Conversational Breakdown in a Customer Service Chatbot: Impact of Task Order and Criticality on User Trust and Emotion

ASBJØRN FØLSTAD, SINTEF, Oslo, Norway
EFFIE L.-C. LAW, Durham University, Durham, UK
NENA VAN AS, boost.ai, Sandnes, Norway

While chatbots are increasingly used for customer service, there is a knowledge gap concerning the impact of Conversational Breakdown in such chatbot interactions. In a 2 × 4 factorial design online experiment, we studied how Conversational Breakdown impacts user emotion and trust in a chatbot for customer service, given variations in task criticality and breakdown task order. Here, 257 participants were randomly assigned to complete high- or low-criticality tasks with a prototype chatbot for customer service, experiencing Conversational Breakdown for the first, second, third or none of their tasks. The task set was decided from a 63-participant pre-study. We found significant impact of Conversational Breakdown, including a marked order effect on overall trust, as well as a bounce-back effect on task-specific trust and emotion after subsequent successful task completion. We found no post-interaction effect of Task Criticality. Based on our findings, we discuss theoretical and practical implications and suggest future research.

CCS Concepts: • **Human-centered computing** → **Interaction paradigms**;

Additional Key Words and Phrases: Conversational agent, Conversational Breakdown, Task Criticality, Trust, Emotion

ACM Reference format:
Asbjørn Følstad, Effie L.-C. Law, and Nena van As. 2024. Conversational Breakdown in a Customer Service Chatbot: Impact of Task Order and Criticality on User Trust and Emotion. *ACM Trans. Comput.-Hum. Interact.* 31, 5, Article 66 (November 2024), 52 pages.
https://doi.org/10.1145/3690383

1 Introduction

Chatbots are software agents through which users access services and information through natural language interaction [Følstad et al., 2021]. Due to their promise to complement human customer representatives [Sands et al., 2021] and provide cost-efficient and low-threshold support [Gartner, 2019], customer service is an important chatbot use case [Adam et al., 2021]. The market for conversational **Artificial Intelligence (AI)** is increasing with a large number of available conversational platforms, including open-source platforms such as Rasa, low-threshold alternatives

The study has been supported by the Research Council of Norway, Grant no. 270940. This article has been finalized by support of the Research Council of Norway, Grant no. 346762.
Authors' Contact Information: Asbjørn Følstad (corresponding author), SINTEF, Oslo, Norway; e-mail: asf@sintef.no; Effie L.-C. Law, Durham University, Durham, UK; e-mail: lai-chong.law@durham.ac.uk; Nena van As, boost.ai, Sandnes, Norway; e-mail: nena.van.as@boost.ai.

ACM 1557-7325/2024/11-ART66
https://doi.org/10.1145/3690383

such as BotPress, or enterprise solutions such as Kore.ai and boost.ai [Gartner, 2023]. It has been predicted that as much as 10% of contact centre agent interactions at a global level will be automated by 2026 [Gartner, 2022]. In sectors with high chatbot uptake, such as insurance and consumer banking, about 50% of leading companies provide customer service through chatbots [Taylor et al., 2019]. Customer service is also a major domain for chatbot research. In a review published in 2021, customer service was identified as the most frequently reported application domain for published chatbot research [Rapp et al., 2021].

While **large language models (LLMs)** gradually are taken up in chatbots for customer service [Gartner, 2023], chatbots in this domain are overwhelmingly set up as intent-based solutions [Kvale et al., 2021]. Such intent-based chatbots apply machine learning to identify users' intents from a predefined set, based on the free text input in the user messages [Luo et al., 2022]. When an intent is predicted, the chatbot provides the corresponding response to the user through a rule-based approach. Comprehensive chatbots for customer service may include an intent base of several thousand intents [Zhang et al., 2023]. Intent-based chatbots are optimised to provide quality assured responses, in contrast to chatbots based on LLMs, which are more flexible but prone to 'hallucinations', that is, responses without basis in the training data [Li et al., 2023], or responses not in line with company policy [Wester et al., 2024].

In spite of increasing uptake, substantial challenges remain for successful application of chatbots for customer service. While such chatbots may indeed bring benefits for users, specifically in terms of efficient and accessible support for simple requests [Følstad and Taylor, 2021], concern has been expressed regarding subpar user experiences associated with poorly designed or immature chatbots [van der Goot et al., 2020]. Hence, a substantial proportion of potential users are reluctant to use chatbots for customer service [Statista, 2022]. Specifically, users have been found to be concerned about chatbots misinterpreting users or not providing helpful support, causing frustrating or derailed conversations [Drift, 2018]. User perceptions of chatbots not being able to interpret or act on what they are requested are closely associated with chatbot trust [Nordheim et al., 2019].

Troubles that arise during interaction, which obstruct the user and chatbot from understanding each other, are referred to as Conversational Breakdown. Such breakdown is common for any conversation involving humans [Schegloff, 1987] but represents a particular challenge for interactions between users and chatbots [Ashktorab et al., 2019]. And while the flexibility of chatbot responses clearly will be advanced through the availability of LLMs [Brown et al., 2020], Conversational Breakdown is likely to represent a substantial challenge in human-chatbot interaction also for the foreseeable future due to the inherent issues in LLM-powered chatbots [Li et al., 2023; Wester et al., 2024].

Previous research has shown the potential negative effects of Conversational Breakdown on user experience and investigated different approaches towards recovering from such breakdown [Benner et al., 2021]. There is, however, a lack of knowledge concerning key aspects of the effect of Conversational Breakdown in interactions between users and chatbots. First, while Conversational Breakdown has been shown to impact specific aspects of users experience such as trust [Law et al., 2022], there is a lack of research addressing the interplay between key aspects of user experience, such as trust and emotion, and how this evolves in chatbot interactions where breakdown occurs. Such knowledge is important as it can enable more nuanced responses to breakdown, which in turn may lead to improved user experience. Second, while it is clear that Conversational Breakdown impacts user experience [Følstad and Taylor, 2021], there is a lack of knowledge concerning the effect of the timing of breakdown during conversation. This is a pressing knowledge gap as previous work has hinted at a potential bounce-back effect, where users' negative responses to Conversational Breakdown may be mitigated provided later successful task completion [Law et al., 2022]. Third, while chatbots for customer service typically are employed for **low-criticality (LC)** tasks [Følstad and Skjuve, 2019]], this is likely to change as chatbot capabilities evolve. Hence, it is

important to understand how Task Criticality impacts the user experience of Conversational Breakdown.

In response to this lack of knowledge, we have conducted an experimental study to examine the effect of Conversational Breakdown on user trust and emotion. Specifically, we have examined how user trust and emotion evolve throughout a chatbot conversation which includes a breakdown situation, how Task Criticality impacts user trust and emotion, and how trust and emotion are impacted by the timing of the Conversational Breakdown—that is, whether breakdown occurs relatively early or late in the conversation. The study was conducted as an online experiment following a 2×4 (high/low Task Criticality × Conversational Breakdown at four different timings) between-subjects design. The study involved 257 participants interacting with a prototype chatbot for customer service in the consumer banking and insurance domain. To determine the HC and LC tasks to be included, the online experiment was preceded by a pre-study. Here, six tasks were identified following an assessment of a pool of 60 candidate tasks by 63 participants.

The main contributions of our work are twofold: generating new knowledge on how user trust and emotion evolve during chatbot conversation involving breakdown; developing and implementing new methodological approaches to investigating the phenomena of interest. Specifically, the article:

—Provides insight into the surprisingly limited impact of perceived Task Criticality on the effect of Conversational Breakdown;
—Demonstrates a marked order effect of Conversational Breakdown, where more recent breakdowns are associated with lower overall trust in the chatbot;
—Provides new insight into a bounce-back effect, that is, the observed restoration of task-specific trust and emotion after Conversational Breakdown—provided that a subsequent task is completed successfully;
—Offers insights into the relation between trust and emotion as these evolve during chatbot interaction;
—Demonstrates the process of selecting tasks for different levels of criticality, translating them into scenarios, and evaluating them systematically.

The remainder of the article is structured as follows. First, we present relevant background on user trust and emotion in interaction with chatbots for customer service, Conversational Breakdown in chatbot interaction, and relation between trust, and emotion. We then detail the **research questions (RQs)** and study methods, before presenting the findings. In the discussion section, we clarify the contribution with reference to previous work, outline the implications of our findings, and consider limitations and future research.

2 Background

2.1 User Trust and Emotion in Interaction with Chatbots for Customer Service

For chatbots in customer service, user trust and emotion are important constructs to ensure positive user experiences and increased uptake.

User trust is critical to any new interactive technology [Lankton et al., 2015], and previous research has demonstrated the importance of trust for chatbots in general [e.g., Przegalinska et al., 2019] as well as for customer service [Nordheim et al., 2019]. For customer service, user trust has been found to be determined by factors associated with the chatbot (e.g., chatbot performance), the user (e.g., propensity to trust), and the context (e.g., company brand) [Nordheim et al., 2019].

Of particular interest to this study, user trust in chatbots for customer service has been linked to the criticality of the task at hand. Specifically, Mozafari et al. [2021], found that service criticality

negatively impacted trust in a disclosed chatbot, when comparing this to a non-disclosed chatbot. Based on literature from service research [Ostrom and Iacobucci, 1995], Mozafari et al. [2021] understood criticality as related to the significance or magnitude of consequences of a service delivery. Similarly, Chanseau et al. [2018] defined Task Criticality as the 'importance of a task being carried out safely, correctly and with attention to detail' [ibid., p. 1059]. However, Mozafari et al. [2021] also referred to other aspects of relevance for users' perceptions of criticality, such as task complexity or the need for situational involvement [Webster and Sundaram, 1998].

Users emotion has also been shown to be important for chatbots for customer service. For example, Xu et al. [2017], in a study of Twitter conversations between users and customer service agents, noted that 40% of customer requests were emotional. Crolic et al. [2022], in their studies of chatbot and human service interactions, found evidence of an interaction between user emotion and chatbot design, where users entering a service interaction in an angry state responded negatively to a humanlike chatbot. Sands et al. [2021], in an experimental study of chatbot service interactions, found that application of specific service scripts may be beneficial to user emotion.

It is also noteworthy that recent research has shown a close coupling between user emotion and trust in chatbots for customer service. Wang et al. [2023], in an experiment grounded in cognitive appraisal theory, found that chatbots in customer service may elicit different emotional responses than human service agents with negative consequences for trust, specifically for subjective tasks. Likewise, Rajaobelina et al. [2021], in an experimental study of users' interactions with a chatbot for insurance quotes, found users' perceptions of creepiness in a chatbot to negatively impact both trust and emotion. Furthermore, Lappeman et al. [2023], in an experimental study of user disclosure to a banking chatbot, found trust to be positively associated with willingness to disclose and noted the need to consider emotional aspects of trust in design and management of such chatbots.

Given the importance of user trust and emotion during interaction with chatbots for customer service, and the assumed link between user trust and Task Criticality, it is highly important to investigate these aspects in the context of Conversational Breakdown as this is a context where user trust and emotion may be severely impacted [Law et al., 2022]. To lay ground for such investigations, it is crucial to revisit the empirical background on conversational breakdown in chatbots as well as the theoretical basis for understanding user trust and emotion.

2.2 Conversational Breakdown in Chatbot Interaction

The conversational performance of a chatbot for customer service is critical for how it is perceived by users. A survey of chatbot users identified productivity as a main motivation for chatbot use [Brandtzaeg and Følstad, 2017]. Consequently, a chatbot's ability to comply with the productivity expectation of its users is a substantial determinant of its success [Law et al., 2022]. While user interactions with well-designed chatbots tend to yield relevant responses and helpful dialogue outcomes [Følstad and Taylor, 2021], chatbot conversations may also involve breakdown. Conversational Breakdown in chatbot interaction typically occur if the chatbot makes erroneous predictions of users' intents (false positives) or when it fails to recognise known intents (false negatives), that is, to fail to predict the intent even though the true intention of the user matches an intent in the chatbot model [Følstad and Taylor, 2019].

Conversational Breakdown is, however, a phenomenon not limited to chatbot interaction. As detailed in the field of conversation analysis [Schegloff et al., 1977], breakdown in the form of conversational trouble is to be expected in any conversation involving humans and several repair mechanisms are available. Initiating repair may, for example, be conducted through one of the parties in the conversation expressing uncertainty or expressing their failure to understand. The knowledge base from the field of conversation analysis, including advice on how to initiate and conduct repair, has been taken up in chatbot research and development to guide chatbot interaction

design [Moore et al., 2022]. Here, repair may be provided with the aim of successfully completing a specific conversational topic or task at hand or with the aim of moving the conversation along to a novel topic or task.

When Conversational Breakdown happens, further progress in the conversation depends on the ability of the chatbot and the user to engage in repair, that is, to take the needed conversational actions to get the conversation back on track towards the user's intended goal. As reviewed by Benner et al. [2021], a substantial volume of research has addressed strategies for mitigating breakdown through repair actions [e.g., Ashktorab et al., 2019; Følstad and Taylor, 2019]. There is also an emerging body of research on the high-level implications of Conversational Breakdown [e.g., Law et al., 2022, 2023; Li et al., 2020; Mozafari et al., 2021; Rapp et al., 2021].

Of particular interest to this study, Mozafari et al. [2021] in an experimental study addressing trust and task criticality found breakdown to substantially impact users' trust in a chatbot, and also found evidence for this effect to be impacted by what they referred to as service criticality. Adding to this, Law et al. [2022], in an experimental study of the effect of chatbot human likeness, found Conversational Breakdown to hold potentially different implications for task-specific and overall trust in the chatbot. Their findings suggested the potential relevance of the task order in which Conversational Breakdown occurs (a breakdown task order effect), as well a tendency for task-specific trust to restore after Conversational Breakdown provided the chatbot was able to successfully solve a later task (a bounce-back effect). However, the nuances of how Task Criticality and breakdown order impact the negative effect of Conversational Breakdown on trust and emotion remains unexplored.

2.3 User Trust—Theoretical Basis

To investigate the impact of Conversational Breakdown on user trust and emotion, a thorough understanding of the theoretical basis for trust is required. While multiple definitions and models exist, trust is commonly understood as the willingness of a trustor to '*accept vulnerability based on positive expectations of the intentions or behaviour of the other*' [Rousseau et al., 1998]. In a much-applied model of trust in organisations, Mayer et al. [1995], trust is determined by the trustor's perceptions of key characteristics of the trustee—specifically their ability, benevolence, and integrity—in addition to the trustor's propensity for trust. Depending on the perceived risk implied in the context or task at hand, the resulting trust allows the trustor to engage in risk-taking behaviour involving the trustee.

The trust model of Mayer et al. has been used as a basis for theoretical adaptations in the technology domain [e.g., Lankton et al., 2015; McKnight et al., 2011]. In these adaptations, the concept of trust is typically interpreted as the user's willingness to depend on the technology, i.e., trusting intention [Lankton et al., 2015; McKnight et al., 2011], or their reliance on the technology not to perform actions that go against the benefit of the user [Hancock et al., 2011, 2021]. Trust in technology is seen as determined by the user's perceptions of key characteristics of the technology—such as functionality, reliability, and supportiveness [McKnight et al., 2011]—as well as the users' perceptions of context characteristics such as the task at hand and risk involved, as well as the user's propensity to trust [Hancock et al., 2011, 2021; Lankton et al., 2015; McKnight et al., 2011]. As such, unreliable or unsupportive chatbot behaviour, as experienced during Conversational Breakdown, and also task criticality, as a telltale of the risk involved, are clearly relevant for an investigation of trust in chatbots for customer service.

User trust in technology is continuously calibrated [de Visser et al., 2020]. Wischnewski and colleagues [2023], in their comprehensive review on trust calibrations for automated systems, identified intriguing patterns regarding trust fluctuation and recovery, referred to as trust resilience. Specifically, they found that users are more sensitive to decreases in system reliability than to

increases; the decay in trust is steeper than the increment in trust despite the comparable degree of reliability change in either of the directions, as also noted by Wiegmann et al. [2021]. Users have also been found to trust automated agents more following recovery from minor reliability drops but not so for larger ones [Lu and Sarter, 2019]. Furthermore, the type of error responsible for reliability change could play a key role in trust calibration, for instance, false alarms have been found to dampen trust more than misses [Chen et al., 2021].

Users' trust in technology has typically been measured in terms of their perceptions of the technology's effectiveness for its intended purpose [e.g., Lankton et al., 2015; McKnight et al., 2011; Nordheim et al., 2019]. In addition, measurement instruments for trust determinants such as perceived functionality, helpfulness, and reliability [McKnight et al., 2011] or perceived expertise and efficiency [Nordheim et al., 2019] have been developed. For our purpose, trust measurements addressing users' perceptions of technology effectiveness for intended purpose are particularly relevant. Such measurements have previously been applied by Law et al. [2022, 2023] to investigate users overall trust in chatbots for customer service, as well as their trust in the chatbot for particular tasks. Here, trust has been measured by asking users whether they depend, rely, or count on the chatbot.

2.4 Trust and Emotion

User trust in technology has been analysed more from the cognitive rather than affective processes [Wischnewski et al., 2023], and only a few studies investigated the emotion-trust relation in the context of automated systems [e.g., Fahim et al., 2021].

As a complement to this dominance of a cognitivist perspective, Komiak and Benbasat [2006] presented a complementary model. Here, user trust is conceptualised as including an emotional component, *emotional trust*, in addition to cognitive trust, where emotional trust may mediate cognitive trust in users' decisions to take up a particular technology.

This coupling of trust and emotion is foreseen in theories of emotion. For instance, in Plutchik's [2001] multidimensional model of emotion (i.e., wheels of emotion), trust (or acceptance) together with fear, joy, sadness, anger, and others belong to the vocabulary of the subjective language for describing emotional states. In Hoff and Bashir's [2015] three-layered trust model, emotional states are implicitly mentioned as a personality trait under dispositional trust and as a factor influencing internal validity of situational trust, but they play no role in learned trust.

