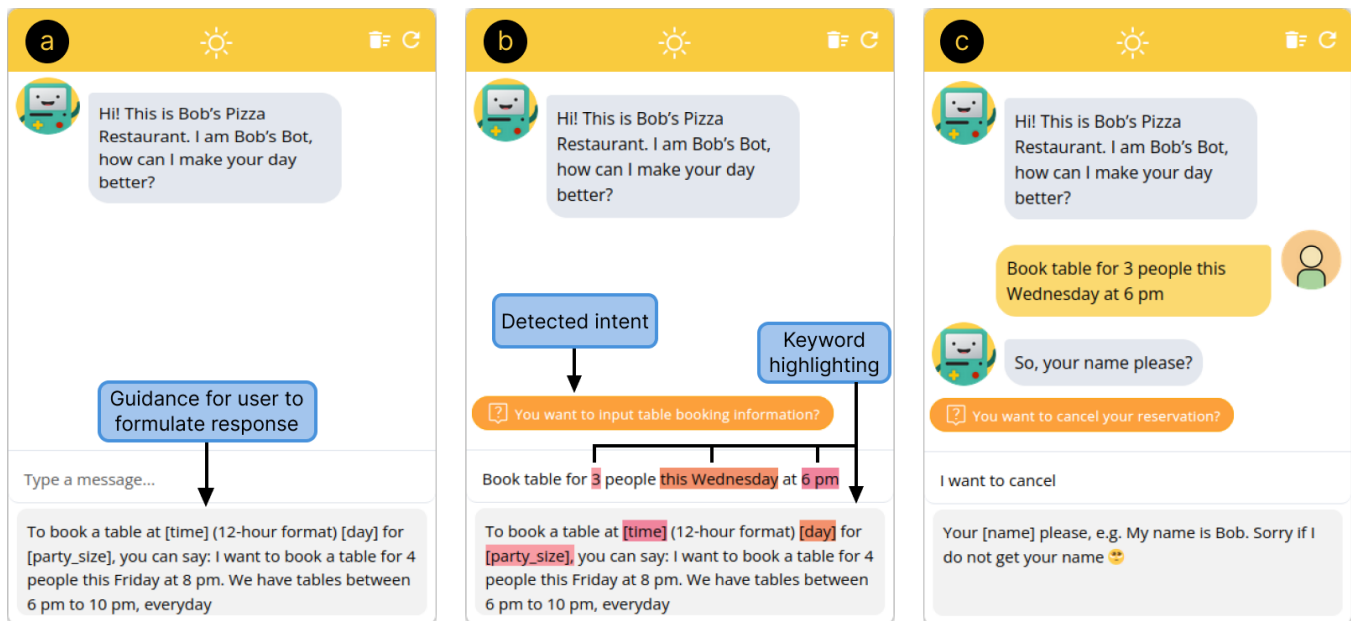


Figure 1: Illustration of in-situ assistive features providing real-time feedback during conversational agent (CA) interactions. a) The CA opens the conversation and displays a guidance box below the input field. b) As the user types, detected intent and relevant keywords are highlighted. c) Guidance updates dynamically based on the ongoing context.

Abstract

While conversational agents have become increasingly capable, users still often struggle to understand their capabilities and anticipate their responses. To address this, we employed a three-stage approach: a) a participatory design workshop (N=9), b) an online survey (N=26), and c) an online study (N=30). From these, we derived key design considerations for developing chat assistive features for text-based conversational agents that enhance user interactions in both task-oriented and social-conversational contexts. We implemented and evaluated a novel in-situ chat assistance interface that displays detected intents, highlight keywords, and provides dynamic response guidance. Our findings show that these features

improve transparency, reduce user effort, and support more effective conversations, although user perceptions differed between the two conversational contexts. This work contributes validated design insights and an adaptable implementation to inform the development of more supportive chat interfaces.

CCS Concepts

• **Human-centered computing** → **Interactive systems and tools.**

Keywords

conversational agent, chat assistance, text-based, chatbot, in-situ assistance

ACM Reference Format:

Michelle Xiao Lin Foo, Ilhan Aslan, and Sven Mayer. 2025. Designing and Evaluating In-Situ Assistive Features to Anticipate Text-Based Responses of Conversational Agents. In *CHIItaly 2025: 16th Biannual Conference of the Italian SIGCHI Chapter (CHIItaly 2025)*, October 06–10, 2025, Salerno, Italy. ACM, New York, NY, USA, 14 pages. <https://doi.org/10.1145/3750069.3750140>



This work is licensed under a Creative Commons Attribution 4.0 International License. *CHIItaly 2025, Salerno, Italy*

© 2025 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2102-1/25/10
<https://doi.org/10.1145/3750069.3750140>

1 Introduction

With the rapid advancement of Large Language Models (LLMs), we are witnessing a surge in sophisticated conversational agents (CAs) [37, 39] over simple, rule-based chatbots [2]. However, some users still struggle to communicate their intentions effectively with CAs. One of the key challenges lies in the skill mismatch between users and CAs, as users lack an understanding of the agents' capabilities. Since CAs with limited capabilities typically recognize only specific terms, users frequently have to experiment with different reformulations to obtain the desired information. This trial-and-error process can be frustrating, making it difficult to explore the system's capabilities without encountering numerous unsupported commands [44]. Further, users of LLMs-based CAs often engage in iterative prompt refinement to obtain responses that more accurately align with their desired outputs [28, 34, 46]. This process underscores the susceptibility of LLMs-based CAs to ambiguity and misalignment in interpreting user intent [42]. Therefore, there exists a need for mechanisms that help users navigate and understand CA's capabilities more effectively, and to craft inputs that elicit more accurate and aligned responses.

As defined by Deriu et al. [18], task-oriented systems facilitate structured conversations to help users accomplish specific tasks efficiently, while social-oriented systems support open-ended interactions without a predefined goal. Previous research has explored various methods to enhance user understanding of CAs, such as integrating chat assistance within conversational turns [3, 45] or offering selectable response suggestions [21, 24, 29]. However, to the best of our knowledge, no existing implementation has provided real-time, in-situ guidance and feedback on user input to help anticipate the CA's response. This gap presents an opportunity to explore a novel approach to chat assistance. We hypothesize that offering real-time guidance and feedback before a user submits their input can serve as an alternative to traditional reply suggestions in the form of selectable options. We assume such an approach can enhance user exploration and improve long-term anticipation of CAs responses, leading to more effective interactions.

First, we identified user needs to anticipate text-based responses of CA by conducting a participatory design workshop [30] (N=9). We asked them to design and discuss assistive features for task-oriented and social conversational contexts. This informed us on establishing key design considerations and criteria for developing in-situ chat assistive features. Second, we designed an in-situ chat assistive prototype that allows users to anticipate the responses. Through an online survey (N=26), we evaluated the design and evaluated the potential usefulness of our non-functional prototype. Third, we implemented a functional prototype using the Rasa framework [8]. Finally, we conducted an online study (N=30) to compare the proposed in-situ chat assistance against the conventional approach of providing reply suggestions in the form of predefined options in a task-oriented and social conversational context.

From design workshops, we extracted key design considerations using affinity diagramming. We verified our design via a non-functional prototype before implementing and evaluating a functional prototype of the chat assistance. Together, we reinforced the notion that task-oriented and social conversational contexts have different needs; however, we could show that our prototype

supported users. Our contributions are as follows: (1) a set of key design considerations and criteria for developing chat-assistive features to support user interactions with text-based CAs. These insights help designers to understand user needs and guide them in designing chat-assistive features. (2) by translating these insights into a chat assistance system for both task-oriented and social-conversational contexts, we provide initial insights into how such features influence user experience with CAs.

2 Related Work

We review previous research on breakdowns, repairs, and guidance in human-conversational agent communication. We then examine studies on support interfaces with CAs, highlighting why more research is needed to understand user needs better and develop effective chat-assistive features.

2.1 Breakdowns, Repairs, Guidance in Human-Agent Communication

A conversational breakdown occurs when there is a mismatch in understanding between the user and the CA. To address this issue, various repair and guidance strategies have been proposed and studied to assess their effectiveness in preventing or resolving such breakdowns [4]. In task-oriented contexts, Ashktorab et al. [3] explored user preferences for eight repair strategies derived from communications theories. Their findings indicate that users favor strategies that provide actionable resources, such as options or explanations, to resolve conversational breakdowns with CAs effectively. Based on the analysis of simulated user interactions with task-oriented CAs, Dippold [19] suggested that explicitly communicating interaction rules with CAs can significantly improve the success of breakdown repair strategies. Beyond repair strategies, some approaches focus on preventing conversational breakdowns altogether. For example, Yeh et al. [45] proposed design recommendations for guidance strategies in task-oriented CAs, emphasizing factors like guidance type and timing. Their study found that example-based guidance is particularly effective when provided at the beginning of a task, though they also noted that the interplay between guidance type and timing is highly scenario-dependent. Braggaar et al. [10] examined various repair strategies and found that deferring to human customer support had the strongest positive impact on trust and brand attitude, followed by the strategy options, with the repeat strategy having the least effect.