Two theoretical frameworks are of particular relevance for interpreting user emotions that arise from human-technology interaction: Expectation Confirmation Theory and Cognitive Appraisal Theory. From the cognitive perspective, Expectation Confirmation Theory suggests that people form expectations about the outcomes of a situation, and their emotional response is based on whether those expectations are met or not [Bhattacherjee, 2001]. If the actual outcome matches or exceeds their expectations, people experience positive emotions such as happiness and satisfaction. On the other hand, if the actual outcome falls short of their expectations, people experience negative emotions such as frustration and anger. From the emotional perspective, Cognitive Appraisal Theory posits that our emotional and responses to a situation are determined by how we interpret and evaluate the situation based on our beliefs, goals, and prior experiences [Ellsworth, 2013; Lazarus, 1991; Scherer, 2005; Smith and Ellsworth, 1985]. Positive or negative emotions are elicited when a goal is achieved or denied respectively. The more important (relevant) an event or an object is to an appraiser, the stronger the emotional response to it will be. Cognitive Appraisal Theory has recently been applied in analysis of customer acceptance towards AI-powered services [e.g., Gursoy et al., 2019].

However, Cognitive Appraisal Theory can also lead to contradictory predictions whether certain emotions impact trust. According to the Appraisal Tendency Framework [Lerner and

Keltner, 2001], which is rooted in Cognitive Appraisal Theory, emotion can give rise to an implicit cognitive predisposition to interpret what has happened and to appraise future events. One salient appraisal concerns control. For this *control appraisal* [Myers and Tingley, 2016], the questions are to what extent the 'individual self', 'individual other', or 'exogenous factor' is responsible for a situation and which agent will be in control of future events. As long as the 'individual self' is in control, emotional states like guilt and pride will have no impact on trust in the agent with which one interacts. However, once the 'individual other' is in control, emotional states such as anger and happiness will increase or decrease trust in the agent depending on the valence.

Measurement of emotion is challenging, both theoretically and in practice. A decades-long debate in emotion research concerns whether emotion should be conceptualised and measured as distinct states (categories) or relative points along certain dimensions [e.g., Barrett and Westlin, 2021; Russell and Barrett, 1999]. According to the distinct-state approach, each emotion should be examined as unique [Izard, 1993]. The major issue with the distinct-state approach is that there are obvious overlaps and resemblances across states. The alternative dimensional approach identifies basic dimensions that account for the similarities and differences among emotional states [Osgood, 1962; Russell, 1978]. Here, emotion may be analysed as consisting of three orthogonal dimensions—evaluation (valence or pleasure), potency (dominance or control), and activity (arousal or activation)—though the orthogonality of the dimensions has been debated [e.g., Grgić et al., 2022]. The dimensional approach to measurement of emotion evolved into the Semantic Differential Measures of Emotional States [Russell and Mehrabian, 1977] and **self-assessment manikin (SAM)** [Bradley and Lang, 1994]. SAM is a widely used pictorial tool for measuring emotions in the field of **Human-Computer Interaction (HCI)**, with an underlying assumption that the users are the best source of information on their emotional experiences [Mahlke and Minge, 2008]. Apart from SAM, a plethora of tools for measuring emotional responses have been adapted or developed for HCI purposes, including PANAS [Watson et al., 1988], PrEMO [Desmet, 2018], and User Experience Questionnaire [Laugwitz et al., 2008], to name a few. However, for the specific purposes of our study, where the aim is to measure emotion on a small set of key emotion dimensions, SAM was considered a preferable choice.

3 RQs and Hypotheses

In this section, we present the four key research questions (RQ1–4) and eight hypotheses (H1–H8) of our study. For each RQ, we detail the assumptions guiding the formulation of the questions and related hypotheses.

3.1 Effect of Task Criticality on Trust

Based on theories of trust [Mayer et al., 1995; McKnight et al., 2011] and service criticality [Ostrom and Iacobucci, 1995], Task Criticality is assumed to have substantial implications for users' trust requirements. However, the main effect of Task Criticality has not been demonstrated in larger studies though it has been predicted in the literature [Mozafari et al., 2021; Rossi et al., 2020]. In consequence, the following RQ is explicated:

RQ1: How does Task Criticality impact user trust in a chatbot for customer service?

Previous work has found that when users are made aware that a conversational partner is a chatbot (chatbot disclosure), this has a more negative effect for **high-criticality (HC)** than LC tasks [Mozafari et al., 2021]. Such awareness of the machine character of a conversational partner may also be the result of Conversational Breakdown leading to an effect resembling that observed by Mozafari et al. [2021]. Task Criticality has also been linked to trust implications in social robots

in a small-scale study [Rossi et al., 2020]. Drawing on this previous work, we posit two hypotheses on the impact of Task Criticality:

Hypothesis 1: Task Criticality will impact users' overall trust in a chatbot for customer service, observable as a main effect of Task Criticality.

Hypothesis 2: Task Criticality will impact the effect of Conversational Breakdown on users' overall trust, observable as an interaction effect between Task Criticality and Conversational Breakdown.

3.2 Effect of Conversational Breakdown on Trust

In line with established theories on trust [Lankton et al., 2015; Mayer et al., 1995; McKnight et al., 2011], Conversational Breakdown is assumed to hold substantial implications on users' trust in a chatbot for customer service. In consequence, it is relevant to pose the following RQ:

RQ2: How does Conversational Breakdown impact user trust in a chatbot for customer service?

Previous work has demonstrated marked negative effects of Conversational Breakdown on trust in chatbots for customer service [Law et al., 2022; Mozafari et al., 2021]. This negative effect is foreshadowed in theories on trust as Conversational Breakdown is likely to reduce users' positive expectations of the chatbot's behaviour. In consequence, the following hypothesis is explicated:

Hypothesis 3: Conversational Breakdown will have a significantly negative effect on user's trust, observable as a main effect of Conversational Breakdown.

3.3 Change of User Trust and Emotion in Chatbot Interaction with Conversational Breakdown

While previous work has addressed the impact of Conversational Breakdown on trust [Law et al., 2022; Mozafari et al., 2021], there is a lack of research on how trust dynamically changes during chatbot interactions where Conversational Breakdown occurs. Furthermore, while the importance of emotion to trust have been addressed in previous work [e.g., Komiak and Benbasat, 2006], changes in trust and emotion throughout chatbot interactions have not yet been addressed. We therefore set up the following RQ.

RQ3: How do user trust and emotion change in chatbot interactions with Conversational Breakdown?

Concerning RQ3, previous work [Law et al., 2022] suggests that it may be particularly relevant to investigate (a) an order effect of Conversational Breakdown on overall and task-level trust and emotion and (b) a bounce-back effect on task-level trust and emotion. Motivated by this, we therefore split RQ3 into two sub-questions addressing these complementary aspects, detailed in Sections 3.3.1 and 3.3.2 below.

3.3.1 Order Effect. First, we consider a possible order effect of Conversational Breakdown in the following sub-question:

RQ3a: How does the position of the task where the breakdown occurs determine the extent of trust and emotion changes in users?

The theoretical basis for hypotheses concerning RQ3a may be drawn from theory on emotion as well as theory on trust. According to Expectation Confirmation Theory, if Conversational Breakdown occurs in an initial task, the extent of emotion changes will be smaller than when a breakdown occurs in later tasks. In later tasks, if the user's expectation for the chatbot's performance is set to be positive by successful initial task completion, a breakdown will disconfirm the expectation,

potentially causing disappointment, frustration, or confusion. Arguing along this line, the extent of emotion changes will be relatively larger when breakdown occurs at a relatively late point in the conversation, given the successful completion of previous tasks.

Changes in user trust are expected to follow a similar pattern to that of emotion. The order in which the breakdown occurs is assumed to impact user expectation, where tasks completed without breakdown will increase positive expectations, leading to a more severe breach in expectations when breakdown occurs and a corresponding more severe reduction in trust. The negative effect of breakdown on trust is also assumed to be exacerbated by the negative impact of emotion.

Emotion theory may also motivate hypotheses concerning Task Criticality as a possible moderator of an order effect of Conversational Breakdown. According to Cognitive Appraisal Theory, the extent of emotion changes for the HC tasks should be significantly larger than that for the LC tasks. Users arguably will attach more importance to the former (i.e., anticipating more severe personal consequences) than the latter. In consequence stronger emotional responses may be expected. On this basis, we detail the following hypotheses.

Hypothesis 4 a and b: There will be significant differences in the changes of (a) emotions (valence, activation and control) and (b) trust when Conversational Breakdown occurs at different positions in a set of multiple tasks (IV—task breakdown position).

Hypothesis 5 a and b: There are significant differences in the changes of (a) emotions (valence, activation and control) and (b) trust between the high-criticality and low-criticality tasks depending on task breakdown position (IV—Task Criticality).

3.3.2 Bounce-Back Effect. Second, we consider a possible bounce-back effect of task-specific trust and emotion in the following sub-question:

RQ3b: How does the position of the task where the breakdown occurs determine the level of users' emotion and trust resilience?

The working definition of emotion and trust resilience for our study is the ability of the user to bounce back to an initial, more positive (higher) emotional and trust level, which has been lowered due to negative user experience with Conversational Breakdown. Trust resilience is expected to occur given that users are able to make nuanced trust judgements depending on the task at hand, as suggested both in organisational trust theory [Mayer et al., 1995] and theories of trust in technology [Lankton et al., 2015; McKnight et al., 2011]. Hence, it may be expected that decreases in trust caused by Conversational Breakdown will be offset when a chatbot in subsequent tasks improves on reliability. Potentially, the extent of trust resilience will be higher when breakdown occurs in an early task rather than in later tasks, as the experience of a larger number of non-breakdown tasks following Conversational Breakdown may set the stage for stronger restoration of trust.

Emotion resilience is expected to follow a similar pattern as that of trust resilience. For Conversational Breakdown in an early task, later successful task completions may restore emotion to pre-breakdown levels. Likewise, when breakdown occurs in later tasks, restoration may also be expected but potentially to a lesser degree. In consequence, the following hypotheses are posed.

Hypothesis 6 a and b: There will be significant changes in (a) emotion (valence, activation and control) and (b) trust, due to the Conversational Breakdown of a task, but the respective measures will restore to the pre-breakdown levels after completing subsequently a non-breakdown task.

Hypothesis 7: The extent to which the bounce-back effect is demonstrated will be enhanced proportionally with the number of non-breakdown tasks to be completed after the breakdown one.

3.4 Emotion-Trust Relation

Our final RQ for the study concerns the relation between emotion and trust in chatbot interactions including Conversational Breakdown, and the degree to which this relation may be dependent on experimental conditions such as Task Criticality and breakdown position. This RQ is detailed as follows.

RQ4: Whether and to what extent do emotions mediate users' trust across different experimental conditions?

The existing literature suggests the mediating role of emotion on trust, and that such a relation can be sensitive to contextual attributes. Such a mediating role is particularly foreshadowed in theories of the role of emotion in trust [Komiak and Benbasat, 2006]. Furthermore, theories of emotion, in particular Expectation Confirmation Theory and Cognitive Appraisal Theory, suggest that the emotional impact of Conversational Breakdown may be impacted by contextual aspects such as Task Criticality and breakdown order. In consequence, we posit the following hypothesis.

Hypothesis 8: The extent to which emotions can predict trust will vary significantly with context, operationalised as the different experimental conditions.

4 Methodology

In response to the RQs and hypotheses, our empirical work was set up as an online experiment, following a 2×4 factorial design with Task Criticality and task order for Conversational Breakdown as the **independent variables (IVs)**. Prior to the online experiment, we conducted a pre-study to identify the tasks to be included in the online experiment and define the measurement instrument for Task Criticality.

In this section, we first present the pre-study method and outcome (Section 4.1). Following this, we present the details of the online experiment (Section 4.2). Our empirical work is set in the context of customer service in consumer banking. This context was chosen as consumer banking is a domain where service providers have been early adopters of chatbots for customer service [Taylor et al., 2019].

4.1 Pre-Study

In the pre-study preceding the online experiment, we (a) identified a short list of relevant tasks for the experimental study, (b) established our measurement for Task Criticality and (c) defined the sets of HC and LC tasks to be included in the experimental procedure.

4.1.1 Identification of a Shortlist of Relevant Tasks. In an intent-based chatbot for customer service, relevant tasks are closely associated with available intents. To ensure that the banking tasks selected for the experimental study were realistic, a set of 200 intents was randomly drawn from an original pool of more than 1,700 intents belonging to a chatbot service provider's module for consumer banking. The third author, working for this service provider, was granted access to this 200-intent set by its legal advisor and selected 60 out of the set. This selection was conducted to reduce the number of intents, thereby limiting the resources required for the pre-study and ensuring task relevance for the online experiment. Specifically, the subset was so selected that the tasks reflected in the intents were feasible to be set up for the experiment and credible for participant engagement. After selecting the 60 relevant intents, these were paraphrased and translated into concrete tasks as seen from the user perspective. The 60 resulting tasks are listed in Appendix A.

4.1.2 Establishment of the Task Criticality Measurement Instrument. As a starting point for establishing the Task Criticality measurement instrument, we identified seven potential attributes

Table 1. Task Criticality Attribute and Attribute Descriptions for Rating the Tasks

Attribute	Attribute description
Complex	require highly specialised knowledge or expertise to do appropriately
Difficult	need much effort to carry out
Tedious	are tiresome or boring to do
Specific	the customer service one receives needs to be adapted to one's particular situation
Personalisation	the customer service you receive needs to take into account your detailed characteristics and preferences.
Risk	substantial negative consequence is likely if you do not get adequate help or advice.
Trust	you need to rely on the customer service

of Task Criticality from the literature (Table 1). To assess the relations among these attributes, thereby determining which ones would be most relevant to be included in the measurement instrument, a pilot test was conducted. Here, 20 participants were recruited from a crowdsourcing platform, Prolific, and were asked to rate each of the attributes per task for a set of 15 tasks randomly drawn from the subset of 60 tasks listed above. The tasks were to be rated on seven-point Likert scales on the extent to which each of the attributes were applicable (1: not at all, 7: very much). Hence, each participant made 120 ratings (15 tasks*8 ratings). Each attribute was explained to the participants with the following structure, completed with the attribute names and descriptions of Table 1:

'*Some of the banking tasks for which you can seek customer service are/require/involve <attribute>, that is, <attribute description>*'

In addition, participants were asked to indicate, for each of the tasks, whether they would consider using a chatbot (`useChatbot`). Here, the following definition of chatbot was provided to clarify the meaning of the term: '*Chatbots are automated chat robots that answer questions through a message dialogue.*' This question was posed as the last question to avoid any bias on evaluating the task critically attributes.

Pairwise bivariate correlations (Spearman's rho; Shapiro–Wilk test, $p < .05$) were computed at task level (Table 2). On this basis, the following implications for the attribute relevance to the task selection were drawn:

- `complex`, `difficult`, and `tedious` were highly significantly correlated, suggesting only one of these three be retained for the subsequent study.
- `specific` had a unique correlation pattern; it was significantly correlated with `difficult` and `tedious`, but not with `complex`. It was mildly (non-significantly) correlated with `personalisation`, although semantically they were similar. It implied that participants might interpret this attribute differently from the definition given.
- `personalisation` was highly significantly correlated with `risk` and `trust`, although they were semantically different.
- `risk` and `trust` were highly significantly correlated. If a task is perceived to carry higher risk, more trust is required. While risk could be considered as one of the defining elements of trust, it would be useful to measure both to further analyse their relation.
- `useChatbot` had low, non-significant correlations with `complex`, `difficult` and `tedious`, but had highly negatively significant ones with `personalisation`, `risk` and `trust`. These

Table 2. Bivariate Correlations at Task Level ($N = 15$) among the Seven Task Criticality Attributes and useChatbot

	difficult	tedious	specific	personal	risk	trust	useChatbot
Complex	**0.93****	**0.77****	0.36	−0.39	−0.01	−0.19	0.12
Difficult		**0.79****	**0.58***	−0.15	0.19	−0.04	−0.05
Tedious			**0.54***	−0.27	0.02	−0.15	0.04
Specific				0.31	0.35	0.11	−0.15
Personal					**0.67****	**0.64****	**−0.71****
Risk						**0.94****	**−0.9****
Trust							**−0.91****

$^{**}p < .01$, $^{*}p < .05$.

Table 3. Demographics of Pre-Study Participants

	Gender			Age		Education		Country			Banking task familiarity
	N	M	F	Mean	Range	High School	Higher Ed.	UK	Ireland	USA	
Group 1	20	9	11	30.85	21–43	4	16	14	4	2	3.40
Group 2	21	13	8	33.08	22–54	3	18	20	1	0	3.43
Group 3	22	6	16	34.91	18–62	4	18	17	1	3	3.77
Group 4	20	11	9	32.15	19–56	2	18	14	4	2	3.25
Overall	63	39	44	32.75	18–56	13	70	65	10	7	3.46

intriguing results suggested that participants tended to use chatbots for tasks requiring low levels of personalisation, risk and trust, irrespective of their complexity.

Based on the above analysis, the three Task Criticality attributes retained for the subsequently used measurement instrument were: complexpersonalisation, and risk. In addition, trust and useChatbot were retained to observe their relations to the other attributes, though not included as defining attributes for Task Criticality.

4.1.3 Definition of the Task Set to be Used in the Experiment. To define the task set to be used in the experiment, 63 participants were invited from a crowdsourcing platform, Prolific, and were randomly assigned to four groups. None of the participants had taken part in the pilot test. Participants in each group were asked to rate a set of 15 tasks (Appendix A) against the Task Criticality attributes complex, personalisation and risk, as well as trust and useChatbot, i.e., each participant made 75 ratings. The demographics of each group are shown in Table 3. While the gender distribution was imbalanced in Group 3, the other characteristics, including age, education, country of residence and banking task familiarity were comparable. Countries of residence were limited as only English native speakers were recruited, given the language of the chatbot user interface. Banking task familiarity was explained as the extent to which the participants were familiar with the tasks in the study, rated from 1 (not at all) to 5 (very familiar).

To identify clusters of tasks with comparable Task Criticality scores, we applied two cluster analysis methods—TwoStep and *k*-means—with SPSS v28. TwoStep supports automatic determination of the optimal number of clusters with Schwartz's Bayesian information criterion. *k*-means provides results on the distance to the cluster centre for each case.

Table 4. ANOVA Results of the Two Clusters with Three Measures

	Cluster		Error			
	Mean Square	**df**	**Mean Square**	**df**	**F**	**Sig.**
Z-score (Complex)	28.310	1	0.529	58	53.501	<.001
Z-score (Personal)	29.710	1	0.505	58	58.834	<.001
Z-score (Risk)	23.332	1	0.615	58	37.942	<.001

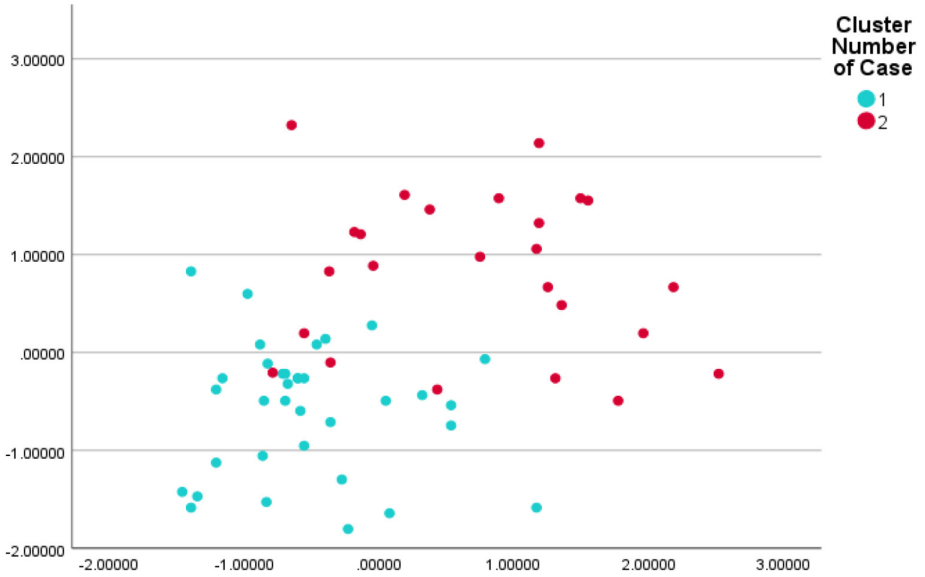

Fig. 1. Two clusters resulting from *k*-means analysis with the axes representing standardised measures.

Based on the data of the Task Criticality attributes—`complex`, `personalisation` and `risk`—TwoStep returned a model of 2 clusters as good fit. Results on cluster sizes (33: 27; the largest to smallest cluster size ratio: 1.22) and cluster memberships were largely consistent with those of *k*-means (35:25; ratio: 1.4), which we will report on subsequently.

Results of an ANOVA indicate that all three attributes contribute significantly to the two clusters (Table 4). Figure 1 illustrates the distribution of the 60 tasks. Cluster 1 ($n = 35$; blue dots) represents LC tasks with lower standardised scores for the three measures whereas Cluster 2 ($n = 25$; red dots) represents HC tasks.

Based on the metric 'distance to centroid' returned by *k*-means for each task (i.e., the lower the value, the higher the representativeness of that task for the cluster), three tasks for each cluster were selected that were closest to the centroid. The range of the distance to centroid metric was 0.30–0.43 for Cluster 1 and 0.53–0.78 for Cluster 2.

The six tasks selected (three HC and three LC) were deployed in the online experiment. For the related task descriptions, see Table 6 (Section 4.2.3).

On a final note from the pre-study, results of bivariate correlation at task level ($N = 60$) indicated a significantly positive correlation between `risk` and `trust` ($r = 0.68$, $p < .001$) and a significantly negative correlation ($p < .001$) between the following pairs: `complex` and `useChatbot` ($r = -0.48$), `personalisation` and `useChatbot` ($r = -0.79$), `risk` and `useChatbot` ($r = -0.77$), `trust`

Table 5. Arrangement of the Tasks in the Experimental Groups

	HC tasks: A, B, C	LC tasks: X, Y, Z
No Breakdown (Bd0)	A1-**B0**-C3	X1-**Y0**-Z3
Breakdown in Task 1 (Bd1)	**B1**-A2-C3	**Y1**-X2-Z3
Breakdown in Task2 (Bd2)	A1-**B2**-C3	X1-**Y2**-Z3
Breakdown in Task3 (Bd3)	A1-C2-**B3**	X1-Z2-**Y3**

Bd, Breakdown.

and `useChatbot` ($r = -0.66$). These findings, along with the cluster analyses, suggest that LC tasks—perceived as low in complexity, need for personalisation, and risk—are also tasks for which users have lower trust requirements and are more willing to use chatbots. The converse is true for HC tasks.

4.2 Online Experiment

4.2.1 Experimental Design. In response to the RQs and hypotheses, the experiment was set up with random assignment to groups in a 4 × 2 (Conversational Breakdown × Task Criticality) factorial design. The participants engaged with a customer service chatbot to resolve three tasks. They could encounter a Conversational Breakdown in one or none of the three tasks.

Prior to interacting with the chatbot, the participants responded to a pre-interaction questionnaire concerning their perception of the tasks they were to use the chatbot for (manipulation check) and to assess their current emotion (`valence`, `activation` and `control`). After each task, the participants responded to a post-task questionnaire concerning `task-specific trust` in the chatbot as well as repeated response on their current emotion. After all tasks were completed, the participants responded to a post-interaction questionnaire concerning `overall trust` in the chatbot. The entire experiment was conducted online.

IVs

The experiment included two IVs:

— *Conversational Breakdown (four conditions).* The participants encountered a Conversational Breakdown in Task 1 (Bd1), Task 2 (Bd2), Task 3 (Bd3) or no breakdown at all (Bd0). No participant encountered more than one breakdown. The breakdowns were designed as part of the conversation, and the participants were not able to recover from these (i.e., no repair was given).

— *Task Criticality (two conditions).* The participants were requested to conduct three tasks. These were either HC or LC, as identified in the pre-study. All participants in the same Task Criticality condition (high or low) completed the same tasks, but the task order varied.

Given the conditions of the two IVs, the experiment involved eight groups as specified in Table 5. This table also shows the ordering of tasks associated with each group. The tasks for the HC and LC conditions are denoted as A, B, C and X, Y, Z, respectively. To signify task order, each task is suffixed with a number indicating the order in which the task is completed. For instance, A1 indicates that Task A is performed as the first task and C2 denotes that Task C is performed as the second task. For the HC condition, only Task B could involve a breakdown, for the LC condition only Task Y. These two tasks, hence, had two states: with no breakdown (B0 and Y0) or with breakdown, which could be in the position 1 (B1, Y1), position 2 (B2, Y2) or position 3 (B3, Y3).

Dependent Variables (DVs)
The experiment included the following DVs:

—`Task-specific trust`. The participants' degree of trust in the chatbot for the specific task just completed.
—`Overall trust`. The participants' degree of trust in the chatbot for customer service in general.
—*Emotion.* The participants' experienced emotion (`valence`, `activation` and `control`) at the time of reporting.
—*Pre-interaction assessment of Task Criticality, trust requirements and willingness to use chatbot.* This assessment was conducted for a manipulation check in the pre-interaction questionnaire, using the Task Criticality measurement from the pre-study (`complex`, `personalisation` and risk), as well as the measurements `trust` and `useChatbot` (Section 4.1).
—*Participant prior chatbot experience.* Measured after chatbot interaction as self-reported `previous chatbot use` and `general satisfaction with chatbots`.
—*Demographic variables.* Measured after chatbot interaction, including `age`, `gender`, `education level` and `country of residence`.

Details on the specific measurement instruments are presented in Section 4.2.3.4.

4.2.2 Participants and Recruitment. Participants were recruited through the Prolific platform which helps crowdsource participants for research studies. As part of their invitation, the participants were informed that the study concerned how chatbots for customer service are experienced and that they would use a chatbot for three banking tasks as part of their participation.

In total, 334 potential participants entered the study. The inclusion criteria were that the participants should have English as their first language and that their participation was through a desktop computer. To avoid confounds due to regional variation in customer service demands, all participants in the experiment were recruited from the UK.

All potential participants' chatbot dialogues were scrutinised. This served as an attention check as well as an opportunity to exclude participants who had experienced unforeseen issues. Following the dialogue checks, 257 participants were included in the final dataset.

Participant incentives were set to 3 GBP. The duration of participation was estimated to be 20 minutes prior to recruitment. The median duration of participation, including participants failing to be included in the final dataset, was 11:33 minutes as measured by Prolific. All participants were awarded the incentive, irrespective of whether their data was included in the final dataset.

4.2.3 Materials and Setup. Participation and data collection was conducted in an online environment set up for the purpose of the experiment. The environment included a study website with participant instructions, a chatbot and a questionnaire service.

Upon the recruitment, the participants were forwarded to the webpage assigned to one of the eight experimental groups (random assignment), each implemented as a separate webpage providing the instructions for participation. The instruction webpage presented the three tasks the participant was to complete as part of the study, provided a link to the pre-interaction questionnaire, and, after the completion of this, provided access to the chatbot to be used. From here, the participant used the chatbot to resolve the assigned tasks. Following each task, the chatbot invited the participant to respond to a post-task questionnaire. After the final post-task questionnaire, the participant answered a final post-interaction questionnaire. The flow of each participant's study interaction is presented in Figure 2.

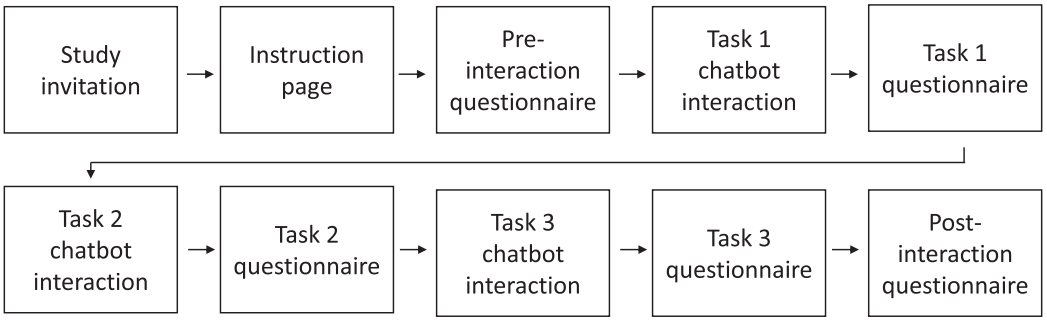

Fig. 2. Flow of each participant's study participation.

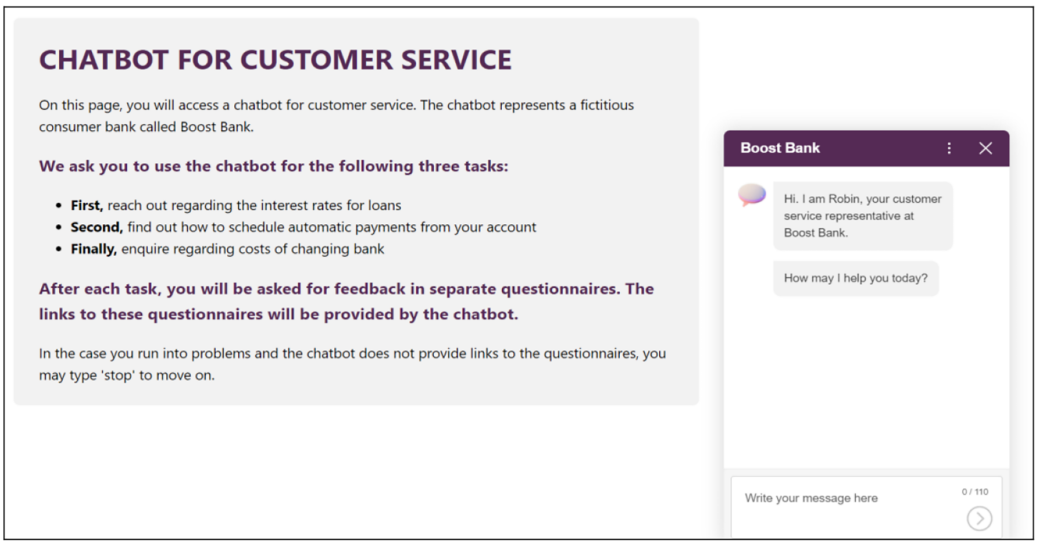

Fig. 3. Screenshot of an example experiment instruction page (HC condition) with the chatbot activated.

4.2.3.1 Chatbot. A customer service chatbot for a fictitious consumer bank was set up. The chatbot was implemented as an intent-based chatbot [Luo, 2022], specifically for this study. The implementation was conducted using *boost.ai,* an enterprise chatbot platform used by clients in the private and public sector with implementation scalability to cover thousands of intents [Gartner, 2023]. The chatbot was represented as a speech bubble avatar and the participants interacted with it through requests in free text and, for some follow-up questions, predefined answer alternatives. The chatbot was set up as a separate screen element overlaying the instruction webpage, aligning with the conversational design style guide of the platform provider. The size of the chatbot screen element was calibrated to allow the participant to see the task descriptions while interacting with the chatbot (Figure 3).

The chatbot style of communication was intended to be pleasant and friendly while keeping up with the professionalism expected from a consumer bank chatbot. The conversational design of the chatbot was developed through iterations within the team of authors, drawing on the experience from conversation design projects with several consumer banks.

The participant's interaction with the chatbot was initiated by a welcome message from the chatbot. The interaction associated with each task was initiated by the participant making the relevant enquiry in their own words. To make the dialogues more extensive than merely providing

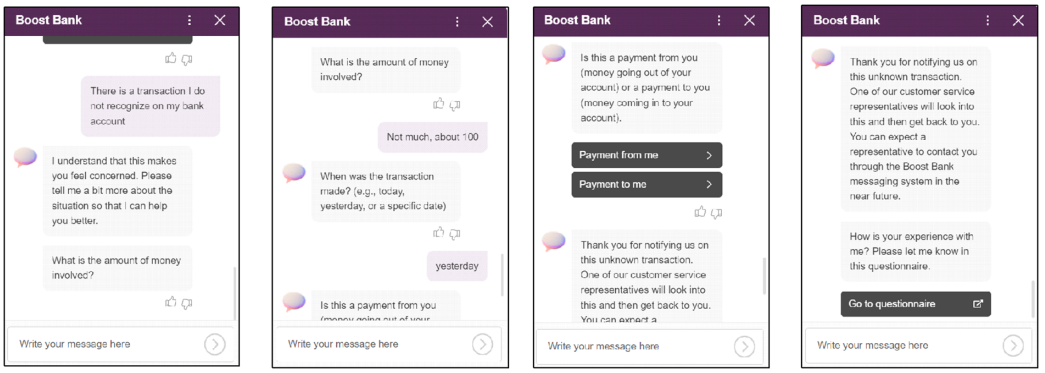

Fig. 4. Example chatbot interaction, Task B0.

Table 6. Overview of the Online Experiment Tasks

Criticality	Task	Task description
High	A	Reach out to the bank because your card was taken by an ATM (cash machine)
	B	Find out what to do if you have found an unknown transaction in your account
	C	Enquire regarding insurance for your car if you want to use it abroad
Low	X	Reach out regarding the interest rates for loans
	Y	Find out how to schedule automatic payments from your account
	Z	Enquire regarding costs of changing bank

a brief response to the participant's request, each task was designed to involve several follow-ups for clarification to give the participant a better ground for assessing the chatbot. As an example of such a task interaction, Figure 4 presents the interaction for Task B0 (HC, no breakdown).

For each participant, the log from the chatbot interaction was saved in the chatbot platform. This log allowed us to verify the completeness of the task interactions.

4.2.3.2 Tasks. For the experiment, three HC and three LC tasks were prepared. The tasks were selected based on the pre-study (Section 4.1). For each task, an intent and a corresponding dialogue flow was established. The intent was predicted from the free text requests of the participant, based on a machine learning model for intent recognition trained as part of the study design. Apart from the initial intent-prediction at the onset of each task, the interactions during the task flows were predetermined. However, the task flows included options for input and choices from the participants. An overview of the six tasks is provided in Table 6.

For examples of the dialogue flows associated with the tasks, the flows for Task B0 and Y0, the non-breakdown versions of the breakdown tasks are presented in Table 7.

4.2.3.3 Conversational Breakdown. In six of the eight experimental groups, the participants encountered a Conversational Breakdown in one of the tasks they were requested to use the chatbot for. During breakdown, the participants were not able to complete the corresponding task. Upon trying to complete the breakdown task for the second time, the chatbot would note its inability to help with the task and ask the participant to continue.