To summarize, prior work focused on investigating various strategies to prevent or repair breakdowns. However, none of them explore the design to display those strategies from CA users' perspective. This leads us to ask the following research question (**RQ1**): *What are the key design factors for developing in-situ chat assistive features for CAs in task-oriented and social-conversational contexts?*

2.2 Supporting Interaction through CAs' Design

Effective interface design is essential to facilitate seamless communication between humans and CAs. Jain et al. [26] introduced a context view at the top of the chat interface to help users better understand the CA's dialogue state during conversations. Their form-based interface allowed users to provide precise input using interactive elements. Gao and Jiang [21], Hohenstein and Jung

[24], Jiang and Ahuja [29] explored the use of CAs to offer reply suggestions in human-to-human conversations. Li et al. [33] developed a multi-modal interface combining speech and direct manipulation, leveraging existing mobile application GUI resources for grounding, enabling users to discover, diagnose, and recover from conversational breakdowns effectively. Additionally, Khurana et al. [31] proposed in-application explainable CA interfaces that provide insights into the CA’s underlying processes during breakdowns. Their design improved users’ understanding of the causes of breakdown and positively impacted perceptions of usefulness, transparency, and trust. Brachman et al. [9] explored design recommendations for supporting interactions with and understanding natural language systems. They found that users are likely to benefit from a system that provides specific terms and phrases, along with explanations tailored to their current context and mental framework.

Thus, prior work explored various designs to support the use and understanding of CAs. However, they have not examined chat assistance that offers real-time, in-situ guidance and feedback on user input to allow users to understand CAs and anticipate their responses. Thus, we address this with the research question (RQ2): *How can we help users better understand the capabilities of text-based conversational agents and anticipate their responses?*

3 In-situ Chat Assistance Design Exploration: Participatory Design Workshop

To identify design factors for in-situ chat assistance features (RQ1), we conducted a participatory design workshop. This approach actively engaged CA users in the design process, ensuring their needs and preferences were addressed in the feature development process.

3.1 Procedure

After welcoming all participants, we gave them an overview. After answering all open questions, we asked participants to sign an informed consent form. We then started the recording. The workshop was structured into three parts. First, we facilitated a general discussion on text-based CAs. Second, we asked participants in groups of 2-3 to create design sketches envisioning possible chat assistive features. To support this activity, we provided conversational scenarios. Once the sketches were completed, each group presented their designs, and we asked other participants to evaluate them, focusing on their strengths and weaknesses. Third, we asked participants to propose design criteria they considered important for in-situ chat assistive features. The participants then rated the proposed design criteria on a seven-point Likert scale, assessing their importance in both task-oriented and social-oriented contexts. The workshop lasted ~ 140 min, which we compensated with 20€.

3.2 Materials

The conversational scenarios used in the design workshop varied across several dimensions, including the type of context (task-oriented or social-oriented), the formality of the language used in the dialogue, and the user profiles associated with them. For each scenario, we presented two versions of a dialogue: one showing how the conversation would unfold without the assistance feature and another illustrating how the conversation would flow with the feature. We carefully crafted the dialogues to provide participants with

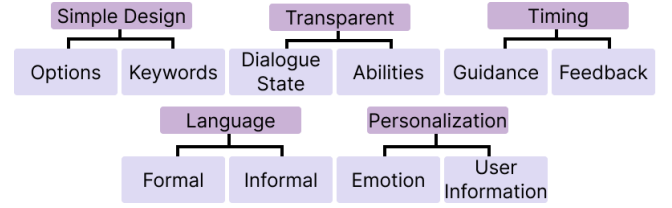


Figure 2: Design considerations for chat assistance based on participants’ design sketches and discussions.

concrete examples to inspire their design sketches. To further guide the design process, we created two distinct user profiles. These profiles highlighted characteristics such as the users’ experience with CAs, their level of trust in such systems, and their familiarity with chat applications. This approach aimed to encourage participants to create designs that addressed these diverse aspects.

3.3 Participants

We recruited 9 participants through our university’s mailing list. Their ages ranged from 19 to 34 years ($M = 25.8$, $SD = 4.2$), and they identified as female (3), male (5), and non-binary (1). All participants are university students with prior experience interacting with text-based CAs on mobile and web interfaces. All participants stated they did not view themselves as CA experts. Three participants reported interacting with CAs more than 5 times in the past 12 months, while the remaining participants interacted less than 5 times or not at all. When asked to rate their most recent interaction with a text-based CA, participants’ experiences varied widely, ranging from very good to extremely poor. The results of the pre-study screening questionnaire revealed that participants had highly diverse experiences with CAs, ensuring a broad range of perspectives and minimizing potential bias in their feedback.

3.4 Results of the Design Workshop

We recorded a total of 2.4 hours of video data. We used affinity diagramming [23] to analyze the discussions. To examine the design sketches, we compiled all designs and identified key design considerations based on participants’ remarks and feedback. We then synthesized the results and identified five recurring themes to explore design considerations for in-situ chat assistance in CA interactions, as illustrated in Figure 2.

3.4.1 Design Considerations for Chat Assistance.

Simple Design with Options and Keywords. All participants created chat assistance designs that incorporated options. P1, P3, P5, and P8 designed features that utilized keywords. For option-based designs, P2 and P6 found them easy to understand and use. However, P6 noted that while presenting options works well for short queries, it may be insufficient for more detailed inputs. P5 expressed concern that this design could overly restrict user interactions with the CA, and suggested that additional interaction assistance should be provided beyond simply displaying options. As for keyword-based designs, P1 and P3 envisioned a design where the CA builds context from user-provided keywords and prompts the user to select the most relevant option. They also suggested iteratively refining the

information displayed based on the keywords chosen by the user. However, P4 and P5 argued that typing keywords felt too similar to a search engine and seemed somewhat unnatural.

Transparent Dialogue State and Abilities. P5, P6, and P8 emphasized the importance of making the CA’s abilities and dialogue state transparent to users as part of the assistive features. P6 noted, “*The CA seems to know what it’s capable of and what it isn’t capable of. This means the user will always know that the answer they receive isn’t a redirection just so that the agent can say it has given an answer. You always trust someone who knows when they don’t know.*” Similarly, P8 added, “*If you ask it and it gives you what sort of questions that it can solve and what sort of functionality that it has so that I can at least know whether it will ever be able to solve that problem or not.*” Regarding how the CA should convey its understanding of user input, P5 and P8 proposed an assistive feature that highlights or underlines keywords. P8 explained, “*If it just copies my statement from earlier. It does not feel like it understood. So, it should also highlight the keywords that it takes.*”

Timing of Guidance and Feedback. P9 stated that the timing of the guidance and feedback shown to the user should be fast. P8 noted that the system should be aware of the user’s current state and provide appropriate guidance when the user hesitates too long, without the need for the user to ask for it.

Language Style Used for In-Situ Chat Assistance. Regarding the formality of language used in the feature, P2 and P7 sketched designs that uses informal language to build user trust. P6 supported this approach, noting that informality fosters trust, as people are more likely to trust those who appear “friendly.” However, P5 argued that the language style is less important as long as the information is conveyed clearly to the user. P4 added, “*I think this is a cultural thing. If it is English, it’s fine with one because they don’t distinguish, but for other languages that do this kind of thing, then you would need completely two different versions.*”

Personalization of In-Situ Chat Assistance. P1, P2, P3, P5, and P8 believed that chat assistance should adapt to the user based on both interaction data and user-provided information. Personalization of the feature could enhance engagement with the CA (P3, P5, P6, P8), but P5 was uncertain whether it would also increase trust, stating, “*For me personally, a personalized design I think makes it less trustworthy, due to privacy reasons.*” Further personalization through user emotion detection was highlighted by P4 and P5. However, P9 cautioned that such detection could lead to misunderstandings between the user and the CA – for example, when the user displays emotions not directed towards the CA, but the system misinterprets it and takes action. P1 and P9 also raised concerns about potential privacy issues associated with emotion detection.

3.4.2 Identified Design Criteria. To identify key design criteria essential for in-situ chat assistance across different conversational contexts, we first compiled a list of design criteria through discussions with the participants, as shown in [Table 1](#). Next, we asked the participants to assess their importance using a seven-point Likert scale. [Figure 3](#) shows the participant assessments of the design criteria across two conversational contexts. In the task-oriented context, two design criteria received mean ratings below 4: C4 ($M = 3.56$,

$SD = 2.3$) and C5 ($M = 3.89$, $SD = 2.42$). In the social-oriented context, only C8 ($M = 3.89$, $SD = 2.32$) was rated as unimportant by the majority of the participants.

We first conducted Shapiro-Wilk tests to assess the normality for each design criterion and identify the appropriate statistical method. The Shapiro-Wilk tests showed that ratings were not normally distributed. Based on this outcome, we used a non-parametric test to compare participants’ evaluations of each criterion between task-oriented and social-oriented contexts. Wilcoxon signed-rank tests indicated that conversational context had a significant impact on C4: Rated significantly more important in social-oriented context ($Mdn = 7$) than task-oriented context ($Mdn = 4$); $W = 3$, $p = .035$, and C9: Rated significantly more important in task-oriented context ($Mdn = 7$) than social context ($Mdn = 6$); $W = < .001$, $p = .041$. For the remaining design criteria, the differences in importance ratings between the two conversational contexts were not statistically significant. See [Section A.4](#) for the complete table.

4 Design of the In-situ Chat Assistance

Building on the insights from the participatory design workshop, we developed the initial design of the in-situ chat assistance, see [Figure 1](#). We focused on creating a design that works for both task-oriented and social-oriented contexts. Our goal was to build a unified in-situ assistance framework that supports both scenarios. To ensure familiarity and usability, we based our design on the common chat widget and enhanced it with additional in-situ assistance components. Participants wanted actionable resources in the chat assistance. Thus, inspired by Yeh et al. [45], we presented guidance as rules and examples. We displayed this guidance in a box below the user input field, helping users formulate their responses. The guidance provides examples of expected responses and suggestions for crafting answers effectively (see [Figure 1](#)). Participants also

Table 1: The nine design criteria for in-situ chat assistance suggested by the participants, along with their explanations.

Criteria	Explanation
C1 Contextual info	Keep track of user interaction contextual information.
C2 Feedback right after interaction	Provide feedback right after user types something.
C3 Guidance when user hesitates	Provide guidance when user does not respond.
C4 User emotion detection	Consider user’s emotional state during interaction.
C5 Embodied agent	Communicate in a human-like manner using facial expressions and gestures.
C6 Beyond text input	Support additional input options, such as images and data files, to better understand user intents.
C7 Awareness of website info	Reference elements within the website for more context-aware assistance.
C8 Function to export chat info	Provide options to save or email chat information.
C9 Access to internet	Search online when additional information is needed.

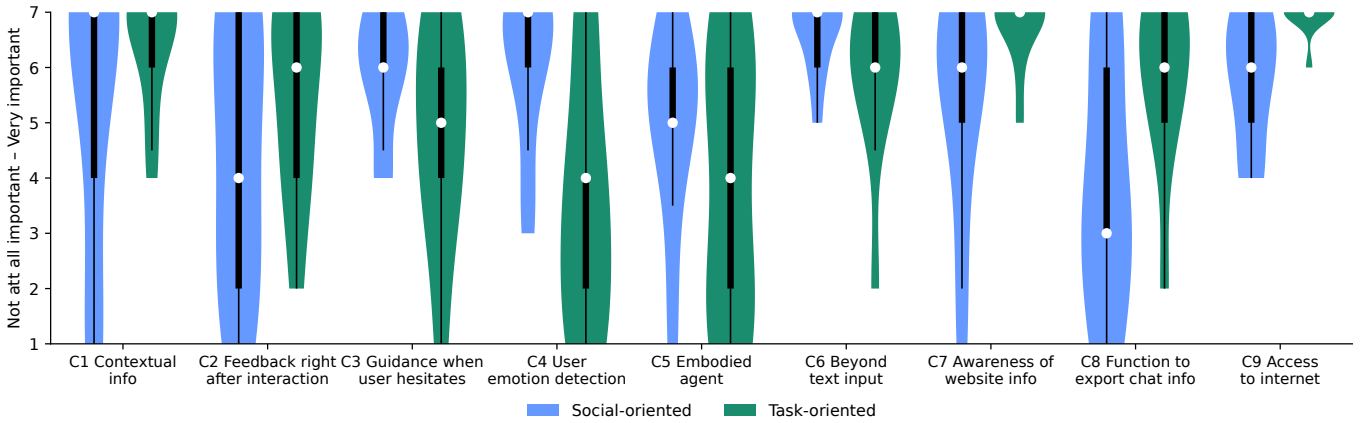


Figure 3: Participants' responses (N=9) on the importance of the nine design criteria across different contexts.

emphasized the importance of making the CA's abilities and dialogue state transparent and providing immediate feedback. Thus, we added a feedback chip above the input box to show the detected intent from the user's input. We also highlighted keywords detected in real-time, both in the user input field and in the guidance box below. This feature gives instant feedback as users type, helping them understand how the CA interprets their input.

4.1 Design Evaluation Online Study

We conducted an online survey to gather feedback on the design of the in-situ chat assistance and explore how CA users perceive it.

4.1.1 Procedure and Survey Design. We first asked participants to interact with a task-oriented CA without the proposed assistive features. We then introduced the proposed design (see Figure 1) along with detailed descriptions of the chat assistance functionality. To demonstrate these features, we used a task-oriented example dialogue similar to the one the participants had experience earlier. To evaluate the task-oriented CA and proposed design, we adapted question items from Cheng et al. [12], which was based on the Technology Acceptance Model (TAM) [16, 17], to measure the following constructs: *Perceived Ease of Use (PEOU)*: The degree to which a person believes that using a particular system would be free of effort [17], *Perceived Usefulness (PU)*: The degree to which a person believes that using a particular system would enhance his job performance [17], and *Intention to Use (INT)*: The participants' intention to use the proposed features with the CA.

There are four question items each for PEOU and PU, whereas there are only three for INT. See Section A.2 for the complete questionnaire. We did not include attitude towards usage (ATU) in this initial study because, according to [38], in the context of adopting information technology innovation, attitude can be synthesized from perceived characteristics of innovating. The revised TAM by [17] excluded ATU as a measured construct. The evaluation questionnaire consists of seven-point Likert scale questions that measure the aforementioned constructs and two open questions on how the chat assistive features might influence the user experience and user behavior while using the CA.

4.1.2 Participants. We distributed the survey through a mailing list and received 26 responses. None of the respondents had participated in the initial participatory design workshop. Among the participants, 20 identified as male and 6 as female. They were aged between 19-30 (1), 31-45 (15), and 46-60 (10). Additionally, 12 participants considered themselves experienced CA users.

4.2 Design Evaluation Results

We analyzed the quantitative data using Python and R, while the open-ended responses were examined using thematic analysis with Atlas.ti [7]. Figure 4 shows the participants' ratings of PEOU, PU, and INT for the rule-based chatbot and proposed in-situ chat assistance design. We identified 2 themes, *impact on user experience of CA* and *impact on user interaction behavior with CA*.

Positive PEOU, PU, and INT. We calculated Cronbach's alpha [14] for the three constructs to evaluate the extent to which the question items measure their intended constructs. All calculated Cronbach's alpha values exceeded the minimum acceptable value of 0.7, ranging from 0.95 to 0.98, indicating strong correlations among the items.

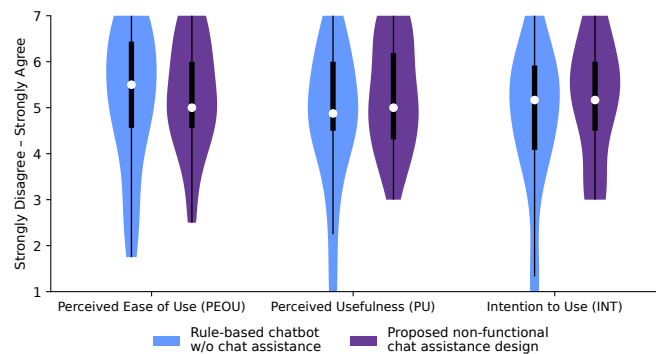


Figure 4: Participants' ratings of Perceived Ease of Use (PEOU), Perceived Usefulness (PU), and Intention of Use (INT), for the rule-based chatbot and proposed in-situ chat assistance design. (N=26)

This suggests that the question items effectively represent the constructs they aim to measure. Similarly, we performed Shapiro-Wilk tests on the ratings. The results indicated that the rating distributions for PEOU, PU, and INT were significantly non-normal. Thus, we used Wilcoxon signed-rank tests to compare the ratings for the rule-based CA and proposed chat assistance design. The Wilcoxon signed-rank test showed a significant difference in the PU ratings for rule-based CA ($Mdn = 4.88$) and proposed design ($Mdn = 5.0$); $W = 41$, $p = 0.03$. See Section A.5 for the complete table. For the proposed design, the overall mean and standard deviation of each construct were as follows: PEOU ($M = 5.26$, $SD = 1.21$), PU ($M = 5.29$, $SD = 1.19$), INT ($M = 5.29$, $SD = 1.28$). These results suggest that participants generally believe using the CA with chat assistance would be free of effort, enhance their task performance, and to some extent, intend to use the proposed assistive features.

Impact on user experience of CA. 19 out of 26 participants agreed that the novel features would be helpful in guiding users to better understand the CA's abilities and limitations, as well as in formulating responses in the "CA's language" (P21). P18 similarly noted, "It will make it easier for me to formulate questions that the CA could understand." P12 highlighted the value of the assistive features for new users, stating, "It will assist users, especially novice users, to understand the interactions more quickly and thus enhance the productivity." Likewise, P3 believed that understanding how to interact with the CA more effectively would enable users to use it more efficiently. However, not all feedback was positive. P6 expressed concern that the chat assistive features might restrict the scope of interaction, potentially leaving the user with an incomplete experience. P7 added, "A CA should be clear enough in its conversation (as if with another human), to not warrant these extra instructions."

Impact on user interaction behavior with CA. 16 out of 26 survey participants answered that the chat assistive features would change user behavior when interacting with the CA, while the remainder either disagreed or had no opinion. P3 noted that these features would boost user confidence during interactions with the CA. Additionally, P4, P12, P15, and P21 pointed out that the chat assistance would help users learn to navigate the CA faster and express their needs more clearly, enabling the CA to understand them better. However, P7 noted that the assistive features might slow down interactions and make them less fluid.