Breakdowns were implemented by providing the participants a fallback message when the intent for the breakdown task was predicted. For example, if a participant in a no-breakdown condition

Table 7. Example Dialogue Flows—Tasks B0 and Y0 (No Breakdown, HC/LC)

Task B0—dialogue flow	Task Y0—dialogue flow
User: [Free text request concerning Task B] **Chatbot:** I understand that this makes you feel concerned. Please tell me a bit more about the situation so that I can help you better. What is the amount of money involved? **User**: [free text, unprocessed] **Chatbot**: When was the transaction made (e.g., today, yesterday or a specific date) **User**: [free text, unprocessed] **Chatbot**: Is this a payment from you (money going out of your account) or a payment to you (money coming in to your account). —Payment from me [button] —Payment to me [button] **User:** [Chooses among answer alternatives] **Chatbot:** [If user chooses 'Payment from me'] Thank you for notifying us on this unknown transaction. One of our customer service representatives will look into this and then get back to you. You can expect a representative to contact you through the Boost Bank messaging service in the near future. [If the user chooses 'Payment to me'] Thank you for getting in touch regarding unknown transactions. Note that it may take some days before all details are available on payments into your account. You will be able to find your updated payment details in your account overview. How is your experience with me? Please let me know in this questionnaire. —Go to questionnaire [button]	**User:** [Free text request concerning Task Y] **Chatbot:** I can help you schedule automatic payments. To get started, I first need to know a little more about the recipient. How have you paid the recipient so far? —Not yet done before (a new payment) [button] —Manual bank payment [button] —Payment through other bank [button] **User:** [Chooses among answer alternatives] **Chatbot**: Could you please tell us the main reason for you to use scheduled automatic payments? —Convenient [button] —Safe [button] —Free service [button] **User:** [Chooses among answer alternatives] **Chatbot**: Perfect. Do you have the recipient's account number and a customer identification number? —I have the needed details —I do not have the needed details **User:** [Chooses among answer alternatives] **Chatbot:** [If the user chooses 'I have the needed details'] Great! Then, on your account webpage, please select "payments" and then "scheduled payments." Here you enter the requested details, and you should be all set. [If the user chooses 'I do not have the needed details'] You find the needed details on your latest invoice from the recipient. Then, on your account webpage, please select "payments" and then "scheduled payments". Here you enter the requested details, and you should be all set. How is your interaction with me? Please report on your experience so far in this questionnaire. — Go to questionnaire [button]

(Bd0) asked the chatbot '*There is a transaction in my bank account I do not recognize*' (initiating Task B), the chatbot would provide the following adequate response: '*Is this a payment from you (money going out of your account) or a payment to you (money coming in to your account) […].*' However, if a participant in a breakdown condition (Bd1, Bd2 and Bd3) asked the same question, the chatbot would instead respond '*I am sorry, I am not sure if I understood you correctly. Did you mean: make a transaction?*' Further follow-ups from the user would lead to new fallback messages and after two attempts of follow-up, the chatbot would state that it cannot help with this task and then request the participant to complete the post-task questionnaire.

To avoid bias concerning task content, the breakdown always occurred in the same tasks (Tasks B and Y), however, the order of the task with the Conversational Breakdown changed according to the experimental condition, as explained in Section 4.2.1. The dialogue flows for the breakdown tasks in the breakdown conditions (B1-3 and Y1-3) are presented in Table 8.

4.2.3.4 Questionnaires and Measurements. The study measurements were taken with five questionnaires, presented at different points in time, as outlined in Figure 2 above. The questionnaires were implemented in the tool *SurveyMonkey*. The measurement scheme is depicted in Figure 5. Subsequently, we detail the specific measurements.

Table 8. Dialogue Flows for Conversational Breakdown in Task B1-3 and Y1-3

Task B1-3—breakdown dialogue flow	Task Y1-3—breakdown dialogue flow
User: [Free text request concerning Task B] **Chatbot:** I am sorry, I am not sure if I understood you correctly. **Did you mean:** make a transaction? **User:** [Free text in response to erroneous response] **Chatbot:** I am sorry that I was not able to understand your question. Please ask again using slightly different words. Remember to keep questions brief and to the point. **User:** [Free text in response to request for repetition] **Chatbot:** I am sorry, but it seems that I am not able to help you with this question. How is your experience with me? Please let me know in this questionnaire. —Go to questionnaire [button]	**User:** [Free text request concerning Task Y] **Chatbot:** I am sorry, I am not sure if I understood you correctly. **Did you mean:** when can I expect payments? **User:** [Free text in response to erroneous response] **Chatbot:** I am sorry that I was not able to understand your question. Please ask again using slightly different words. Remember to keep questions brief and to the point. **User:** [Free text in response to request for repetition] **Chatbot:** I am sorry, but it seems that I am not able to help you with this question. How is your experience with me? Please let me know in this questionnaire. —Go to questionnaire [button]

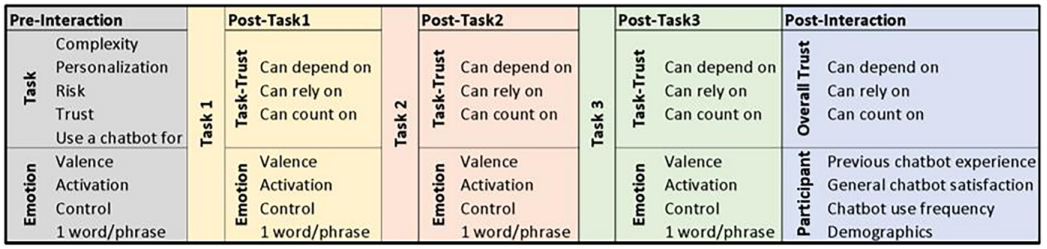

Fig. 5. An overview of the measurement scheme of the online experiment.

Task Criticality

`Task criticality` was measured pre-interaction using the measurement instrument on complexity, personalisation, and risk established in the pre-study (for details, see Table 1). The participants responded to the items with seven-point scales from (1) 'Not at all [...]' to (7) 'Very [...]' where only endpoints were labelled.

Task-Specific and Overall Trust

`Task-specific trust` was measured in each of the three post-task questionnaires, each time using the same three items. The items were based on the Lankton et al. [2015] measurement of trusting intent and concerned whether the participants could depend, rely, and count on the chatbot for the specific task. The participants reported on seven-point Likert scales from (1) 'Strongly disagree' to (7) 'Strongly agree' with only endpoints labelled. `Overall trust` was only measured in the post-interaction questionnaires. The same three items used to measure `task-specific trust` were used also to measure `overall trust` but phrased so as to refer to the chatbot in general. An overview of the questionnaire items for `task-specific trust` and `overall trust` is provided in Table 9.

Table 9. Questionnaire Items for Overall Trust and Task-Specific Trust

Overall trust	Task-specific trust—example from Task B
When in need of customer service, I feel I can depend on the chatbot.	Considering the chatbot's answer regarding [an unknown transaction in my bank account], I feel I can depend on it.
I can always rely on the chatbot to provide good customer service.	I can rely on the support provided by the chatbot regarding [an unknown transaction in my bank account].
I feel I can count on the chatbot for my customer service needs.	I feel I can count on the chatbot for questions regarding [an unknown transaction in my bank account].

Bracketed text in the task-specific trust example is replaced when the respective questionnaire items are used for the other tasks.

Table 10. The Applied Nine-Point Likert Scales for Measuring Emotion

Emotion dimension	Scale
Valence (Pleasure)	It is unpleasant. I am unhappy (1–9) It is pleasant. I am happy.
Activation (Arousal)	I am calm and relaxed (1–9) I am excited and activated.
Control (Dominance)	It is not in my control. I cannot affect it (1–9). It is in my control. I can affect it

Emotion

Participants' emotion in terms of `valence`/pleasure, `activation`/arousal and `control`/dominance, i.e., the SAM scale of Bradley and Lang [1994] was measured at four time points: Pre-interaction (baseline), post-Task 1, post-Task 2 and post-Task 3 (Figure 5).

In line with the recommendations of Bradley and Lang [1994], nine-point Likert scales were used with left and right anchor descriptors (Table 10). Note that the original SAM pictorial representations (curving mouth for pleasure, exploding chest for arousal, body size for dominance] were not used, given that the pictographic format has been criticised for being too sketchy, oversimplified, inaesthetic [Sonderegger et al., 2016], gender-biassed [Sainz-de-Baranda Andujar, 2022] and outdated [Liu et al., 2023]. Nevertheless, the conceptual framework of the underlying model remains viable, and we applied it with an empirically proven verbal scale [Bartosova et al., 2019].

Participants were also asked to describe their current emotion with a keyword or phrase. This mixed (quantitative and qualitative) data approach allowed both the dimensional and categorical measures of emotions to be taken.

Participant Prior Chatbot Experience

The participant's prior experience with chatbots was measured with questionnaire items previously used by Law et al. [2023] and Hobert et al. [2023] (Table 11). The measures applied seven-point Likert scales, from (1) 'Strongly disagree' to (7) 'Strongly agree'; only endpoints labelled. Specifically, we measured participants' previous use of chatbots [Hobert et al., 2023; Law et al., 2023] and their satisfaction in general with chatbots for customer service [Law et al., 2023].

4.2.4 Analysis. Following data capture, data from the five questionnaires (pre-interaction, post-task 1, post-task 2, post-task 3 and post-interaction) were merged by use of a common participant ID across all questionnaires. Upon merging the questionnaires, dialogues for all participants were checked to ensure compliance with the study protocol. About 77 responses with missing tasks were discarded, leaving 257 valid responses. Of these valid responses, 30 were found to have minor

Table 11. Questionnaire Items Used to Measure Previous Use of Chatbots and General Satisfaction with Chatbots for Customer Service

Previous use of chatbots	**General satisfaction with chatbots for customer service**
I frequently use chatbots for customer service.	Chatbots for customer service typically provide good help.
I use chatbots for customer service when this is provided as a service alternative.	In general, chatbots for customer service are an efficient way to get support.
I have used chatbots for customer service for a long time.	I usually find chatbots for customer service pleasant to use.

unexpected events during their dialogues, such as the need to repeat a question to get a relevant response from the chatbot. None of these, however, experienced non-recoverable Conversational Breakdown as was the case in the experimental breakdown conditions.

Analyses of the quantitative data were conducted using SPSS v28. Following an initial descriptive overview, we conducted a manipulation check, based on the pre-interaction questionnaire data. After these introductory analyses, the analyses pertaining to the study hypotheses were conducted as follows:

- *Analysis of effects of Task Criticality and Conversational Breakdown on trust (Hypotheses 1–3)*: The effect of the IVs (Breakdown and Task Criticality) on participants' `overall trust` in the chatbot was investigated through a two-way ANOVA with `overall trust` as DV.
- *Analysis of the order effect of Conversational Breakdown on trust and emotion (Hypotheses 4–5)*: The changes in trust and emotion after breakdown, depending on their position in the conversation, were investigated by one-way ANOVA and one-way ANCOVA, with the task breakdown position as IV.
- *Analysis of the bounce-back effect after breakdown on trust and emotion (Hypotheses 6–7)*: The changes in trust and emotion after successful task resolution following a previous breakdown were investigated with paired-samples t-test.
- *Analysis of the prediction of trust by emotion (Hypothesis 8)*: The extent of prediction was computed with linear regression analysis, regressing each of the three emotion dimensions (`valence`, `activation` and `control`) on the task-specific trust for each experimental condition.

In addition, the analysis of the complementary qualitative data (i.e., one word/phrase description of emotion) was performed using word-cloud visualisation technique [e.g., Heimerl et al. 2014].

4.2.5 Ethics. The ethical aspects of the study were carefully considered prior to the study start-up. Data collection was designed so as to be comfortable for participants, and care was taken to avoid unnecessary data collection or participant activities. Participation was conditioned on the provision of the informed consent where participants were informed on the study's purpose, the implications of their participation, and the anonymity of data collection. The study design and data collection followed the ethical guidelines of the leading research institution, SINTEF, and complied with the relevant external guidelines of the National Committees for Research Ethics in Science and Technology [NENT, 2019] and Social Sciences and Humanities [NESH, 2022].

Table 12. Participant Distribution across the Experiment Groups ($N = 257$)

		Breakdown			
		No breakdown	**First task (Task 1)**	**Second task (Task 2)**	**Third task (Task 3)**
Criticality	High	Group 1 (31) HCBd0	Group 2 (33) HCBd1	Group 3 (30) HCBd2	Group 4 (36) HCBd3
	Low	Group 5 (30) LCBd0	Group 6 (32) LCBd1	Group 7 (32) LCBd2	Group 8 (33) LCBd3

Table 13. Participant Demographics and Prior Experience with Chatbots

Group	***N***	**Gender**			**Age**		**Education**			**Prior chatbot use**		**Prior chatbot satisfaction**	
		M	**F**	**Prefer not to say**	**Mean**	**SD**	**Elem. School**	**High School**	**Higher Ed.**	**Mean**	**SD**	**Mean**	**SD**
1	31	20	11	0	38.2	9.1	0	5	26	4.7	1.1	4.6	1.5
2	33	12	21	0	39.5	12.1	0	5	28	4.4	1.3	4.2	1.4
3	30	12	17	1	37.1	10.4	0	3	27	4.2	1.6	4.0	1.8
4	36	21	14	1	38.1	12.6	0	9	27	3.7	1.6	3.8	1.6
5	30	17	13	0	37.1	13.0	0	4	26	4.7	1.3	4.4	1.5
6	32	17	15	0	37.3	11.9	0	4	28	4.5	1.8	4.3	1.5
7	32	18	14	0	36.4	11.1	1	7	24	4.2	1.3	4.2	1.3
8	33	16	17	0	38.7	13.3	0	7	26	4.0	1.8	4.0	2.0
Overall	257	133	122	2	37.8	11.7	1	44	212	4.3	1.5	4.2	1.6

5 Results

5.1 Descriptive Overview

5.1.1 Participant Demographics. In total, 257 participants provided valid responses. The participants were randomly assigned to the eight experiment groups as outlined in Table 12. An overview of participant demographics, as well as their prior experience with chatbots, is provided in Table 13.

5.1.2 Manipulation Check. The manipulation check was conducted to ascertain whether the HC tasks (A, B and C) were perceived significantly different from the LC tasks (X, Y and Z). As the data were not normally distributed per group level, as shown by the results of Shapiro–Wilk tests ($p < .05$), we applied non-parametric Mann-Whitney tests to compare the two groups of participants; those randomly assigned to work on A, B, and C ($N = 130$) and those on X, Y and Z ($N = 127$) on the Task Criticality measurement instrument (complexity, required personalisation and risk) as well as task trust requirements and likelihood to use a chatbot for the task. All variables showed highly significant differences, as shown by the respective Z values ($p < .001$) (Figure 6). HC tasks were perceived as significantly more complex, more in need of personalisation and more risky—and also having higher trust requirements and being less likely to be done with a chatbot.

5.2 Effects of Task Criticality and Breakdown on Overall Trust (Hypotheses 1–3)

`Overall trust` was measured by three items. The items were found to have acceptable inter-item reliability (Cronbach alpha $\alpha = .94$). At the group level, Shapiro–Wilk tests indicated normal distribution of data ($p > .05$) for all groups but one. A descriptive overview of `overall trust` scores for the eight groups is provided in Table 14.

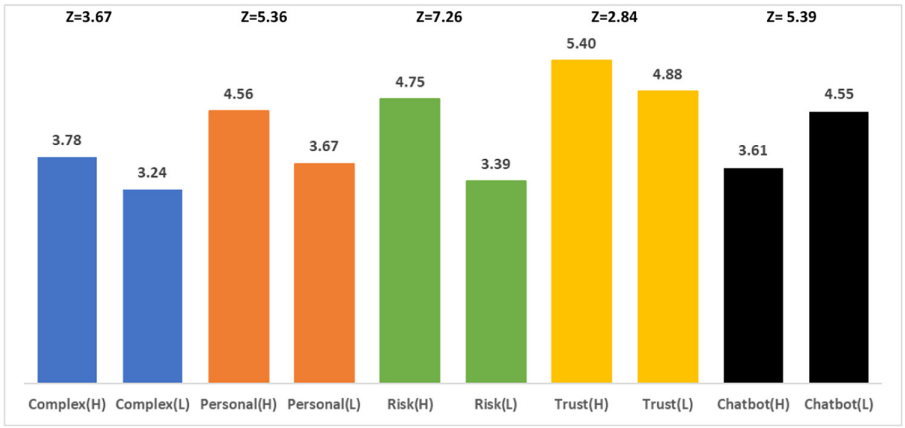

Fig. 6. Mann-Whitney tests between participants of HC and LC groups on the averages of the five pre-interaction variables.

Table 14. Mean (SD) Scores for Overall Trust across the Eight Experiment Groups

		Conversational Breakdown							
		No breakdown		Task 1		Task 2		Task 3	
Criticality	High	5.1	(1.4)	4.3	(1.3)	3.8	(1.4)	3.3	(1.4)
	Low	5.3	(1.2)	4.4	(1.4)	3.9	(1.1)	3.7	(1.3)

The effects of Task Criticality and Conversational Breakdown on `overall trust` were investigated in a two-way ANOVA. A significant main effect was identified for Conversational Breakdown ($F(3,249) = 20.9$, $p < .001$), but not for Task Criticality ($F(1,249) = 1.6$, $p = .21$). No significant interaction effect was found ($F(3,249) = .7$, $p = .04$), as shown in Figure 7. Hence, hypotheses 1 and 2 were rejected, whereas Hypothesis 3 was accepted.

The effect of Task Criticality and Conversational Breakdown on `overall trust` also suggested an impact of the order of breakdown, in line with our assumption in Hypothesis 5b. While our main investigation of an order effect of Conversational Breakdown was conducted by way of the DVs of `task-specific trust` and emotion (detailed in Section 5.3 below), it is a noteworthy finding that a marked order effect is identified for our overall trust measure. Specifically, `overall trust` was reduced for all groups encountering Conversational Breakdown compared to the no-breakdown conditions, and the impact of breakdown on `overall trust` was found to be more severe when breakdown happened relatively late in the dialogue. This was demonstrated in a Tukey HSD pairwise comparison of the four Conversational Breakdown conditions, the results of which are shown in Figure 8.

5.3 How User Trust and Emotion Change in Chatbot Interactions with Conversational Breakdown (Hypotheses 4–7)

In this section, we first present the descriptive statistics of `task-specific trust` and the three emotion measures, followed by results of inferential statistics for the hypotheses concerning a breakdown order effect (Hypotheses 4–5), and a bounce-back effect for `task-specific trust` and emotion (Hypotheses 6–7).