5 In-situ Chat Assistance Implementation

With the design considerations in mind (Section 4), we implemented the in-situ chat assistance prototype to support users in their interactions with CAs by providing guidance for response formulation and real-time feedback on their input, as illustrated in Figure 1.

5.1 Conversational Agent

We developed the CA using Rasa [8], an open-source Python-framework for building custom AI-powered chatbots. Rasa consists of two main components: Rasa NLU and Rasa Core. Rasa NLU interprets user inputs by recognizing intents, extracting entities, and structuring information, while Rasa Core handles state tracking, dialogue management, and response generation. Specifically, Rasa

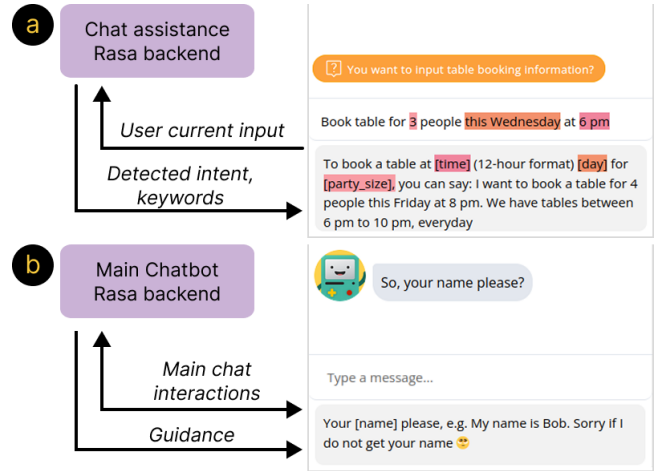


Figure 5: Two Rasa CA backends are used for the chat assistive features, one dedicated to the primary conversation (b) and another for the chat assistive features (a).

Core predicts the next action to perform from a predefined set of options based on the input processed by Rasa NLU.

To evaluate the chat assistive features in CA interactions, we selected scenarios that reflect those frequently used in prior chatbot research [1, 27, 32, 45]. For the task-oriented CA, we implemented a table-booking scenario, while for the social-oriented CA, we developed a simple mental health CA that engages users in conversations about their emotions. Both CAs were designed to carry out conversations with at least five conversation turns.

5.2 Proposed Chat Assistive Features

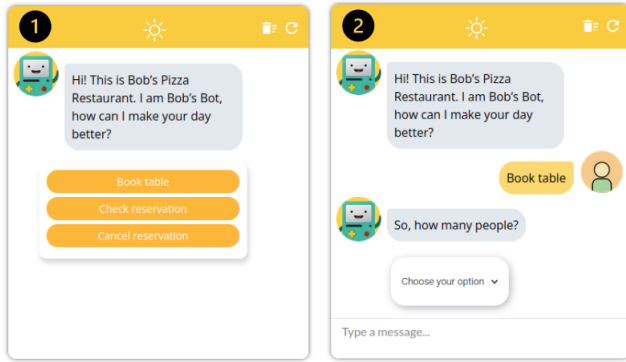
For the functional prototype of the proposed assistive features, we focused on implementing features such as keyword highlighting, user intent explanations, and response guidance. We ensured the chat assistive features share the same knowledge base as the CA by using two separate Rasa CAs – one dedicated to the primary conversation and another for the chat assistive features. Since the response guidance updates dynamically as the user types and after the CA replies, this separation prevents disruptions to the CA's state management. Figure 5 shows the information flow between the chat interface and two separate Rasa backends.

5.3 Front-end and Back-end

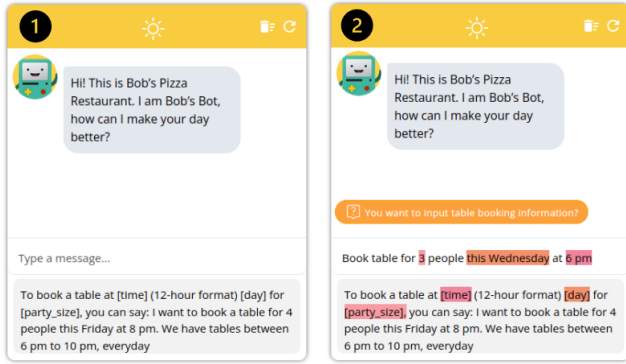
We developed the front-end chat widget using HTML, CSS, and JavaScript, and hosted it on Netlify and the CAs' servers on a local server with 32GB RAM. In addition, we linked the CAs to an online database, MongoDB for logging the interaction data.

6 In-situ Chat Assistance Evaluation

To answer (RQ2), we conducted an online within-subject user study using our in-situ chat assistance. In addition to comparing the new chat assistive features in both social- and task-oriented scenarios, we also evaluated them against the widely used option-based chat assistive features. For each evaluated implementation, we employed a convergent mixed-methods design [13]. We collected



(a) Example of "option-based" chat assistance.



(b) Example of "proposed" chat assistance.

Figure 6: The different chat assistive features for task-oriented conversational context used in the study.

the quantitative data through the Technology Acceptance Model (TAM) [16], as well as trust, reliance, and engagement. As for the qualitative data, we used open-ended questions.

6.1 Designs and Procedure

The online study consists of four main sections, where participants interact with four different CAs across both CONTEXT (*task-oriented* and *social contexts*), with "option-based" or "proposed" chat assistive features (FEATURES) (see Figure 6). Compared to the implementation of the CA with proposed chat assistive features, as shown in Figure 5, we used a single Rasa backend for each implementation of the CA with option-based chat assistance across both conversational contexts. At the start of the study, we presented a statement of consent to the participants, which they must agree to in order to proceed. After answering some demographics questions, a brief guide about the chat assistive features and interface is shown. Next, we asked the participants to use the chat assistive features while interacting with the CA. Since the CA follows a limited set of conversation paths, we provide a conversation flow guide to help participants navigate their interactions. They are free to explore the implementation without any time restrictions. After each interaction, they are asked to rate their experience and respond to open-ended questions. This process is repeated four times, with the estimated total study duration 30 minutes per participant.

6.2 Survey Design

We designed the questionnaire similarly to the online survey described in Section 4.1.1 to assess how the chat assistive features impact usability, ease of use, and intention to use the CA. Additionally, we incorporated questions on user attitude, trust, and engagement with the CA. We further included the attitude toward usage (ATU) construct [16] to evaluate their positive or negative perceptions of using the CA with chat assistance. To gain a more in-depth understanding of user impressions of the chat assistance, we added open questions that probe what they like and dislike about the new features, their overall experience, and the perceived impact of the assistive features on their interaction with the CA.

We presented each rating question as a statement to which the participants had to state their agreement using a slider ranging from *Strongly disagree* to *Strongly agree* on a 101-point scale without ticks in order to reduce response clustering [36]. We opted for visual analog scales over Likert scales, as they provide more precise responses, and thus a higher data quality [20]. Furthermore, they allow for more statistical analyses due to their continuous nature [40]. To identify careless respondents, we had an attention check item after each interaction block [5, 15, 25], randomly instructing participants to move the slider to the left or right. We also counterbalanced the order of CA interface presentation to mitigate sequence effects [41]. See Section A.3 for the complete questionnaire.

6.3 Participants

Out of the 39 participants we recruited, we accepted 30 as valid responses for further analysis. Nine responses were excluded due to the participants failing the attention checks or missing data. The 30 participants (21 from Prolific and 9 from our university mailing list) ranged in age from 19 to 51, ($M = 26.3$, $SD = 6.8$), with the following gender distribution: female (10), male (19), and non-binary (1). Out of the total participants, 16 were students, 12 were full-time employees from various professions and 2 were unemployed. Three participants had previously taken part in the first participatory design workshop. When asked about their expertise with CAs, only 4 participants stated they would consider themselves as experts. The mean expertise level was 3 on a seven-point scale ($SD = 2.0$).

7 Results of Online Study

We processed the quantitative data using Python and R, while the open-ended responses were analyzed through thematic analysis in Atlas.ti [7]. We identified 3 themes.

7.1 Quantitative Results

Participants engaged in more than five conversational turns with the CAs on average, with each interaction lasting between 1.85 and 2.72 minutes. We calculated Cronbach's alpha of the following subscales, PEOU, PU, INT, and ATU, to determine whether the question items are representative of the individual subscales for each interaction context and assistance type. The subscale question items are highly reliable in each evaluated scenario. The Cronbach's alpha values range from .85 to .98 for all measured items.

Our data are not normally distributed. Thus, we then ran an aligned rank transform using ARTool [43] before performing two-way analysis of variance (ANOVAs) to investigate how the context

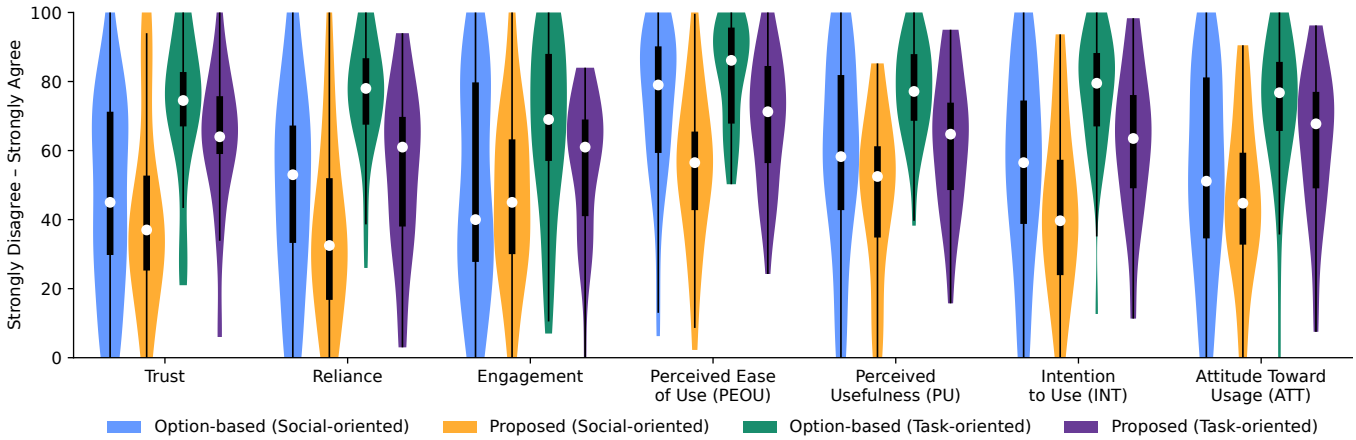


Figure 7: Users' ratings (N=30) of the CA with different types of chat assistance in different interaction contexts.

of the interaction with the CA and the type of assistive features affect the evaluators' scoring of the different constructs.

The two-way ANOVAs in Table 2 show no statistically significant interaction between the effects of the type of assistive features and conversational context across any of the measured constructs. However, the main effects analysis showed that the conversational context did have a statistically significant effect on all constructs ($p < .001$), suggesting that participants evaluated task- and social-oriented CAs differently across all measured dimensions. In contrast, the main effects analysis showed that the type of assistance used in the interaction had a more limited influence. It did not have a statistically significant effect on the engagement ratings, and showed weaker effects on other constructs.

The individual results indicate that task-oriented CAs generally receive higher scores than social-oriented CAs (see Figure 7). However, an exception is observed in PEOU, where social-oriented CA with option-based assistive features received a slightly higher average rating than task-oriented CA with the proposed chat assistance. Additionally, users rated CAs with option-based assistive features

higher than those with the proposed features across all constructs in both conversational contexts.

7.2 User Experience with Proposed Features

Most participants had an overall positive experience interacting with the CAs with the support of the proposed chat assistive features (73% of participants). They found the features helpful, and easy to use. However, many felt that the features were more beneficial in task-oriented interactions than in social conversational contexts (43% of participants). Some participants (P11, P14, P15, P18, P22, P28) noted that the assistive features provided limited support in the social scenario due to the broad nature of the topics. For instance, P18 and P22 pointed out that there were limited suggestions on how to respond to the CA.

Despite these limitations, P2, P4, P6, P10, and P28 found the proposed chat assistance intriguing and recognized their potential in enhancing communication with CAs. P28 remarked, "A pleasant surprise. AI chatbots often misinterpret inputs and you don't always know what they're missing, or what is misunderstood or not interpreted." P10 added, "I felt that when it is fully developed, it can be a huge benefit for users." Additionally, P1 and P17 highlighted that typing responses instead of selecting predefined options felt more like a real conversation.

7.3 Behavioral Impact of Assistive Features

Most participants reported that the proposed chat assistive features changed their interaction behavior with the CAs, regardless of the conversational contexts (90% of participants in task-oriented context and 80% of participants in social context). The features increased their awareness of the CA's understanding limitations, prompting them to adjust their responses based on the guidance and feedback. As P1 explained, "I changed my answer before sending it, trying to make it work for the CA." The participant further commented that with the chat assistance, there were fewer wrong answers. Similarly, P6 noted, "Since I could see what the bot understands by highlighting it, I knew I was on the right track. It gave me confidence in writing out a longer sentence with multiple information."

Table 2: Two-way ANOVAs determined how the type of assistive feature and conversational context influence the measures (N=30). Results show that conversational CONTEXT has a significant effect on all the measures, whereas the FEATURE has a significant effect on all measured constructs except engagement. There are no interactions between the two factors.

Measures	FEATURE			CONTEXT			Interaction		
	F	p	η^2	F	p	η^2	F	p	η^2
Trust	4.863	.03	.053	40.21	<.001	.316	.149	.701	.002
Reliance	16.277	<.001	.158	31.671	<.001	.267	.686	.41	.008
Engagement	2.624	.109	.029	13.306	<.001	.133	1.216	.273	.014
PEOU	31.154	<.001	.264	12.089	<.001	.122	1.316	.254	.015
PU	14.23	<.001	.141	19.704	<.001	.185	.261	.61	.003
INT	17.334	<.001	.166	30.869	<.001	.262	.037	.849	<.001
ATU	7.298	.009	.077	26.201	<.001	.231	.102	.75	.001

However, P30 remarked that the features can lead to over-explaining and extending the conversation beyond CA’s capabilities.

Regarding the option-based assistive features, in a social-oriented context, many participants commented that the options made the conversation paths clearer and simplifies the interactions as they eliminated the need to type responses. P30 noted, *“With options there, I automatically want to just choose from them rather than type.”* However, some felt restricted by the predefined options, which sometimes prevented them from expressing responses that better reflected their thoughts. As P28 commented, *“It made me choose options that were not applicable to me.”* In task-oriented context, many respondents found the option-based assistive features made interactions with the CA more straightforward and faster. P21 noted that the options made it easier to make decisions.

7.4 Design of Proposed Chat Assistive Features

Many participants liked that the proposed chat assistance provided guidance on how to respond to the CAs (P1, P3, P5, P7, P12, P16, P18, P22, P23). In particular, P1, P3, P5, and P18 expressed that they liked having suggestions and examples to assist them in formulating their responses. Additionally, P26 noted that the feature eliminated the need to guess how to interact with the CA.

Several participants (P3, P4, P6, P11, P21, P26, P28) remarked that they appreciated the live feedback provided through intent detection and keyword highlighting. P28 added, *“I like that I could see what the agent thought I meant by my phrases and tells when it is uncertain, or got the wrong idea from my message.”* Similarly, P4 said, *“I liked the feature, which tells you if the agent could actually process your data you are trying to provide.”*

Participants highlighted design concerns regarding the assistive features. P3, P6, and P21 commented that they clicked on the intent bubble above the input text field, expecting it to be interactive. P5 felt that the guidance box was visually unappealing, while P12 found the changing text color distracting and P21 noted that the constantly changing assistive text felt overwhelming.

8 Discussion

To address our research questions (RQ1) and (RQ2), we derived key design considerations through a participatory design workshop and translated these insights into a set of novel chat assistive features. These were then implemented and evaluated.

8.1 Positive Effects of Chat Assistive Features

Our findings show that the proposed assistive features, such as displaying detected user intent, highlighting detected keywords in real-time, and providing response guidance on what the CA expects, improves user understanding of the CAs’ capabilities and allow users to adapt their responses accordingly. While prior work has emphasized the value of providing simple, actionable resources to support user interaction, and to avoid and recover from conversational breakdowns [3, 33], our approach goes beyond static suggestions. By surfacing the CA’s real-time understanding of user intent, we provide users with greater transparency and feedback, allowing them to better calibrate their mental model of the CA. This interpretive layer acts as a conversational mirror that reduces ambiguity, helping users align their communication more effectively.

Our chat assistance design reduced user effort, minimized misunderstandings, and facilitated more effective interactions between users and the CA. Although similar outcomes have been noted in previous studies [3, 26, 33, 45], our work introduces a new way for making the CA’s dialogue state and expectations explicitly visible.

Participants responded positively to both rule-based and example-based guidance at the beginning of every conversational turn. In line with Yeh et al. [45], users reported improved confidence, reduced cognitive effort, and higher accuracy in response formulation. Here, both guidance types aimed to balance quick usability with long-term understanding, enabling users to respond efficiently while developing an intuitive grasp of the CA system’s logic.

Thus, carefully designed, in-situ assistive features can significantly improve user experience by promoting understanding, trust, and alignment in CAs.

8.2 Conversational Context Matters

Our findings revealed that not all design criteria identified by participants are important in both conversational contexts. Participants rated C4–*user emotion detection* and C5–*embodied agent* as particularly important in social-oriented scenarios. This aligns with prior work demonstrating that emotionally responsive systems foster positive affect in social conversational contexts [6, 11, 22]. However, users found C4 and C5 less relevant in task-oriented contexts, where precision, efficiency, and goal completion take precedence. This variation highlights a key design challenge: assistive features must adapt not only to user needs but also to the nature of the interaction, whether it is instrumental or expressive in purpose.

Our findings also showed that conversational context played a more substantial role than feature type in shaping user perceptions. Participants perceived the CA’s limitations more acutely in social interactions, likely due to the open-ended, less structured nature of such conversations. In contrast, task-oriented scenarios offered a clear framework for assistance, which made the proposed features more beneficial and better received. This suggests that users have context-sensitive expectations and that assistive features must be attuned to the context itself, particularly the need for chat assistance to support a broader range of topics in social-oriented contexts.