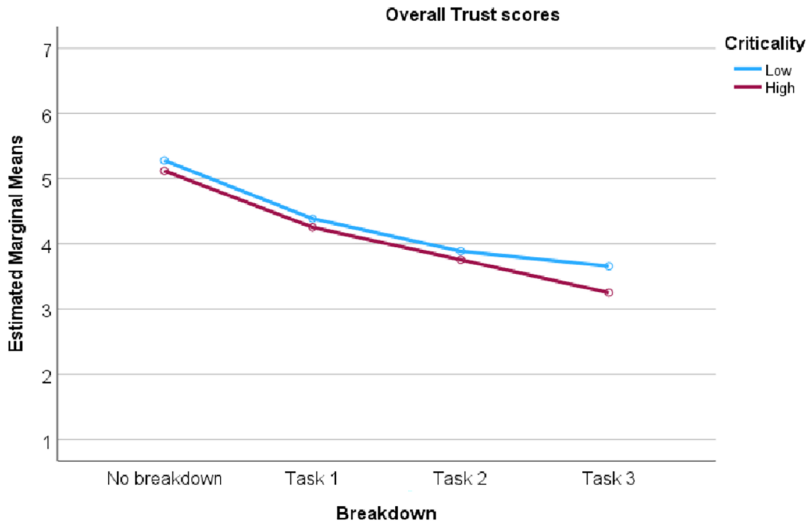

Fig. 7. Effect of Task Criticality and Conversational Breakdown on overall trust.

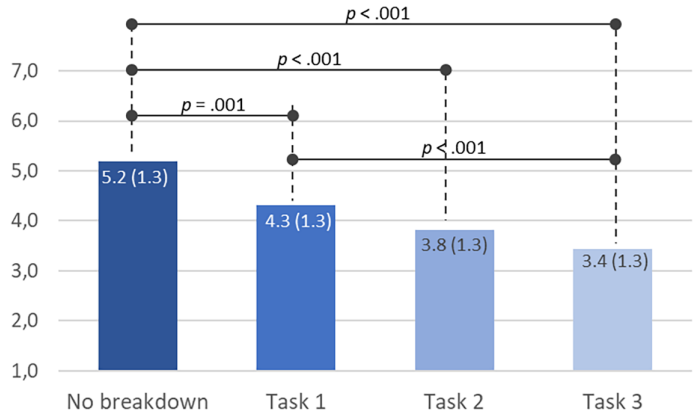

Fig. 8. Results from Tukey HSD pairwise comparisons of overall trust for different Conversational Breakdown conditions. Means and SDs are provided for each condition. Only significant differences are indicated.

5.3.1 Task-Specific Trust—Overview. `Task-specific trust` was measured after a participant's completion of each of the three tasks. The inter-item reliability of the three items constituting the `task-specific trust` measurement was acceptable with Cronbach alpha α (<0.94) for each of the three tasks. However, the `task-specific trust` measure was not found to be normally distributed for most of the experimental groups, given the results of Shapiro–Wilks tests ($p < .05$).

While `task-specific trust` was significantly reduced for tasks with Conversational Breakdown (see Figure 9), there was no evidence of an overall change in `task-specific trust` scores over the three tasks (Table 15—the means across all groups were 4.5, 4.4 and 4.2 for Task 1, 2 and 3, respectively). Note that Task 1, 2 and 3 are different for different groups, for instance, Task 1 was Task A for HCBd0 and Task B for HCBd1, given the order effect investigated.

Table 15. Mean (*M*) and SD of Task-Specific Trust for Each Post Task Measurement, by Group and Across Conditions ($N = 257$)

Group	Post Task 1		Post Task 2		Post Task 3	
	M	**SD**	***M***	**SD**	***M***	**SD**
1 (HCBd0)	5.9	1.3	4.5	1.8	5.1	1.4
2 (HCBd1)	1.1	0.4	6.1	0.8	4.3	1.3
3 (HCBd2)	5.8	1.0	1.0	0.2	3.8	1.4
4 (HCBd3)	5.8	1.4	5.3	1.8	3.3	1.4
5 (LCBd0)	5.4	0.9	5.7	1.3	5.3	1.2
6 (LCBd1)	1.0	0.2	5.8	1.3	4.4	1.4
7 (LCBd2)	5.5	1.2	1.1	0.6	3.9	1.1
8 (LCBd3)	5.3	1.5	5.7	1.4	3.7	1.3
Across all groups	4.5	2.3	4.4	2.3	4.2	1.5

Bd, Breakdown; HC, High Criticality; LC, Low Criticality.

Table 16. Mean (*M*), SD of the Three Measures of Emotion over the Four Time Points over All Conditions ($N = 257$)

	Pre-Interaction (baseline)			**Post-Task 1**			**Post-Task 2**			**Post-Task 3**		
	M	SD	ITC[a]	*M*	SD	ITC[a]	*M*	SD	ITC[a]	*M*	SD	ITC[a]
Valence	6.61	1.67	0.24	6.26	1.95	0.43	6.17	1.95	0.37	6.00	2.14	0.50
Activation	4.11	2.24	0.14	4.32	2.21	0.03	4.39	2.09	0.12	4.40	2.16	0.09
Control	6.49	1.81	0.19	6.07	1.99	0.40	5.99	1.97	0.43	5.91	2.16	0.47
Cronbach's α	0.13			0.4			0.33			0.42		

[a]ITC, Corrected item-total correlation.

5.3.2 Emotion Scores—Overview. Emotion was measured through three single-item scales (`valence`, `activation` and `control`). The low inter-item reliability, as indicated by low Cronbach alpha ($\alpha < 0.4$, see Table 16), indicated that the items of the three scales were not highly correlated with each other and should be analysed independently.

Table 16 displays the mean values for `valence`, `activation` and `control` over all conditions. No substantial changes can be observed. `Valence` and `control` had the highest baseline and decreased over tasks whereas `activation` showed the opposite trend.

Figure 9 also shows the changes in the emotional measures. Although the changes for `valence` and `control` are smaller than that of task-specific trust, they follow a similar pattern: when breakdown occurs, both `valence` and `control` show significant negative change. Interestingly, `activation` shows no significant change regardless of whether and when breakdown occurs—with the exception of the HCBd3 condition: here, there was a significant increase in activation when breakdown occurred.

5.3.3 Order Effect on Trust and Emotion. The order effect of breakdown task on `task-specific trust` and emotion (`valence`, `activation and control`) was investigated by one-way ANOVAs with the Conversational Breakdown position being the IV. Analyses were performed separately for

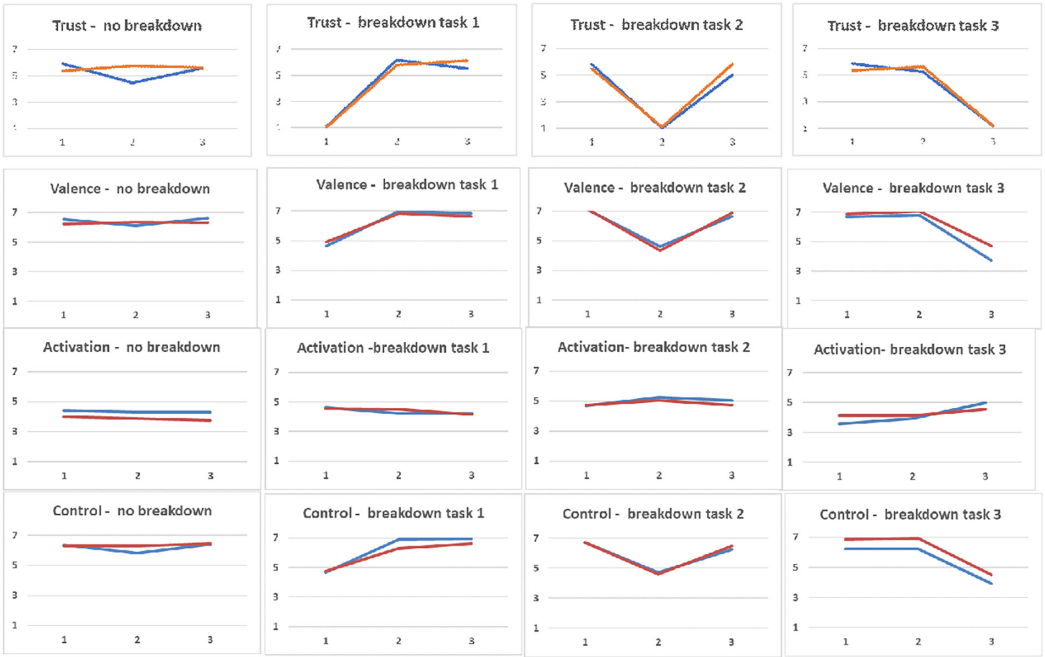

Fig. 9. Changes in ratings (y-axis) of task-specific trust (top), valence (second), activation (third) and control (bottom) with the three tasks (x-axis). Blue lines: HC tasks; Orange lines: LC tasks.

Table 17. Results of ANOVA on the Three Post-Task Measures of Trust and Emotion among the Three Groups with Breakdown at Different Positions, Separately for HC and LC Tasks

	Task-specific trust	Valence	Activation	Control
HC	$F(2,96) = 1.497$ $p > .05$ (= .229)	$F(2,96) = 1.81$ $p > .05$ (= .169)	$F(2,96) = 1.221$ $p > .05$ (= .299)	$F(2,96) = 1.377$ $p > .05$ (= .257)
LC	$F(2,94) = 0.792$ $p > .05$ (= .456)	$F(2,94) = 0.600$ $p > .05$ (= .551)	$F(2,94) = 0.853$ $p > .05$ (= .429)	$F(2,94)$—0.092 $p > .05$ (= .912)

HC and LC tasks. The analysis of the order effect on emotion was complemented with one-way ANCOVAs, with the baseline measures of emotion as covariates.

Results indicated that there were no significant differences among the experimental conditions in emotion measures (Tables 17 and 18). Furthermore, there were no significant differences among conditions in `task-specific trust`. The implication was that the extent of emotion changes was not determined by when the Conversational Breakdown was experienced, be it in the first, second or third task. In short, there was no order effect. Hence, hypotheses 4a and 4b were rejected.

Furthermore, we evaluated Hypothesis 5a and 5b to investigate the effect of the Task Criticality on task-specific emotion and trust by performing pairwise comparisons. To do so, we computed the change in the three emotion measures (`valence`, `activation` and `control`) from baseline to post-Task 1 for both the HC and LC groups, and compared them to each other. For example, that would be Group 2 (HCBd1) versus Group 6 (LCBd1), both of which had breakdown at Task 1 (see Table 12 for grouping). The resulting differences were t-tested (Shapiro–Wilk test, $p > 0.5$). The same procedure was applied to the differences between Group 3 and 7 (HCBd2 and LCBd2), and

Table 18. Results of ANCOVA on the Three Emotion Measures with Baseline as Covariate among the Three Groups with Breakdown at Different Positions, Separately for HC and LC Tasks

	Valence	Activation	Control
HC	F(2,95) = 1.149 p > .05 (= .321)	F(2,95) = 1.102 p > .05 (= .336)	F(2,96) = 1.243 p > .05 (= .293)
LC	F(2,93) = 1.077 p > .05 (= .345)	F(2,93) = 0.680 p > .05 (= .509)	F(2,93)—0.144 p > .05 (= .866)

Table 19. Results of *t*-Tests on the Three Emotion Measures Adjusted with Baseline and Task-Specific Trust for Pairwise Comparison between the High and Low Task Criticality Group, with the Breakdown Task (Bd) at the Same Position

	Valence			Activation			Control			Trust		
	Grp2–6 Bd Task1	Grp3–7 Bd Task2	Grp4–8 Bd Task3	Grp2–6 Bd Task1	Grp3–7 Bd Task2	Grp4–8 Bd Task3	Grp2–6 Bd Task1	Grp3–7 Bd Task2	Grp4–8 Bd Task3	Grp2–6 Bd Task1	Grp3–7 Bd Task2	Grp4–8 Bd Task3
t/Z	−0,99	0,31	−0,94	0,65	0,43	1,48	0,11	−0,46	0,8	−0,45	−0,96	−0,44
n	63	60	67	63	60	67	63	60	67	63	60	67
p	0,32	0,76	0,35	0,52	0,67	0,07	0,92	0,65	0,43	0,65	0,34	0,66

Group 4 and 8 (HCBd3 and LCBd3). As no baseline measure was involved for task-specific trust, the averages of the three items were used for Mann-Whitney tests (Shapiro–Wilk test, $p < .05$). Results (Table 19) indicate that none of the comparisons were significant for any of the three emotion measures or task-specific trust, suggesting that participants' emotion and trust changes were not sensitive to the Task Criticality.

5.3.4 Bounce-Back Effect on Trust and Emotion. To verify the bounce-back effect on trust and emotion (Hypotheses 6–7), we analysed the changes in the `task-specific trust` and three dimensions of emotion (`valence`, `activation and control`) across the eight experimental conditions. Specifically, we computed pairwise differences between the two consecutive post-task measures to find out to what extent the trust and emotion experienced in the preceding task would be masked by that in the following one.

We evaluated the significance of the differences (Diff) in the following pairs: Task 1-baseline (for emotion data only as there is no baseline for trust), Task 2-1 and Task 3-2. In addition, for emotion data, we computed Task 3-baseline to check whether the emotion measures would bounce back to the initial baseline level. Paired-samples *t*-tests were applied to the emotion data (Shapiro–Wilk test, $p > .05$) whereas Wilcoxon Signed Rank tests were used for the trust data (Shapiro–Wilk test, $p < .05$). Results are presented in Table 20 (`task-specific trust`), Table 21 (`valence`), Table 22 (`activation`) and Table 23 (`control`). Notes for reading Tables 20–23: high/low = Task Criticality; bd0 = no breakdown; bd1 = breakdown at Task 1; bd2 = breakdown at Task 2 and bd3 = breakdown at Task 3. Task 1/2/3 refers to the order that the task is done, but not the exact task, which is indicated by the respective column heading.

Below, we structure the results along the four categories of Conversational Breakdown: No breakdown, breakdown in Task 1, breakdown in Task 2 and breakdown in Task 3. In addition, we report the order of the tasks for each of those conditions in brackets.

Table 20. Analysis of Bounce-Back Effect in Task-Specific Trust over the Three Tasks Across the Experimental Conditions

	Task1	**Task2**	**Task3**	**Diff-Task2-1**	**Diff-Task3-2**
	Trust_A1/X1	Trust_B0/Y0	Trust_C3/Z3		
high-bd0	5.90	4.46	5.58	$Z = -3.67$, $p < .001$	$Z = -3.33$, $p < .001$
low-bd0	5.37	5.74	5.61	n.s.	n.s.
	Trust_B1/Y1	Trust_A2/X2	Trust_C3/Z3		
high-bd1	1.10	6.14	5.53	$Z = -5.06$, $p < .001$	$Z = -2.83$, $p = .005$
low-bd1	1.04	5.78	6.13	$Z = -4.97$, $p < .001$	n.s.
	Trust_A1/X1	Trust_B2/Y2	Trust_C3/Z3		
high-bd2	5.83	1.03	5.02	$Z = 4.81$, $p < .001$	$Z = -4.72$, $p < .001$
low-bd2	5.49	1.14	5.85	$Z = -4.95$, $p < .001$	$Z = -4.96$, $p < .001$
	Trust_A1/X1	Trust_C2/Z2	Trust_B3/Y3		
high-bd3	5.84	5.27	1.18	$Z = -2.68$, $p = .007$	$Z = -5.08$, $p < .001$
low-bd3	5.34	5.67	1.19	$Z = -1.56$, $p = .119$	$Z = -5.03$, $p < .001$

5.3.4.1 *Patterns in Task-Specific Trust.*

— *No breakdown (A1-B0-C3/X1-Y0-Z3).* For the HC tasks, there was a significant drop in `task-specific trust` from Task A1 to B0, even though there was no Conversational Breakdown. It might be down to the nature of Task B (i.e., unknown transaction in the account), which was perceived higher in `complexity`, `personalisation` and `risk` than Task A and C (Appendix B). `Task-specific trust` significantly recovered after completing Task C3. In contrast, `task-specific trust` changed to a limited extent only for the three LC tasks.

— *Breakdown in Task 1 (B1-A2-C3/Y1-X2-Z3).* Both HC and LC groups had a very low level of `task-specific trust` for the breakdown task (B1/Y1) and a significant increase after the completion of non-breakdown Task 2 (A2/X2). Interestingly, for the HC group, there was a significant drop in `task-specific trust` from Task A2 to Task C3, whereas there was a (non-significant) increase for the LC group (X2 to Z3). The observations may be explained in terms of the pre-interaction ratings (Appendix B), showing that Task A and C were perceived to be significantly different whereas Task X and Z were similar.

— *Breakdown in Task 2 (A1-B2-C3/X1-Y2-Z3).* Both criticality groups showed the expected trend, with a drop of `task-specific trust` for the breakdown tasks (B2/Y2) and a significant increase after the successful completion of subsequent Tasks C3/Z3.

— *Breakdown in Task 3 (A1-C2-B3/X1-Z2-Y3).* The HC group showed a significant drop in trust from Task A1 to Task C2, though both tasks had no breakdown; it dropped further for B3 with breakdown. A similar drop in trust from Task A to C was also observed for the other group that had Task C following Task A (B1, HC). This might be due to the nature of the tasks, which were perceived differently even prior to the interaction (cf. Appendix B).

Overall, the results demonstrate the bounce-back effect in `task-specific trust`.