8.3 Option-based vs. Proposed Chat Assistance

When comparing option-based assistance with our proposal within the same conversational context, users reported higher ratings for the former across constructs such as trust, reliance, perceived ease of use (PEOU), perceived usefulness (PU), intention to use (INT), and attitude toward usage (ATU). One likely reason is familiarity, because option-based assistance is common and intuitive, whereas the proposed features require users to interpret new feedback. This initial skepticism is consistent with findings by Marangunić and Granić [35], which show that unfamiliar technologies often receive lower early-stage ratings. As users grow accustomed to the new system, their perceptions typically improve. Thus, the lower acceptance scores may reflect a transitional phase rather than a fundamental usability issue. This also raises a key challenge in chat assistance design: how to balance simplicity and interpretability without overwhelming the user or introducing friction.

While option-based assistance was rated higher, users felt it constrained their interaction. Suggesting a design paradox: familiar systems are easier to use but may limit flexibility, while novel designs offer richer interactions but require a learning curve.

Our findings do not suggest these two approaches are mutually exclusive. On the contrary, our proposed design is flexible enough to incorporate option-based assistance where appropriate. This adaptability allows designers to blend both approaches to support users, depending on the conversational needs.

8.4 Implications for Chat Assistive Designs

User feedback from the online study pointed to several actionable design improvements. Many participants found the option-based feature easy to use and highly effective in supporting interactions with the CA. To build on this, we could incorporate options directly in the CA's replies for seamless continuation of the conversation. Additionally, we could make the detected intent chips clickable, enabling users to select an intent to respond to the CA.

While most participants valued the concept of response guidance and found it helpful, one criticized its visual design. In future iterations, we could redesign the guidance box to better align with the user interface to minimize cognitive and aesthetic friction. Similarly, for keyword highlighting, some participants found it distracting. To mitigate this and maintain clarity without distraction, we could explore less intrusive alternatives, such as underlining keywords with subtle colored lines or changing their text color.

Beyond surface-level refinements, our evaluations reveal a deeper insight that users expect different kinds of support depending on the conversational context. In social-oriented contexts, participants desired more flexible and nuanced assistance that could accommodate a wider range of topics and emotional tones. This presents a design challenge, as such conversations lack a fixed goal or structured flow that typically guide assistive interventions.

To address this, we envision integrating LLMs to enhance the assistance, enabling more dynamic conversations. Unlike rule-based systems, LLMs can identify a broader range of user intents and extract keywords from complex natural language due to their generalization capabilities. Additionally, LLMs enable the dynamic generation of guidance tailored to each user and dialogue state.

These suggestions converge on a broader insight that effective chat assistive features must be unobtrusive, context-aware, and seamlessly integrated into the chat interface. Users are particularly sensitive to cognitive and visual disruptions, especially in fluid conversations. The adaptability of our chat assistance design positions it well for integration with LLMs, enabling more flexible, intuitive support across both task-oriented and social interaction contexts.

8.5 Limitations and Future Work

The online survey for the initial evaluation of the proposed assistive features design may be subject to desirability bias, as participants could identify the version with proposed improvements. To mitigate this risk and strengthen the internal validity, we could have counterbalanced the presentation order.

Participants observed that the CAs struggled to understand user inputs and engage in topics beyond the predefined scope. As we

wanted the participants to focus on the assistive feature, this limitation was not a concern as both versions were the same. However, these constraints impacted the user experience. While many participants found the assistive features beneficial, their overall ratings were lower due to the CAs' limited conversational capabilities.

Our study did not include a comparison with task-oriented and social CAs without assistive features, as this was not the focus of this paper. However, such a comparison could have provided a more balanced evaluation. Moreover, we evaluated the chat assistance only using one scenario per conversational context, which limits the generalizability. We plan to include multiple scenarios in future evaluations to ensure the applicability of the results. Thus, we plan to assess the improved and extended in-situ chat assistive features, which leverage LLMs to provide in-situ assistance to the user besides providing option-based assistance to users, in combination with LLM-based CAs, across different conversational contexts.

Option-based assistive features received higher ratings for both conversational contexts over our proposed assistive features, which shows that our proposed features still fall short in comparison with today's standard. However, some participants highlighted that text-based interactions, instead of selecting predefined options, felt more like a real conversation, and with the advancement in LLMs, we believe that the proposed chat assistance can prove valuable in supporting text-based interactions with CAs.

While we evaluated user acceptance of the new features through survey questions based on the Technology Acceptance Model, it does not explore cognitive load in-depth. We acknowledge this limitation, particularly in understanding the mental effort required to engage with the assistive features. However, our primary goal was to determine whether users find the feature promising and usable. Future work should conduct longitudinal studies, incorporating cognitive load assessments, to provide a more comprehensive understanding of how the new features affect user interaction.

While our work focuses on in-situ chat assistive features for text-based CAs, the findings may have broader implications, such as voice-based interfaces, accessibility-focused systems, or emergency response CAs, where users may also struggle to understand system capabilities. Future research could explore how the proposed assistive features translate across these modalities or domains.

9 Conclusion

Text-based CAs have seen significant growth in recent years, driven by advancements in LLMs. Despite this progress, many users still face challenges in communicating effectively with CAs. To support user understanding of CA capabilities and anticipation of CA responses, we adopted a three-stage approach to explore design factors for developing in-situ chat assistive features for both task-oriented and social-conversational contexts. Through a participatory design workshop, we identified key design considerations and criteria. Based on these insights, we then implemented and evaluated in-situ chat assistive features to assess their impact on user-CA interactions, uncovering additional design needs specific to different conversational contexts. Our findings can serve as a foundation for designing effective in-situ assistive features for text-based CAs in different conversational contexts.

Open Science

We encourage readers to review, reproduce, and extend our findings. To achieve this goal, we will make our data, code, and analysis script available via OSF <https://osf.io/u2cwf/>.

Author Contributions

Michelle Xiao-Lin Foo: Conceptualization, Formal Analysis, Investigation, Software, Visualization, Writing – original draft, Writing – review & editing; **Ilhan Aslan:** Conceptualization, Supervision, Writing – review & editing; **Sven Mayer:** Conceptualization, Formal Analysis, Funding acquisition, Supervision, Visualization, Writing – original draft, Writing – review & editing

Acknowledgments

This work has been partly supported by the Research Center Trustworthy Data Science and Security (<https://rc-trust.ai>), one of the Research Alliance centers within the UA Ruhr (<https://uaruhr.de>).

References

- [1] Alaa A. Abd-alrazaq, Mohammad Alajlani, Ali Abdallah Alalwan, Bridgette M. Bewick, Peter Gardner, and Mowafa Househ. 2019. An overview of the features of chatbots in mental health: A scoping review. *International Journal of Medical Informatics* 132 (Dec. 2019), 103978. doi:10.1016/j.ijmedinf.2019.103978
- [2] Bayan Abu Shawar and Eric Atwell. 2007. Chatbots: Are they Really Useful? *Journal for Language Technology and Computational Linguistics* 22, 1 (July 2007), 29–49. doi:10.21248/jlcl.22.2007.88
- [3] Zahra Ashktorab, Mohit Jain, Q. Vera Liao, and Justin D. Weisz. 2019. Resilient Chatbots: Repair Strategy Preferences for Conversational Breakdowns. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland UK) (CHI '19). Association for Computing Machinery, New York, NY, USA, 1–12. doi:10.1145/3290605.3300484
- [4] Dennis Benner, Edona Elshan, Sofia Schöbel, and Andreas Janson. 2021. What do you mean? A Review on Recovery Strategies to Overcome Conversational Breakdowns of Conversational Agents. In *International Conference on Information Systems (ICIS)*. 18. https://aisel.aisnet.org/icis2021/hci_robot/hci_robot/13/
- [5] Adam J. Berinsky, Michele F. Margolis, and Michael W. Sances. 2013. Separating the Shirkers from the Workers? Making Sure Respondents Pay Attention on Self-Administered Surveys. *American Journal of Political Science* 58, 3 (Nov. 2013), 739–753. doi:10.1111/ajps.12081
- [6] Timothy Bickmore and Justine Cassell. 2005. Social dialogue with embodied conversational agents. In *Advances in natural multimodal dialogue systems*. Springer, Berlin, German, 23–54. doi:10.1007/1-4020-3933-6
- [7] Ann Blandford, Dominic Furniss, and Stephann Makri. 2016. *Qualitative HCI Research: Going Behind the Scenes*. Morgan & Claypool, San Rafael, CA, United States. <https://api.semanticscholar.org/CorpusID:38190394>
- [8] Tom Bocklisch, Joey Faulker, Nick Pawlowski, and Alan Nichol. 2017. Rasa: Open Source Language Understanding and Dialogue Management. doi:10.48550/arXiv.1712.05181 arXiv:1712.05181 [cs.CL]
- [9] Michelle Brachman, Qian Pan, Hyo Jin Do, Casey Dugan, Arunima Chaudhary, James M. Johnson, Priyanshu Rai, Tathagata Chakraborti, Thomas Gschwind, Jim A Laredo, Christoph Miksovich, Paolo Scotton, Kartik Talamadupula, and Gegi Thomas. 2023. Follow the Successful Herd: Towards Explanations for Improved Use and Mental Models of Natural Language Systems. In *Proceedings of the 28th International Conference on Intelligent User Interfaces* (Sydney, NSW, Australia) (IUI '23). Association for Computing Machinery, New York, NY, USA, 220–239. doi:10.1145/3581641.3584088
- [10] Anouck Braggaar, Jasmin Verhagen, Gabriëlla Martijn, and Christine Liebrecht. 2024. *Conversational Repair Strategies to Cope with Errors and Breakdowns in Customer Service Chatbot Conversations*. Springer Nature Switzerland, Cham, Switzerland, 23–41. doi:10.1007/978-3-031-54975-5_2
- [11] Justine Cassell. 2000. Embodied Conversational Interface Agents. *Commun. ACM* 43, 4 (apr 2000), 70–78. doi:10.1145/332051.332075
- [12] T. C. E. Cheng, David Lam, and Andy Yeung. 2006. Adoption of Internet banking: An empirical study in Hong Kong. *Decision Support Systems* 42 (12 2006), 1558–1572. doi:10.1016/j.dss.2006.01.002
- [13] John W Creswell and Vicki L Plano Clark. 2017. *Designing and conducting mixed methods research*. Sage publications, Thousand Oaks, CA, United States.
- [14] Lee J Cronbach. 1951. Coefficient alpha and the internal structure of tests. *psychometrika* 16, 3 (1951), 297–334.
- [15] Paul G. Curran. 2016. Methods for the detection of carelessly invalid responses in survey data. *Journal of Experimental Social Psychology* 66 (Sept. 2016), 4–19. doi:10.1016/j.jesp.2015.07.006
- [16] Fred Davis. 1985. *A Technology Acceptance Model for Empirically Testing New End-User Information Systems*. Ph.D. Dissertation. Massachusetts Institute of Technology.
- [17] Fred Davis and Fred Davis. 1989. Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. *MIS Quarterly* 13 (09 1989), 319–. doi:10.2307/249008
- [18] Jan Deriu, Alvaro Rodrigo, Arantxa Otegi, Guillermo Echegoyen, Sophie Rosset, Eneko Agirre, and Mark Cieliebak. 2021. Survey on Evaluation Methods for Dialogue Systems. *Artif. Intell. Rev.* 54, 1 (jan 2021), 755–810. doi:10.1007/s10462-020-09866-x
- [19] Doris Dippold. 2023. “Can I have the scan on Tuesday?” User repair in interaction with a task-oriented chatbot and the question of communication skills for AI. *Journal of Pragmatics* 204 (2023), 21–32. doi:10.1016/j.pragma.2022.12.004
- [20] Frederik Funke and Ulf-Dietrich Reips. 2012. Why Semantic Differentials in Web-Based Research Should Be Made from Visual Analogue Scales and Not from 5-Point Scales. *Field Methods* 24, 3 (2012), 310–327. doi:10.1177/1525822X12444061 arXiv:https://doi.org/10.1177/1525822X12444061
- [21] Zihan Gao and Jiepu Jiang. 2021. Evaluating Human-AI Hybrid Conversational Systems with Chatbot Message Suggestions. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management* (Virtual Event, Queensland, Australia) (CIKM '21). Association for Computing Machinery, New York, NY, USA, 534–544. doi:10.1145/3459637.3482340
- [22] Asma Ghandeharioun, Daniel McDuff, Mary Czerwinski, and Kael Rowan. 2019. Towards Understanding Emotional Intelligence for Behavior Change Chatbots. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, New York, NY, USA, 8–14. doi:10.1109/ACII.2019.8925433
- [23] Gunnar Harboe and Elaine M. Huang. 2015. Real-World Affinity Diagramming Practices: Bridging the Paper-Digital Gap. In *Proc. 33rd Annual ACM Conf. Human Factors in Computing Systems*. ACM, New York, NY, USA, 95–104. doi:10.1145/2702123.2702561
- [24] Jess Hohenstein and Malte Jung. 2018. AI-Supported Messaging: An Investigation of Human-Human Text Conversation with AI Support. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) (CHI EA '18). Association for Computing Machinery, New York, NY, USA, 1–6. doi:10.1145/3170427.3188487
- [25] Jason L. Huang, Nathan A. Bowling, Mengqiao Liu, and Yuhui Li. 2014. Detecting Insufficient Effort Responding with an Infrequency Scale: Evaluating Validity and Participant Reactions. *Journal of Business and Psychology* 30, 2 (April 2014), 299–311. doi:10.1007/s10869-014-9357-6
- [26] Mohit Jain, Ramachandra Kota, Pratyush Kumar, and Shwetak N. Patel. 2018. Convey: Exploring the Use of a Context View for Chatbots. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) (CHI '18). Association for Computing Machinery, New York, NY, USA, 1–6. doi:10.1145/3173574.3174042
- [27] Mohit Jain, Pratyush Kumar, Ramachandra Kota, and Shwetak N. Patel. 2018. Evaluating and Informing the Design of Chatbots. In *Proceedings of the 2018 Designing Interactive Systems Conference* (Hong Kong, China) (DIS '18). Association for Computing Machinery, New York, NY, USA, 895–906. doi:10.1145/3196709.3196735
- [28] Ellen Jiang, Kristen Olson, Edwin Toh, Alejandra Molina, Aaron Donsbach, Michael Terry, and Carrie J Cai. 2022. PromptMaker: Prompt-based Prototyping with Large Language Models. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) (CHI EA '22). Association for Computing Machinery, New York, NY, USA, Article 35, 8 pages. doi:10.1145/3491101.3503564
- [29] Jiepu Jiang and Naman Ahuja. 2020. Response Quality in Human-Chatbot Collaborative Systems. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, China) (SIGIR '20). Association for Computing Machinery, New York, NY, USA, 1545–1548. doi:10.1145/3397271.3401234
- [30] Finn Kensing and Andreas Munk-Madsen. 1993. PD: Structure in the Toolbox. *Commun. ACM* 36, 6 (1993), 78–85.
- [31] Anjali Khurana, Parsa Alamzadeh, and Parmit K. Chilana. 2021. ChatrEx: Designing Explainable Chatbot Interfaces for Enhancing Usefulness, Transparency, and Trust. In *2021 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*. IEEE, New York, NY, USA, 1–11. doi:10.1109/VL/HCC51201.2021.9576440
- [32] Heeyoung Kim, Sunmi Jung, and Gihwan Ryu. 2020. A study on the restaurant recommendation service app based on AI chatbot using personalization information. *International Journal of Advanced Culture Technology* 8, 4 (2020), 263–270.
- [33] Toby Jia-Jun Li, Jingya Chen, Haijun Xia, Tom M. Mitchell, and Brad A. Myers. 2020. Multi-Modal Repairs of Conversational Breakdowns in Task-Oriented Dialogs. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology* (Virtual Event, USA) (UIST '20). Association for Computing Machinery, New York, NY, USA, 1094–1107. doi:10.1145/3379337.3415820

- [34] Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegrefe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. SELF-REFINE: iterative refinement with self-feedback. In *Proceedings of the 37th International Conference on Neural Information Processing Systems* (New Orleans, LA, USA) (NIPS '23). Curran Associates Inc., Red Hook, NY, USA, Article 2019, 61 pages.
- [35] Nikola Marangunić and Andrina Granić. 2014. Technology acceptance model: a literature review from 1986 to 2013. *Universal Access in the Information Society* 14, 1 (Feb. 2014), 81–95. doi:10.1007/s10209-014-0348-1
- [36] Justin Matejka, Michael Glueck, Tovi Grossman, and George Fitzmaurice. 2016. The Effect of Visual Appearance on the Performance of Continuous Sliders and Visual Analogue Scales. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) (CHI '16). Association for Computing Machinery, New York, NY, USA, 5421–5432. doi:10.1145/2858036.2858063
- [37] Michael McTear. 2020. Conversational ai: Dialogue systems, conversational agents, and chatbots. *Synthesis Lectures on Human Language Technologies* 13, 3 (2020), 1–251.
- [38] G.C. Moore and I. Benbasat. 1991. Development of an instrument to measure the perceptions of adopting an information technology innovation. *Information systems research* 2, 3 (1991), 192–222.
- [39] Amon Rapp, Lorenzo Curti, and Arianna Boldi. 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 (2021), 102630. doi:10.1016/j.ijhcs.2021.102630
- [40] Ulf-Dietrich Reips and Frederik Funke. 2008. Interval-level Measurement with Visual Analogue Scales in Internet-based Research: VAS Generator. *Behavior research methods* 40 (09 2008), 699–704. doi:10.3758/BRM.40.3.699
- [41] Fritz Strack. 1992. "Order Effects" in Survey Research: Activation and Information Functions of Preceding Questions. Springer New York, New York, NY, 23–34. doi:10.1007/978-1-4612-2848-6_3
- [42] Hari Subramonyam, Roy Pea, Christopher Pondoc, Maneesh Agrawala, and Colleen Seifert. 2024. Bridging the Gulf of Envisioning: Cognitive Challenges in Prompt Based Interactions with LLMs. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 1039, 19 pages. doi:10.1145/3613904.3642754
- [43] Jacob O. Wobbrock, Leah Findlater, Darren Gergle, and James J. Higgins. 2011. The Aligned Rank Transform for Nonparametric Factorial Analyses Using Only Anova Procedures. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vancouver, BC, Canada) (CHI '11). Association for Computing Machinery, New York, NY, USA, 143–146. doi:10.1145/1978942.1978963
- [44] Xi Yang, Marco Aurisicchio, and Weston Baxter. 2019. Understanding Affective Experiences with Conversational Agents. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland UK) (CHI '19). Association for Computing Machinery, New York, NY, USA, 1–12. doi:10.1145/3290605.3300772
- [45] Su-Fang Yeh, Meng-Hsin Wu, Tze-Yu Chen, Yen-Chun Lin, XiJing Chang, You-Hsuan Chiang, and Yung-Ju Chang. 2022. How to Guide Task-Oriented Chatbot Users, and When: A Mixed-Methods Study of Combinations of Chatbot Guidance Types and Timings. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) (CHI '22). Association for Computing Machinery, New York, NY, USA, Article 488, 16 pages. doi:10.1145/3491102.3501941
- [46] J.D. Zamfirescu-Pereira, Richmond Y. Wong, Bjoern Hartmann, and Qian Yang. 2023. Why Johnny Can't Prompt: How Non-AI Experts Try (and Fail) to Design LLM Prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) (CHI '23). Association for Computing Machinery, New York, NY, USA, Article 437, 21 pages. doi:10.1145/3544548.3581388