5.3.4.2 *Patterns in Valence.*

— *No breakdown (A1-B0-C3/X1-Y0-Z3).* For HC tasks, none of the differences were significant except the change from B0 to C3 where the `valence` was increased after a dip. While B0 had no breakdown, the nature of this task might cause a degree of unpleasantness in participants. Nonetheless, the drop was non-significant (n.s.). The completion of the subsequent C3 led

Table 21. Analysis of Bounce-Back Effect in Valence (Val) over the Three Tasks across the Experimental Conditions

	Baseline	Task1	Task2	Task3	Diff-Task1-Base	Diff-Task2-1	Diff-Task3-2	Diff-Task3-Base
	Val_Base	Val_A1/X1	Val_B0/Y0	Val_C3/Z3				
high-bd0	6.35	6.55	6.13	6.61	n.s.	n.s.	$t(30) = 2.5$, $p = .02$	n.s.
low-bd0	6.13	6.23	6.37	6.30	n.s.	n.s.	n.s.	n.s.
	Val_Base	Val_B1/Y1	Val_A2/X2	Val_C3/Z3				
high-bd1	6.88	4.64	7.00	6.85	$t(32) = 5.8^*$	$t(32) = 5.6^*$	n.s.	n.s.
low-bd1	6.66	4.91	6.84	6.66	$t(31) = 5.7^*$	$t(31) = 5.6^*$	n.s.	n.s.
	Val_Base	Val_A1/X1	Val_B2/Y2	Val_C3/Z3				
high-bd2	6.97	7.10	4.60	6.63	n.s.	$t(29) = 5.9^*$	$t(29) = 4.9^*$	n.s.
low-bd2	6.88	7.13	4.34	6.88	n.s.	$t(31) = 7.0^*$	$t(31) = 6.6^*$	n.s.
	Val_Base	Val_A1/X1	Val_C2/Z2	Val_B3/Y3				
high-bd3	6.33	6.69	6.81	3.75	n.s.	n.s.	$t(35) = 6.9^*$	$t(35) = 5.7^*$
low-bd3	6.73	6.88	7.06	4.73	n.s.	n.s.	$t(32) = 5.6^*$	$t(32) = 4.8^*$

$^*p > .001$.

to a significant increase; the task might elicit in participants positive emotion associated with travelling abroad. The C3 rating was higher than the baseline, albeit non-significant. In contrast, for the LC tasks, Y0 increased as compared with Task X1 and decreased in Task Z3. These differences were small and non-significant.

— *Breakdown in Task 1 (B1-A2-C3/Y1-X2-Z3).* B1 with breakdown was completed as the first task, resulting in a notable drop from the baseline, but the rating increased substantially after completing A2. Both changes were statistically significant. The C3 rating was comparable to the baseline. The same pattern was observed for LC tasks: Y1 followed by X2 and Z3. This case demonstrated the bounce-back effect of emotion.

— *Breakdown in Task 2 (A1-B2-C3/X1-Y2-Z3).* B2 with breakdown was completed as the second task, resulting in a notable drop in `valence` as compared with Task A1, and the completion of the subsequent C3 led to a substantial increase. Both changes were statistically significant. While the C3 rating was lower than the baseline, the difference was non-significant. The same pattern was observed for Y2, X1 and Z3. This case demonstrated the bounce-back effect of emotion.

— *Breakdown in Task 3 (A1-C2-B3/X1-Z2-Y3).* B3 with breakdown was completed as the third task, causing a significant drop in `valence` as compared with the C2 and baseline `valence`. Nonetheless, this case cannot be used to demonstrate the bounce-back effect because no other task was done after Task 3.

5.3.4.3 Patterns in Activation. The results in `activation` were perplexing. Almost no notable changes were observed across the four points of measure, irrespective of the Task Criticality or the breakdown condition. While there was an increase for B1/Y1 and B2/Y2, followed by a drop in the subsequent A2/X2 and C3/Z3, respectively, the changes were non-significant.

One exception was Task B3 in the HC condition. There was a notable change in `activation`, which was significantly higher than Task C2 and the baseline. Nonetheless, this case cannot be used to show the bounce-back effect as B3 happened as the last task and there was no subsequent measurement of `activation` to compare to. Overall, given the flat pattern of `activation` ratings, i.e., participants were activated to a comparable level throughout the process, no bounce-effect cannot be concluded.

5.3.4.4 Patterns in Control. The patterns of changes in `control` are the same as those we observed in `valence`. Hence, we argue that the bounce-back effect of `control` can also be demonstrated.

Table 22. Analysis of Bounce-Back Effect in Activation (Act) over the Three Tasks across the Experimental Conditions

	Baseline	Task1	Task2	Task3	Diff-Task1-Base	Diff-Task2-1	Diff-Task3-2	Diff-Task3-Base
	Act_Base	Act_A1/X1	Act_B0/Y0	Act_C3/Z3				
high-bd0	4.42	4.42	4.29	4.29	n.s.	n.s.	n.s.	n.s.
low-bd0	3.83	3.97	3.87	3.73	n.s.	n.s.	n.s.	n.s.
	Act_Base	Act_B1/Y1	Act_A2/X2	Act_C3/Z3				
high-bd1	3.91	4.61	4.24	4.21	n.s.	n.s.	n.s.	n.s.
low-bd1	4.16	4.53	4.50	4.16	n.s.	n.s.	n.s.	n.s.
	Act_Base	Act_A1/X1	Act_B2/Y2	Act_C3/Z3				
high-bd2	4.4	4.70	5.23	5.07	n.s.	n.s.	n.s.	n.s.
low-bd2	4.47	4.75	5.03	4.75	n.s.	n.s.	n.s.	n.s.
	Act_Base	Act_A1/X1	Act_C2/Z2	Act_B3/Y3				
high-bd3	3.61	3.56	3.92	5.00	n.s.	n.s.	$t(35) = 2.4, p < .024$	$t(35) = 2.8, p = .009$
low-bd3	4.18	4.12	4.12	4.55	n.s.	n.s.	n.s.	n.s.

Table 23. Analysis of Bounce-Back Effect in Control (Con) over the Three Tasks across the Experimental Conditions

	Baseline	Task1	Task2	Task3	Diff-Task1-Base	Diff-Task2-1	Diff-Task3-2	Diff-Task3-Base
	Con_Base	Con_A1/X1	Con_B0/Y0	Con_C3/Z3				
high-bd0	6.48	6.39	5.81	6.42	n.s.	$t(30) = 2.3, p < .029$	$t(30) = 3.6, p = .001$	n.s.
low-bd0	6.43	6.30	6.30	6.47	n.s.	n.s.	n.s.	n.s.
	Con_Base	Con_B1/Y1	Con_A2/X2	Con_C3/Z3				
high-bd1	6.67	4.67	6.88	6.94	$t(32) = 5.3^*$	$t(32) = 5.6^*$	n.s.	n.s.
low-bd1	6.81	4.75	6.28	6.63	$t(31) = 4.6^*$	$t(31) = 3.7^*$	n.s.	n.s.
	Con_Base	Con_A1/X1	Con_B2/Y2	Con_C3/Z3				
high-bd2	6.57	6.70	4.70	6.23	n.s.	$t(29) = 5.4^*$	$t(29) = 4.9^*$	n.s.
low-bd2	6.16	6.72	4.59	6.47	n.s.	$t(31) = 4.9^*$	$t(31) = 4.0^*$	n.s.
	Con_Base	Con_A1/X1	Con_C2/Z2	Con_B3/Y3				
high-bd3	5.86	6.22	6.25	3.94	n.s.	n.s.	$t(35) = 4.9^*$	$t(35) = 4.1^*$
low-bd3	6.97	6.88	6.94	4.52	n.s.	n.s.	$t(32) = 5.1^*$	$t(32) = 5.0^*$

$^*p < .001$.

In other words, participants restored the sense of control over the situation when the chatbot performed the requested task successfully.

Nonetheless, for the no breakdown condition, the drop in `control` for Task B0 was statistically significant, which could be due to the nature of B0, as explained in the case of `valence`. The subsequent increase for Task C3 was also statistically significant and restored close to the baseline.

5.3.5 Verbal Expressions of Emotion. In addition to rating the scales of `valence`, `activation` and `control`, participants were asked to give one word or phrase to describe their emotion. The majority responded with one word whereas a handful responded with a phrase or sentence. To analyse these verbal data, a simple word count method was used. Phrases such as '*happy and satisfied*' are counted as two instances of emotion: '*happy*' and '*satisfied*'. Key emotional words are extracted from sentences such as '*Feeling normal, not tired, not distracted, ready to chat with the bot*' and '*Neutral is probably the best way to describe it*', being labelled as 'normal' and 'neutral', respectively. Results are presented with the simple word cloud visualisation technique.

5.3.5.1 Order Effect. As described in Section 5.3.3, the ratings for `valence` and `control` visibly decreased (i.e., less pleasant, less in control) after experiencing Conversational Breakdown to a similar extent, irrespective of the task position. We examined whether emotional words expressed were also comparable, using the word-cloud technique to visualise the relative weight or frequency of individual words. Figure 10 with the data table underneath clearly shows that '*frustrated*' and '*annoyed*'

	frustrated	annoyed	disappointed	irritated	confused	relaxed	calm/ fine/ okay	happy/ content
Bd Task1	8	4	3	3	2	3	3	1
Bd Task2	7	12	1	0	0	1	6	3
Bd Task3	10	7	2	2	2	1	1	3

Fig. 10. Word-clouds and word-counts for HC tasks with breakdown at Task 1 (Left), Task 2 (Middle) and Task 3 (Right).

were the salient emotional responses, followed by '*disappointed*', '*confused*' and '*irritated*'. However, positive emotions, including '*relaxed*', '*calm*' and '*happy*', were also expressed, albeit of low frequency, suggesting that some participants were not emotionally affected by the breakdown. In fact, one participant was '*bemused*' (HCBd2), who might be aware of the experimental context. This resonates with another participant's expression: '*slightly frustrated but relieved it's only an experiment*' (HCBd2). Only one participant showed surprise: '*a bit surprised at the useless chatbot*' (HCBd1).

5.3.5.2 Bounce-Back Effect. To analyse whether the bounce-back effect as described in Section 5.3.4 would also be verified by the data of emotional words, we compared the word-clouds before and after the breakdown for the case of breakdown happening in Task 1 (Bd1) and breakdown happening in Task 2 (Bd2). Breakdown happening in Task 3 (Bd3) is deemed irrelevant for this analysis due to the lack of opportunity to recover). Figure 11 shows the case of Bd1 for HC tasks: the emotional word '*content*' is visibly more frequent in the baseline than Task 2; there are also differences, albeit to a lesser extent, for the words '*relaxed*' and '*calm*.' Furthermore, for Task 2, the unique words '*relieved*', '*reassured*' and '*satisfied*' with each occurring three times suggest that some participants recovered from their negative emotional responses in Bd1, which, for instance, instigated the emotions of '*frustrated*', '*disappointed*' and '*irritated*', but were subsequently replaced with '*reassured*'.

For Bd2, we analysed the emotional words for the tasks before and after it, Task 1 and Task 3, respectively. Figure 12 displays the word-clouds and counts for the five common words. There are small differences between the words '*content*' and '*calm*'. The words used are more diverse for Task 3. Specifically, some participants described their emotions negatively with the words '*apprehensive*', '*annoyed*', '*frustrated*' and '*concerned*', implying that they were still unpleasantly affected by the breakdown that happened in Task 2.

We applied the same process described above to analyse the emotional words/phrases for LC tasks. Similar results were obtained and presented in Appendix C to avoid prolonging the main body of the article.

5.4 Emotion-Trust Relation

To analyse the mediating role of emotion in trust (Hypothesis 8), we applied linear regression with `valence`, `activation` and `control` as predictors for each of the three tasks under each of the eight experimental conditions. The three items of `task-specific trust` (NB: the same three items for all tasks) have high Cronbach alpha ($\alpha > 0.9$). Hence, it is justified to use the average of the three items as the measure to regress on the three emotional dimensions, which are considered as orthogonal (Section 2.4). To contain the complexity, no other potential predictors (e.g., demographic

	content	relaxed	calm	happy	chilled
Before (base)	10	6	5	3	2
After (Task 2)	4	3	4	4	2
After (Task 3)	2	4	4	6	2

Fig. 11. For the HCBd1 condition, word-clouds and word-counts for the emotions measured before (Left: Baseline), and after breakdown (Middle: Task 2 with no breakdown and Right: Task 3 with no breakdown).

	content	calm	okay	happy	fine
Before (Task1)	7	7	3	4	2
After (Task3)	4	5	4	4	2

Fig. 12. For the HCBd2 condition, word-clouds and word-counts for the emotions measured before (Left: Task 1 no breakdown) and after (Right: Task 3 no breakdown).

variables) are included in the regression analysis. In Table 24, only the results with significant predictors are shown. The following patterns are observed:

—For all tasks with breakdown, irrespective of the position the breakdown occurs and regardless of Task Criticality, none of the emotional dimensions predict the `task-specific trust` significantly. This seems counterintuitive because breakdown experience likely elicits emotional responses stronger than non-breakdown ones, influencing the perception of trust (emotional trust; Section 2.4). However, it could be that the breakdown elicited a range of emotions in participants, producing inconsistent effects.

—All but one case have one significant predictor, with `valence` and `control` being more frequent than `activation`. These observations lend support to the Appraisal Tendency Framework (Han et al., 2007; Lerner and Keltner, 2001; Section 2.4) that emotion can lead to an implicit cognitive predisposition to interpret past and future events.

—For the tasks of LC, `task-specific trust` in the chatbot seems less affected by emotions than the tasks of HC.

Table 24. Regression Analysis of the Emotion-Trust Relations

	Condition	Task	Predictor	ANOVA	R^2	Regression Equation
HC	Group 1	Task2 (B0)	Control	F = 10.42; p = .003	23,90%	1.85 + 0.45 (Control)
	No Bd	Task3 (C3)	Control	F = 10.58; p = .003	26,70%	3.02 + 0.4 (Control)
	Group 2	Task2 (A2)	Valence	F = 3.95; p = .05	11,30%	4.56 + 0.23 (Valence)
	BdTask1	Task3 (C3)	Control	F = 24.44, p < .001	44,10%	1.33 + 0.61 (Control)
	Group 3	Task1 (A1)	Valence	F = 11.7; p = .002	29,40%	3.34 + 0.35 (Valence)
	Bd Task2	Task3 (C3)	Valence	F = 8.91, p = .002	24,10%	1.64 + 0.51 (Valence)
	Group 4	Task1 (A1)	Valence Activation	F = 11.57, p < .001	41,20%	4.15 + 0.37 (Valence) – 0.21 (Activation)
	Bd Task3	Task2 (C2)	Valence	F = 24.72, p < .001	42,10%	0.05 + 0.77 (Valence)
LC	Group 5	Task2 (Y0)	Control	F = 13.83, p < .001	33,10%	3.15 + 0.41 (Control)
	No Bd	Task3 (Z3)	Activation	F = 5.57, p = .025	16,60%	4.77 + 0.23 (Activation)
	Group 6	Task2 (X2)	Control	F = 14.42, p < .001	32,50%	3.07 + 0.43 (Control)
	Bd Task1					
	Group 7	Task3 (Z3)	Valence	F = 12.18, p < .002	28,90%	2.42 + 0.5 (Valence)
	Bd Task2					
	Group 8	Task1 (X1)	Valence	F = 18.79, p < .001	37,70%	1.69 + 0.53 (Valence)
	Bd Task3	Task2 (Z2)	Control	F = 11.96, p = .002	27,80%	2.94 + 0.41 (Control)

5.5 Overview of Hypotheses Verification

We summarise the evaluation outcomes of the eight hypotheses (Section 3) in Table 25. Half of the hypotheses were rejected by the empirical data whereas the others were supported, some partially, thanks to the peculiar data of `activation`.

6 Discussion

6.1 The Effect of Task Criticality on User Trust and Emotion

Our data do not support the notion that Task Criticality impacts either trust or emotion - not as a main effect, nor as an interaction effect with Conversational Breakdown. Task Criticality is not found to impact people's `overall trust` in a chatbot for customer service in consumer banking. Moreover, even when the chatbot fails to provide support (in the case of breakdown) and `task-specific trust` drops, the criticality of the task is not found to influence the extent of this drop. The same applies to emotion, where changes in `valence`, `activation` or `control` between the tasks are not influenced by Task Criticality.

It is surprising not to find any such effect of Task Criticality. Especially the lack of an interaction effect between Conversational Breakdown and Task Criticality is intriguing. From existing literature on service criticality [Ostrom and Iacobucci, 1995] and previous work on Task Criticality [Chanseau et al., 2018], we know that HC tasks should carry more significance to users, something that we assumed would make the cost of breakdown higher. However, our data do not support this notion. In light of this, we need to scrutinise whether our findings could be due to aspects of our operationalisation but also whether a nuanced view on the effect of Task Criticality on trust may be required.

Our operationalisation may be scrutinised with regard to (a) our choice of HC and LC tasks and (b) our experimental setup. Concerning our choice of tasks, this was determined through a pre-study where Task Criticality was assessed through participant ratings of perceived `risk`, `complexity`, and required `personalisation`. Task Criticality was then confirmed through a manipulation check in the pre-interaction phase of the experimental procedure. Here, the entire HC set of tasks (A, B and C) was found to be significantly different from the LC set of tasks (X, Y and Z). It should, however, be noted that Task A was rated significantly lower on `complexity` and required `personalisation` than Tasks B and C (see Appendix B). In that sense, Task A may have resembled a LC task for the experiment participants, dampening a potential effect of Task Criticality.