A Appendix

A.1 Participatory Design Workshop: Pre-study Survey

- Q1 Age [number field]
- Q2 What is your gender identity? [single choice]
 - Female
 - Male
 - [text field]
- Q3 What is your English language level? [single choice between A1 to C2]

Q4 What is your highest education level and academic major? [text field]

Q5 What is your current occupation? [text field]

*General questions about your experience with conversational agents

Q6 I am a conversational agents expert. [7-point Likert scale from strongly disagree to strongly agree]

Q6.1 In the past 12 months, how many times have you interacted with any conversational agents? [single choice]

- 0 times
- < 5 times
- ≥ 5 times

Q6.2 If you have interacted with a conversational agent (CA) before, which CA was the most recent text-based CA you have interacted with? You can also provide a link to the CA [text field]

Q6.3 What was the interaction context? e.g. shopping, customer service, chit chat, etc. [text field]

Q6.4 On which platform did you interact with the CA? [single choice]

- Mobile
- Web
- [text field]

Q6.5 The interaction with the text-based conversational agent was extremely good. [7-point Likert scale from strongly disagree to strongly agree]

Q6.6 Please explain your rating.

A.2 Evaluation of In-situ Chat Assistance Proposed Design: Online Survey

*Demographics

Q1 Age [number field]

Q2 What is your gender identity? [single choice]

- Female
- Male
- [text field]

Q3 I am an experienced chatbot user. [7-point Likert scale from strongly disagree to strongly agree]

*Questions after interacting with rule-based chatbot [7-point Likert scale from strongly disagree to strongly agree]

Q4.1 Using the chatbot is easy for me.

Q4.2 I find my interaction with the chatbot clear and understandable.

Q4.3 It is easy for me to become skillful in the use of the chatbot.

Q4.4 Overall, I find the use of the chatbot easy.

Q4.5 Using the chatbot would enable me to accomplish my tasks more quickly.

Q4.6 Using the chatbot would make it easier for me to carry out my tasks.

Q4.7 I would find the chatbot useful.

Q4.8 Overall, I would find using the chatbot to be advantageous.

Q4.9 I would use the chatbot for my needs.

Q4.10 Using the chatbot for handling my needs is something I would do.

Q4.11 I would see myself using the chatbot for handling my needs.

*Questions about new chat assistance feature

[7-point Likert scale from strongly disagree to strongly agree]

- Q5.1 Using the chatbot with chat assistance feature would be easy for me.
- Q5.2 My interaction with the chatbot would be clear and understandable with the help of the chat assistance.
- Q5.3 It would be easy for me to understand the chatbot with the chat assistance feature.
- Q5.4 Overall, I would find the use of the chatbot with the chat assistance easy.
- Q5.5 Using the chatbot with chat assistance would enable me to accomplish my tasks more quickly.
- Q5.6 Using the chatbot with the chat assistance would make it easier for me to carry out my tasks.
- Q5.7 I would find the chatbot with the chat assistance useful.
- Q5.8 Overall, I would find using the chatbot with chat assistance to be advantageous.
- Q5.9 I would use the chatbot with chat assistance for my needs.
- Q5.10 Using the chatbot with chat assistance for handling my needs is something I would do.
- Q5.11 I would see myself using the chatbot with chat assistance for handling my needs.

***Open questions about new chat assistance feature**

- Q6 Do you think this feature would improve the user experience of the chatbot? Why? [text input]
- Q7 How do you think this feature would change the user's behavior while using the chatbot? [text input]

A.3 Evaluation of In-situ Chat Assistance Implementation: Online Study

***Demographics**

- Q1 Age [number field]
- Q2 What is your gender identity? [single choice]
- Female
 - Male
 - [text field]
- Q3 Occupation [text field]
- Q4 Did you participate in the participatory design workshop? [yes/no]
- Q5 I am a conversational agents expert. [7-point Likert scale from strongly disagree to strongly agree]

***Questions after interacting with each chatbot: Task-oriented chatbot for restaurant booking/ social chatbot for sharing emotions (with option-based assistive feature or proposed chat assistance)**

[101-point analog scale from strongly disagree to strongly agree]

- Q6.1 I have high trust in the conversational agent with chat assistive features.
- Q6.2 I could totally rely on the conversational agent with chat assistive features.
- Q6.3 I can strongly engage with the conversational agent with chat assistive features.
- Q6.4 Using the conversational agent with chat assistance is very easy for me.
- Q6.5 My interaction with the conversational agent is totally clear and understandable with the help of the chat assistance.

- Q6.6 It is very easy for me to become skillful in the use of the conversational agent with the chat assistive features.
- Q6.7 Overall, I find the use of the conversational agent with the chat assistance easy.
- Q6.8 Pull the slider all the way to the [attention check-randomized: right/left]
- Q6.9 Using the conversational agent with the chat assistive features would totally enable me to accomplish my interaction goals (reason of interaction with conversational agent) more quickly.
- Q6.10 Using the conversational agent with the chat assistive features would totally make it easier for me to carry out my interaction goals (reason of interaction with conversational agent).
- Q6.11 I would totally find the conversational agent with chat assistance useful.
- Q6.12 Overall, I would find using the conversational agent with chat assistive features to be very advantageous.
- Q6.13 I would totally use the conversational agent with chat assistive features in interactions.
- Q6.14 Using the conversational agent with chat assistive features in interactions is something I would definitely do.
- Q6.15 I would totally see myself using the conversational agents with chat assistance in interactions.
- Q6.16 Pull the slider all the way to the [attention check-randomized: right/left]
- Q6.17 Using the conversational agents with chat assistance is a very good idea.
- Q6.18 I would strongly feel that using the conversational agent with the chat assistive features is pleasant.
- Q6.19 In my opinion, it would be totally desirable to use the conversational agent with chat assistance.
- Q6.20 In my view, using the conversational agent with chat assistance is a very wise idea.
- *Open-ended questions**
- Q7.1 How would you describe your overall experience using the chat assistive features in interactions with the conversational agent? [text input]
- Q7.2 How did the chat assistive features influence your behavior in interactions with the conversational agent? [text input]
- Q7.3 What did you like most about using the conversational agent with the chat assistive features? [text input]
- Q7.4 What did you like least about using the conversational agent with the chat assistive features? [text input]

A.4 Statistics for the Survey on Nine Design Criteria from the Participatory Design Workshop

Table 3: The statistics for the nine design criteria from the design workshop (N=9), see Figure 3. We used Shapiro–Wilk tests for normality testing and Wilcoxon signed-rank tests to compare the results.

	Normality		WSRT	
	W	p	W	p
C1 Contextual info	.636	<.001	1.000	.285
C2 Feedback right after interaction	.842	.006	6.000	.339
C3 Guidance when user hesitates	.855	.010	6.500	.201
C4 User emotion detection	.839	.006	3.000	.035
C5 Embodied agent	.855	.010	5.000	.246
C6 Beyond text input	.677	<.001	1.500	.197
C7 Awareness of website info	.672	<.001	1.500	.104
C8 Function to export chat info	.859	.012	3.000	.062
C9 Access to internet	.687	<.001	<.001	.041

A.5 Statistics for the Online Survey: Evaluation of In-situ Chat Assistance Proposed Design

Table 4: The statistics for the online survey (N=26) , see Figure 4. We used Shapiro–Wilk tests for normality testing and Wilcoxon signed-rank tests (WSRT) to compare the results.

	Normality		WSRT	
	W	p	W	p
PEOU	.938	.009	101.000	.613
PU	.943	.014	41.000	.030
INT	.924	.003	36.500	.057