Concerning our experimental setup, it may be argued that although participants engaged in actual conversations with the chatbot, the tasks that they performed arguably were not of a personally

Table 25. Summary of the Evaluation Outcomes for the Eight Hypotheses

Hypothesis (H)	Status	Comment
H1: The main effect of Task Criticality on overall trust in the chatbot	Rejected	No main effect of Task Criticality
H2: The interaction effect of Task Criticality and Conversational Breakdown on overall trust in the chatbot	Rejected	No interaction effect between the two main factors was observed.
H3: The main effect of Conversational Breakdown on user's trust in the chatbot	Accepted	The Conversational Breakdown significantly reduced the user's task-specific trust and overall trust.
H4: The order effect of the task position with breakdown on (a) emotions; (b) trust	(a) Rejected (b) Rejected	Irrespective of the position of task breakdown, the extent to which the post-task trust, valence and control reduced was comparable.
H5: The main effect of the Task Criticality on (a) emotion; (b) trust	(a) Rejected (b) Rejected	No significant differences in changes of emotion or trust between the HC and LC groups, irrespective of the position of task breakdown.
H6: The bounce-back effect of (a) emotion (b) trust after breakdown	(a) Accepted (b) Accepted	Valence and control clearly demonstrated the bounce-back effect; changes in activation were negligible. Trust recovered from the breakdown task significantly.
H7: The extent of bounce-back effect proportional to the number of non-breakdown tasks after the breakdown.	Rejected	Valence increases by 1.5/1.4 times (high/low) with two non-bd tasks and by 1.4/1.6 times (high/low) with one bd; the corresponding ratios for control are 1.5/1.4 and 1.3/1.4. For Trust, the respective ratios are higher: 5.0/5.9 and 4.9/5.1. Overall, no consistent patterns.
H8: The extent to which emotions predict trust, varying with the conditions.	Accepted	Seven of the eight experimental conditions had only one of three emotion dimensions as a significant predictor. The variance of task-specific trust attributable to emotion ranges from 11% to 44%.

relevant quality due to the artificial nature of the study design. Since the hypothesised effect of Task Criticality depends on direct relevance of the task to the user, the lack of actual relevance may also have limited the potential effect of Task Criticality on trust and emotion. Such relevance, paradoxically, may be more attainable when tasks are described as one-line presentations—as they were in our pre-study and the pre-interaction manipulation check—as such one-liners may easily be interpreted in relation to the participants' personal contexts. In contrast, in the full experimental setup, the specific task details of the chatbot interaction are not personally relevant to the participants. This limitation could potentially explain why we in the pre-study and the pre-interaction manipulation check found significant correlations between participant ratings of Task Criticality and their assessments of trust requirements and whether or not they would use a chatbot for the respective tasks, whereas we did not find such an effect of Task Criticality in the post-interaction measurements of the full experiment.

However, the lack of effect of Task Criticality on the post-interaction measurements of trust and emotion may also indicate the need for a more nuanced view of this topic. In our pre-study, as well as in the pre-interaction manipulation check, participants reported significantly higher trust requirements for a chatbot to help with HC tasks than LC tasks. This is fully in line with what may be expected based on previous research [Chanseau et al., 2018; Mozafari et al., 2021; Ostrom and Iacobucci, 1995]. It is only in post-interaction that the expected effect of Task Criticality is not found. A possible interpretation of this could be that users' experience of successful task completion or

conversational breakdown during task completion overshadows any pre-interaction effect of Task Criticality. Such an interpretation of our findings may be partly in conflict with previous findings in the literature, as Mozafari et al. [2021] found a post-interaction effect of Task Criticality and chatbot disclosure on trust. As such, further research is needed to investigate a nuanced effect of Task Criticality, and whether this effect is more substantial for pre-interaction assessments of trust requirements rather than post-interaction assessments of trust.

Nevertheless, already at this point we have identified findings on Task Criticality that serve to extend current knowledge. First, we find that Task Criticality may impact users' willingness to use a chatbot for a specific task. The higher the Task Criticality, the lower the willingness to use the chatbot for that task. Furthermore, this willingness is correlated with users' pre-interaction trust requirements. This has important implications for practitioners, as particular care is required when designing conversational interactions for HC tasks. This point is detailed in Section 6.4. Second, once users take on interaction with a chatbot for customer service, the quality of the interaction is likely to determine trust far more than the criticality of their specific tasks. That is, any Conversational Breakdown is likely to have an effect on trust that dwarfs any effect of Task Criticality. This finding is compliant with, and extends, previous work by Mozafari et al. [2021], and the effect of Conversational Breakdown will be detailed in full in Section 6.2. As a final point, emotion seems to play a more important role in predicting `task-specific trust` for HC tasks than for LC tasks. Specifically, as can be seen in Table 24 (Section 5.4), `control` and `valence` were found to be significant predictors of `task-specific trust` more often for HC tasks than LC ones. This finding is in line with the assumption that HC tasks carry higher significance to users, thus more likely to elicit emotional response. However, further research is required to fully untangle the interaction of Task Criticality and emotion.

6.2 The Effect of Conversational Breakdown on User Trust and Emotion

Our data support the notion that Conversational Breakdown impacts `overall trust` and `task-specific trust` in chatbots. This impact is influenced by the order with which breakdown occurs, although there is also a clear bounce-back effect. For emotion, the data suggests a more nuanced picture: Although there are some effects, Conversational Breakdown does not impact all measures of emotion (`valence`, `activation` and `control`) in the same way as it impacts trust.

6.2.1 Overall Trust. Our findings extend earlier work by Law et al. [2022] by showing that there is an observable order effect for breakdown: The later breakdown happens in a conversation, the bigger the impact it has on `overall trust`. It is likely that the recency effect [Murdock, 1962] plays a role in this, as a more recent breakdown is easier for participants to recall once they are presented with the `overall trust` questionnaire at the end of the experiment. This implies that any damage to `overall trust` by Conversational Breakdown may be restored simply with the passing of time.

Another possible explanation for this order effect of breakdown on `overall trust` lies in exposure to successful subsequent tasks. In our experimental setup, if breakdown happened in Task 1, the subsequent Tasks 2 and 3 would be completed successfully. This would allow the chatbot to show more competence than when breakdown occurred in Task 2 or even Task 3. In other words, the earlier a breakdown occurs in a conversation, the more possibilities there are for the participant to be exposed to successful task completion, which in turn can repair trust.

The two explanations for this breakdown order effect provide complementary perspectives on the phenomenon, and hold potentially important implications for practitioners which are detailed in Section 6.4.

6.2.2 Task-Specific Trust. Conversational Breakdown negatively impacts `task-specific trust`. This finding replicates and extends previous work. Specifically, as found by Law et al. [2022], Conversational Breakdown strongly impacts users' trust in the chatbot's capability for providing support with a given task. This previous work also indicated a possible bounce-back effect where users' trust in a chatbot's capability to provide support for a task may not be undermined by Conversational Breakdown for a previous task—given that the new task is successfully completed. Our findings support this indication from previous work, which further builds the case that one bad conversational experience may not necessarily impact users' assessment of subsequent interactions. This suggests the importance of designing chatbot interactions in a way where users are nudged towards continued interactions with a chatbot in spite of Conversational Breakdown, as negative trust consequences of such breakdown may be mitigated through successful future interactions.

While our assumption of a bounce-back effect was supported, we found no significant effect of the breakdown order on `task-specific trust`. Hence, the breakdown order seems to have a more profound effect on `overall trust` than trust in a specific task. Still, the data seem indicative of a potential upward trend, where later breakdowns were punished less severely than earlier ones. The two groups that had a breakdown in Task 3 (HCBd3 and LCBd3) gave higher `task-specific trust` ratings for this task than the groups who were faced with breakdown earlier, as can be seen in Table 15 (Section 5.3.1). It is unclear from our data why this would be the case. It may be that participants are reluctant to give the chatbot a bad `task-specific trust` rating if they have given it two good ratings before. Possibly, this phenomenon can be explained in terms of Cognitive Dissonance [Morvan and O'Connor, 2017]. According to this theory, discomfort is instigated when a person's belief clashes with new information perceived; the person may attempt to resolve the contradiction in order to reduce their discomfort by altering their belief.

6.2.3 Emotion. The effect of Conversational Breakdown on emotion is partly in line with our initial assumptions. `Valence` and `control` are negatively impacted by Conversational Breakdown, but `activation` showed no such effect. This indicates that users experience breakdown as something unpleasant and out of their control, but at the same time as neutral rather than exciting or calming. This interpretation of our findings is further corroborated by the qualitative data, showing high frequency occurrences of words like 'frustrated' and 'annoyed' for breakdown tasks. Our data does not support the notion that there is an order effect on any of the emotional measures, so regardless of when breakdown occurs, `valence` and `control` are equally impacted.

However, as soon as participants encountered a successful task completion after breakdown, both `valence` and `control` show a significant bounce-back effect similar to that of `task-specific trust`. Participants described their post-breakdown emotions with words like 'relieved' and 'reassured', which illustrates the fact that there is a certain recovery aspect to this bounce-back effect: Their emotional response to the successful task is influenced by the prior unsuccessful breakdown task. Moreover, both `valence` and `control` bounce back to levels comparable to the baseline, indicating that emotional states are not necessarily impacted long-term after encountering Conversational Breakdown.

Surprisingly, `activation` showed little to no significant changes regardless of experimental condition, and no emotional measure was found to significantly predict `task-specific trust` for breakdown tasks. Both are perplexing results, as tasks with breakdown can be assumed to elicit stronger emotional responses. As already pointed out in Section 6.1, this may have been caused by the laboratory setting of this experiment, which could have made tasks less salient to participants: They were not real for the participants, which could explain why emotional responses were less intense than they would be in real life.

6.3 Implications for Theory

6.3.1 Implications for Theory on Trust in Chatbots. Our study has important implications for theory on users' trust in chatbots for customer service. We find four implications to be of particular interest and detail these in dedicated subsections below.

6.3.1.1 Perceived Task Criticality May Determine Users' Trust Requirements and Inclination to Use a Chatbot. We find that when considering whether to use a chatbot for a specific task, users' considerations may be predicted by their perceptions of Task Criticality and task trust requirements. In the pre-study, the participants consistently rated tasks characterised by higher perceived criticality—that is, higher `risk`, higher perceived demand for `personalisation` and higher `complexity`—as requiring more trust, and as being less likely to seek out help for using a chatbot. Here, the participants' assessments were in line with what may be expected from the organisational trust model of Mayer et al. [1995], as well as trust models for the technology domain [McKnight et al., 2011; Lankton et al., 2015], where trust is considered a basis for risk-taking behaviour. In our study, we expand on this notion, as we demonstrate that Task Criticality may determine users' trust requirements and their willingness to apply chatbots for a given task. In doing this, our study also adds to the insight provided by Mozafari et al.'s [2021] study of service criticality and users' trust in chatbots. Specifically, we demonstrate that users actively consider the task for which they may use a chatbot for customer service in order to assess its criticality and perceived risk.

6.3.1.2 Post-Interaction Trust Assessments of a Chatbot May Be Determined by Users' Experience During Interaction Rather than Perceived Task Criticality. Our findings of the effect of Conversational Breakdown on task-specific and overall trust show that users give their experience during interaction much weight when assessing any post-interaction trust in a chatbot. At the same time, in a post-interaction situation, users seem to a lesser degree to rely on any initial perceptions of Task Criticality. This finding is intriguing, as it shows both that users' trust assessments for a technology may be directly impacted by their immediate experience, and also that this impact may potentially reduce any impact of their pre-interaction perceptions of the criticality of the task at hand or their naïve inclination to use or not use a chatbot for a specific task. This ability of users to update their trust assessments is foreshadowed in established trust models. As noted by Mayer et al. [1995], while trust will be determined by factors of perceived trustworthiness, these factors will in turn be determined by the experienced outcomes of a relation. The notion of prior experiences with a technology determining users trust in the technology is also reflected in technology trust models. For example, McKnight et al.'s [2011] concept of situation normality, where experience-based assumptions of the adequacy of a specific technology may determine user trust, foreshadows our findings regarding the significant impact of experience. Adding to this theory base, our findings suggest that trust in chatbots may be strongly driven by specific experiences following interaction, and that such experiences may limit the effect of contextual factors such as Task Criticality.

6.3.1.3 A Bounce-Back Effect in Trust Following Conversational Breakdown Suggests Task Specificity of Trust Assessments. Our findings corroborate previous work of Law et al. [2022] and provide detailed insight into a bounce-back effect showing that it may appear regardless of whether a breakdown and subsequent successful interaction occurs relatively early or late in a chatbot interaction. The bounce-back effect in `task-specific trust` has important theoretical implications as it suggests that while theoretical models of trust predict that experience may indeed impact future trust assessments—both for trust relations between humans [Mayer et al., 1995] and trust in technology [Lankton et al., 2015; McKnight et al., 2011]—this impact may be highly contextually determined. That is, the impact of a problematic experience when using a chatbot for one specific task may to a substantial degree be contained to that task rather than spilling over to other tasks.

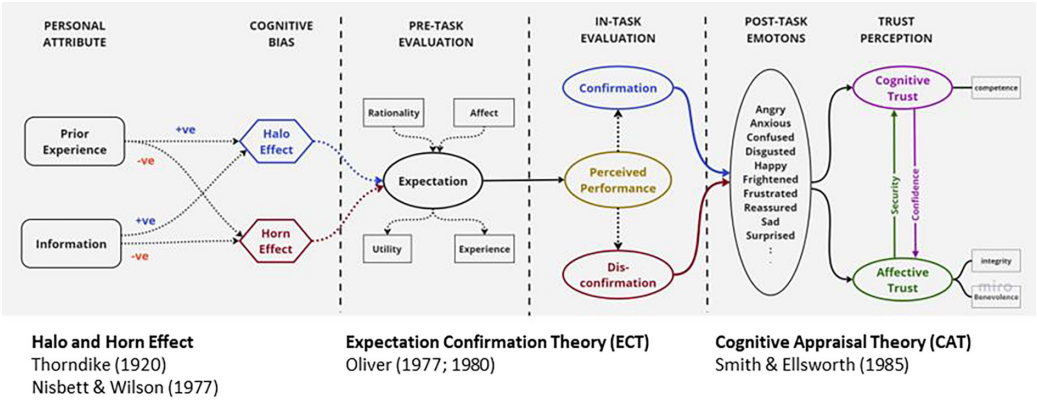

Fig. 13. Phases of research workflow and associated theoretical models on emotion.

Hence, when considering whether to be able to rely or depend on a chatbot for customer service for a specific task, users may rely on their experience with this technology for this type of task more so than their experience with the chatbot for other, unrelated, tasks. This finding may motivate a more nuanced view on trust in technology in general, as well as trust in chatbots for customer service, as it suggests the importance of considering the moderating effect of context. Specifically, when users have mixed experiences with a chatbot, they may well be able to discriminate between tasks for which they would trust it and tasks for which they would not—rather than only relying on a global trust assessment.

6.3.1.4 A Potential Recency Effect in the Impact of Conversational Breakdown on Overall Trust suggests a Nuanced Interplay between Chatbot Trust Assessments Overall and for Specific Tasks. The fourth implication of our study on theory of trust concerns the observed recency effect [Murdock, 1962] of Conversational Breakdown on `overall trust` in the chatbot. While also other explanations may be possible for this finding, a recency effect is likely to explain at least some of this observation from our study. For trust theory, we find the observed recency effect to be particularly important as it suggests a nuanced interplay between `task-specific trust` and `overall trust` in a chatbot. While `task-specific trust` may bounce back, when users successfully accomplish a new task following a previous failed task, this bouncing back does not unconditionally apply to `overall trust`. Rather, `overall trust` is likely to be impacted by breakdown in one of a series of tasks. However, the impact of breakdown arguably is less severe for `overall trust` than `task-specific trust`. Furthermore, the impact of breakdown on `overall trust` is determined not only by individual task-specific experiences but also the order or interrelation between these experiences. This finding extends current theories of trust in technology [Lankton et al., 2015; McKnight et al., 2011], as it indicates that while it may be important to separately consider `task-specific` and `overall trust` in a technology, there is nevertheless a nuanced interplay between these trust aspects. Future research is needed to elaborate on this interplay.

6.3.2 Theoretical Models on Emotions. Emotion and cognition are highly intertwined psychological constructs (Section 2.4). We posit that certain theories can be employed to interpret the phenomena observed in our study; Figure 13 depicts the phases of our research workflow and associated theoretical models.

For the sake of brevity, we do not report the analysis of prior experience for this study. Nevertheless, our previous work confirmed the significant halo and horn effect [Law et al., 2023], imparting positive and negative opinions formed through previous interactions with chatbots in general on the current specific one.

We analysed participants' ratings of the five pre-interaction task-related attributes and three baseline emotions (Section 4.2), which could roughly be mapped to 'rationality/utility' and 'affect/experience' aspects of the Expectation Confirmation Theory (Figure 13), respectively. There were significant differences in the five attributes (the average over the set of three tasks) between the HC and LC groups and among the eight experimental conditions whereas no significant differences in emotion measures were detected (Appendix B). However, when comparing the four groups within the same criticality level (HCBd0–HCBd3 and LCBd0–LCBd3; Table 12), no significant differences in any of the five attributes for any of the three tasks were identified. In other words, participants of the same Task Criticality group started with comparable perceptions (expectations) of the tasks and emotions, and the significant differences in the five pre-interaction attributes were largely explicable in terms of Task Criticality. However, the effect of Task Criticality dissipated in the context of actual task interaction.

While we did not systematically measure the (dis)confirmation of expectations with a scale [e.g., Bhattacherjee, 2001], which would be beyond the scope of this article, we assumed that the impact of Conversational Breakdown would function as negating the expectation about the chatbot's performance, leading to the visible changes in `valence` and `control` and both `task-specific` and `overall trust`. Arguably, the Expectation Confirmation Theory was indirectly supported in this regard. It leaves an intriguing question on how the altered expectation could be recalibrated when the performance of the chatbot returned to its originally expected level. Could the lowered expectation be catapulted by the improved user experience to a much higher level, making it even more difficult to meet this heightened expectation, and thus making it harder to enhance trust in subsequent tasks (e.g., the case of HCBd1; Section 5.3.4.1)? Was the expectation recalibrated with reference to the pre-interaction state or could it be recalibrated to a more realistic level with exposure to more instances of regular performance (cf. Hypothesis 7; Table 25)? The Expectation Confirmation Theory seems not addressing the notion of recalibration and hence has limited power to explicate the observed bounce-back effect of trust and emotion (Section 5.3.5).

The applicability of the Cognitive Appraisal Theory for interpreting emotion-trust relation is not that intuitive. Emotional states engendered by the breakdown, as indicated by high level of `valence` and `control` and verbal expressions (Section 5.3.5), were largely annoyance and frustration, though other non-negative emotions were also elicited. As delineated in Section 2.4, according to the Appraisal Tendency Framework as derived from the Cognitive Appraisal Theory, anger is seen as in control by the other individual (agent), reducing trust in the agent; anger is also seen as high certainty, exerting no influence on trust. The control appraisal model could explain the post-task changes of emotion and trust of which subsequent recovery could be understood in terms of the deactivation of appraisal tendency [Han et al., 2007], given the resolution of the chatbot's performance issue. The deactivation mechanism may imply that emotion plays no further role in trust after such breakdown-resolution episodes, but our results (Table 24) did not support this speculation. In fact, the emotional state of being content/happy, which, same as anger, is also seen as other controlled and high certainty, seems contributing to `valence` and `control` as significant predictors for trust in Task 2 and Task 3 (i.e., the row of HCBd1 in Table 24; Figure 10). However, the emotion and trust levels were plateaued. Overall, the Appraisal Tendency Framework and Cognitive Appraisal Theory entail refinements to interpret the complex emotion-trust relation, especially when multiple emotions can be at play at a particular moment.

6.4 Implications for Practice

The key findings from our study when it comes to practical implications is the effect of Task Criticality, or a lack thereof, as well as the order effect of Conversational Breakdown. Specifically, we find it of high practical interest that Task Criticality mainly is found to have an effect on

pre-interaction perceptions of tasks. Furthermore, there is substantial practical interest in knowing that Conversational Breakdown should not be the last thing that happens in a chatbot conversation. Below, we detail three types of practical implications from these findings: strategic, model-building, and abandonment and measurement.

6.4.1 Strategic Implications. Perceived Task Criticality entails a strategic communication challenge. Even though the effect of Task Criticality may be negligible once users interact with the chatbot, it is still a challenge for practitioners to motivate users to initiate chatbot interaction for tasks perceived as HC. This engagement challenge is a substantial hurdle to realising chatbots' potential and return on investment for practitioners. Our Task Criticality measurement, however, provides some support to tackle this challenge. Specifically, this measurement gives practitioners insight into the different dimensions underlying criticality, which allows them to better understand users' reasons for non-engagement and thereby tailor the chatbot communication to address these. For example, for consumer banking, a task like 'cancelling a double payment you made by mistake' may score particularly high on the `personalisation` dimension of criticality. To show users that the chatbot can actually achieve the required level of personalisation, marketing and in-chat communication may emphasise what kind of back-end systems and/or information the chatbot has access to. Hence, user research on the dimensions of Task Criticality may provide needed information for strategic communication towards user engagement even for HC tasks.

User research on Task Criticality may also support strategic scoping decisions for a customer service chatbot. For example, when expanding on the task set for which a chatbot can provide support—e.g., by expanding on the chatbot intent base [Følstad and Taylor, 2021] or developing new integrations with backend systems [Kvale et al., 2020]—practitioners may prioritise tasks perceived as LC to achieve rapid uptake of new opportunities for support provided through the chatbot. Furthermore, research on user perceptions of Task Criticality may also help determine the scope of tasks for which a chatbot is seen by users as a relevant source of support. For tasks falling outside this scope, practitioners may consider lowering the threshold for other channels of support—e.g., through dedicated customer service personnel. At the same time, since users may not distinguish between HC and LC tasks once they have taken up a chatbot for a particular task, practitioners should be careful not to let perceptions of Task Criticality fully dictate their decision to have a task in- or out of scope.

6.4.2 Model-Building Implications. The importance of the chatbot understanding the question and providing the correct response—in other words, preventing Conversational Breakdown—highlights the need to build and maintain accurate models for intent-prediction and execution of corresponding actions in chatbots for customer service. Before the introduction of LLMs such as ChatGPT, this meant that practitioners had to diligently discover and build intents for all of the different questions users may ask as well as provide the intent-prediction models with sufficient training material for it to understand each intent correctly [Følstad and Taylor, 2021]. This is a time-consuming task for which it is not always clear how much extra value it will yield per additional built intent.

In contrast, LLMs can be connected to existing knowledge bases, covering an extremely wide scope, without the creation of as many intents and training data as required in purely intent-based chatbots. Although there are substantial downsides of using LLMs, e.g., in terms of bias, reliability, and accuracy (as illustrated by e.g., [Weidinger et al., 2022]), which we will not go into for the sake of brevity in this article, there may also be marked benefits: Adding LLM functionality to existing intent-based chatbots has the potential to drastically decrease Conversational Breakdown by expanding the chatbot's existing knowledge at a fraction of the costs associated with such an expansion for intent-based models. Based on our findings, practitioners should look into the

best way to balance benefits and drawbacks of using LLMs in complement with intention-based solutions, for their chatbot setup to limit Conversational Breakdown.

6.4.3 Abandonment and Measurement Implications. Last, our results further strengthen the need for the identification of conversations that show abandonment—an event commonly referred to as 'drop-off' in the industry. Drop-off is usually sought after using process mining techniques and/or manual conversation tagging because they present lost automation potential: Users that abandon a chatbot conversation typically still need assistance and are likely to engage with another customer service channel which may require human resources. For example, when a user tries to order an insurance for their car for a journey abroad but the chatbot replies for domestic insurance packages, this Conversational Breakdown may lead to the user abandoning the chat to find correct information elsewhere—potentially through a human channel like the telephone. However, our results indicate that there is another cost associated with drop-off, especially if it occurs directly after Conversational Breakdown: users will have lower overall trust ratings in the chatbot, which also reduces the likelihood of them using the chatbot again in the future [Law et al., 2022]. Hence, drop-off may substantially reduce the effectiveness of automation for a specific conversation. Furthermore, it may also limit future automation effectiveness as users are likely to seek out alternative channels instead of returning to the chatbot.

6.5 Limitations and Future Research

We can identify several limitations from our current work and suggest interesting avenues for future studies.

6.5.1 Limitation 1: Bounce-Back of Overall Trust: Lack of Longitudinal Data. We have shown that `task-specific trust` ratings tend to restore through bounce-back within one subsequent task completion. Similarly, our findings suggest that `overall trust` also may have a resembling trend towards restoration: when breakdown happens early in the experiment, `overall trust` is found to suffer less than when breakdown happens late. However, our data do not allow us to conclude whether this is simply because of the passing of time, or because participants were exposed to successful task completions after breakdown.

Firstly, it may be the case that `overall trust` restores after breakdown simply with the passing of time. However, we were not able to assess whether this is the case as our study setup is limited to single, brief interactions in a lab-based setting. Future studies should aim to measure trust as it develops over time, in particular as users are likely to assess and recalibrate trust over time [de Visser, 2020].

Secondly, if exposure to successful tasks is required to restore `overall trust`, it would be valuable to understand more about these successful tasks. For example, does a bigger drop in trust require more successful tasks to allow for restoration? Do the successful tasks need to be topically related to the task with breakdown? Can a repair message count as a successful task? These are all interesting questions to pursue in future work.

6.5.2 Limitation 2: Lack of Behavioural Data. Participants were asked to report on the likelihood of them using chatbots for the specific tasks (the `useChatbot` measure), but we did not follow-up to assess whether their future behaviour actually matched their indications. As trust is a prerequisite for people to engage in behaviour that is risky [Mayer et al., 1995], behaviour would be a very strong variable to indirectly measure (changes in) trust. Moreover, from a practical point of view, the resulting behaviour matters more than self-reported attitudes, which means that results would have been easier to translate into practical value.

Future studies should incorporate behavioural trust measures as well as, or instead of, attitudinal ones. This could work well with any approach to tackle the above-mentioned lab-based single study limitations, where measuring behavioural changes over time as well as self-reported assessments on trust could provide a very rich dataset.

6.5.3 Limitation 3: Preventing Abandonment. The order effect of breakdown on `overall trust` highlights the practical importance of preventing chat abandonment directly after breakdown. In our experimental setting, participants were required to complete all three tasks and hence could not leave the conversation, whereas they might have done so in real life. Would a repair message be sufficient to keep people from closing a chat after breakdown? What other kinds of strategies would work to prevent abandonment? Our results do not necessarily give insight into the best way of doing so.

One potential avenue to prevent chat abandonment may be through conversational repair [Ashktorab et al., 2019]. However, as described in Section 2.2, conversational repair can be anything from acknowledging a misunderstanding to actually successfully completing the failed task. It is not clear which tactic is required or sufficient to convince a user to stay within the chat after breakdown. For example, is it sufficient to acknowledge the misunderstanding, or should the chatbot also indicate a way forward, for instance, suggesting the completion of other tasks? Potentially other avenues of abandonment prevention may also be explored in future studies.

6.5.4 Limitation 4: Limited Enrichment from Qualitative Data. Our study's data are skewed towards quantitative data, with only limited qualitative data to enrich our results. However, the qualitative data in our study nevertheless suggests the value of such data in an investigation of trust and emotion implications of Conversational Breakdown. For example, using quantitative data only would not have allowed us to see that the bounce-back effect from `valence` and `control` involved a sense of relief or reassurance.

While we indeed gather and learn from qualitative data associated with our analysis of impact of Conversational Breakdown, it seems reasonable to assume that a more extensive gathering of qualitative data would have served to further enrich our findings. For example, our study does not include qualitative data which might have expanded our understanding of how participants understand Conversational Breakdown or how they themselves consider such breakdowns to impact overall and task specific trust. Future studies should aim to include more qualitative data in study designs to further strengthen our understanding of this important topic.

6.5.5 Limitation 5: Artificial Experimental Setup. Finally, the experimental setup where users engage in tasks not directly related to their own personal context represents a limitation in the study design—particularly in terms of ecological validity [Subramanian et al., 2023] which in HCI research typically concerns aspects of the study setting, users, and research methods.

To address this limitation, we foresee future studies where data from real-world user interactions with chatbots for customer service may be applied for validation and expanding on the current findings. While such use of real-world data may entail limitations in terms of study rigour, such future work could nevertheless complement the knowledge base—in particular is conducted as extensions of research conducted through complementary methods.

7 Conclusion

We have presented a study contributing to the knowledge base on the impact of Conversational Breakdown in user interactions with chatbots for customer service. Specifically, we have studied the impact of such a breakdown on users' emotion and trust in the chatbot in which the breakdown occurs, and also shed light on the importance of when breakdown happens and the characteristics of the task.

In our study, we indeed find a marked effect of Conversational Breakdown on user emotion and trust. In line with previous research, we verify that such a breakdown negatively impacts `task-specific trust` and `overall trust`, albeit to different degrees. Whereas the impact on `task-specific trust` is immediate and extensive, the impact on `overall trust` may depend on other aspects of the chatbot interaction—specifically whether breakdown happened relatively early or late in an interaction sequence. This finding has substantial theoretical and practical implications as it sheds light both on users' capability of distinguishing between tasks when making trust assessments as well as the possibility for service providers to remedy negative trust implications of breakdown in interaction.

The study also provided important insight into the importance of Task Criticality for trust in chatbots. Specifically, we find evidence that users could make consistent assessments regarding Task Criticality and whether they would trust a chatbot with different tasks. However, while such assessments of Task Criticality may be important determinants as to whether users would attempt to use a chatbot for a given task, Task Criticality may not be an important determinant of trust following interaction. Rather, the users' experiences during interaction—positive or negative—seem to outweigh any pre-interaction assessment of Task Criticality when making their post-interaction trust assessment. This finding also has important theoretical and practical implications as it shows how users' assessments of trust depend on hands-on interaction experiences. Furthermore, it suggests to service providers the importance of motivating users to engage in chatbot interaction also for HC tasks, provided that the chatbot can deliver the expected level of quality in interaction.

Finally, the study provided interesting insight into the relationship between emotion and trust during chatbot interaction, and how emotion and trust change in parallel throughout chatbot interaction. This finding suggests that trust assessments, while clearly impacted by rational assessments, also substantially correlate with emotional states, especially for the emotional dimensions of `valence` and `control`. Future research is however needed to fully unpack the intricate interplay between emotion and trust in chatbot interaction.

References

Martin Adam, Michael Wessel, and Alexander Benlian. 2021. AI-based chatbots in customer service and their effects on user compliance. *Electronic Markets* 31, 2 (2021), 427–445. DOI: https://doi.org/10.1007/s12525-020-00414-7

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*. ACM, New York, NY, article 254. DOI: https://doi.org/10.1145/3290605.3300484

Lisa Feldman Barrett and Christina Westlin. 2021. Navigating the science of emotion. In *Emotion Measurement*. Herbert L. Meiselman (Ed.), Woodhead Publishing, 39–84. DOI: https://doi.org/10.1016/B978-0-12-821124-3.00002-8

Michaela Bartosova, Miroslav Svetlak, Martina Kukletova, Petra Borilova Linhartova, Ladislav Dusek, Lydie Izakovicova Holla. 2019. Emotional stimuli candidates for behavioural intervention in the prevention of early childhood caries: A pilot study. *BMC Oral Health* 19, 33 (2019). DOI: https://doi.org/10.1186/s12903-019-0718-4

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 *Proceedings of International Conference on Information Systems (ICIS '21)*. AIS eLibrary, 1–17.

Anol Bhattacherjee. 2001. Understanding information systems continuance: An expectation-confirmation model. *MIS Quarterly* 25, 3 (2001), 351–370. DOI: https://doi.org/10.2307/3250921

Margaret M. Bradley and Peter J. Lang. 1994. Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry* 25, 1 (1994), 49–59. DOI: https://doi.org/10.1016/0005-7916(94)90063-9

Petter Bae Brandtzaeg and Asbjørn Følstad. 2017. Why people use chatbots. In *Proceedings of the Internet Science: 4th International Conference (INSCI '17)*. Springer, 377–392. DOI: https://doi.org/10.1007/978-3-319-70284-1_30

Tom Brown, Bejnamin Mann, Nick Ryder, Melanie Subbiah, jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario

Amodei 2020. Language models are few-shot learners. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS '20)*, 1877–1901.

Adeline Chanseau, Kerstin Dautenhahn, Michael L. Walters, Kjeng Lee Koay, Gabriella Lakatos, and Maha Salem. 2018. Does the appearance of a robot influence people's perception of task criticality? In *Proceedings of the 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN '18)*. IEEE, 1057–1062. DOI: https://doi.org/10.1109/ROMAN.2018.8525663

Jing Chen, Scott Mishler, and Bin Hu. 2021. Automation error type and methods of communicating automation reliability affect trust and performance: An empirical study in the cyber domain. *IEEE Transactions on Human-Machine Systems* 51, 5 (2021), 463–473. DOI: https://doi.org/10.1109/THMS.2021.3051137

Cammy Crolic, Felipe Thomaz, Rhonda Hadi, and Andrew T. Stephen. 2022. Blame the bot: Anthropomorphism and anger in customer–chatbot interactions. *Journal of Marketing* 86, 1 (2022), 132–148. DOI: https://doi.org/10.1177/00222429211045687

Ewart J. de Visser, Marieke M. M. Peeters, Malte F. Jung, Spencer Kohn, Tyler H. Shaw, Richard Pak, and Mark A. Neerincx. 2020. Towards a theory of longitudinal trust calibration in human–robot teams. *International Journal of Social Robotics* 12, 2 (2020), 459–478. DOI: https://doi.org/10.1007/s12369-019-00596-x

Pieter Desmet. 2018. Measuring emotion: Development and application of an instrument to measure emotional responses to products. In *Proceedings of the Funology 2: From Usability to Enjoyment*. Springer, 391–404. DOI: https://doi.org/10.1007/1-4020-2967-5_12

Drift. 2018. *The 2018 State of Chatbots Report.* Technical Report, Drift.

Phoebe C. Ellsworth. 2013. Appraisal theory: Old and new questions. *Emotion Review* 5, 2 (2013), 125–131. DOI: https://doi.org/10.1177/1754073912463617

Md Abdullah Al Fahim, Mohammad Maifi Hasan Khan, Theodore Jensen, Yusuf Albayram, and Emil Coman. 2021. Do integral emotions affect trust? The mediating effect of emotions on trust in the context of human-agent interaction. In *Proceedings of the Designing Interactive Systems Conference (DIS '21)*. ACM, New York, NY, 1492-1503. DOI: https://doi.org/10.1145/3461778.3461997

Asbjørn Følstad, Theo Araujo, Effie L.-C. Law, Petter Bae Brandtzaeg, Symeon Papadopoulos, Lea Reis, Marcos Baez, Guy Laban, Patrick McAllister, Carolin Ischen, Rebecca Wald, Fabio Catania, Raphael Meyer von Wolff, Sebastian Hobert, and Ewa Luger. 2021. Future directions for chatbot research: An interdisciplinary research agenda. *Computing* 103, 12 (2021), 2915–2942. DOI: https://doi.org/10.1007/s00607-021-01016-7

Asbjørn Følstad and Marita Skjuve. 2019. Chatbots for customer service: user experience and motivation. In *Proceedings of the International Conference on Conversational User Interfaces (CUI '21)*. ACM, New York, NY, article 1. DOI: https://doi.org/10.1145/3342775.3342784

Asbjørn Følstad and Cameron Taylor. 2021. Investigating the user experience of customer service chatbot interaction: A framework for qualitative analysis of chatbot dialogues. *Quality and User Experience* 6, 1 (2021), 1–17. DOI: https://doi.org/10.1007/s41233-021-00046-5

Gartner. 2019. *Market Guide for Virtual Customer Assistants.* Technical Report. Gartner. Retrieved from https://www.gartner.com/en/documents/3947357

Gartner. 2022. Gartner Predicts Conversational AI will Reduce Contact Center Agent Labor Costs by $80 Billion in 2026. Retrieved from https://www.gartner.com/en/newsroom/press-releases/2022-08-31-gartner-predicts-conversational-ai-will-reduce-contac

Gartner 2023. *Magic Quadrant for Enterprise Conversational Platforms.* Technical Report. Gartner. Retrieved from https://boost.ai/guides/gartner-magic-quadrant-for-enterprise-conversational-ai-platforms/

Demijan Grgić, Vedran Podobnik, and Arthur Carvalho. 2022. Deriving and validating emotional dimensions from textual data. *Expert Systems with Applications* 198 (2022), 116721. DOI: https://doi.org/10.1016/j.eswa.2022.116721

Dogan Gursoy, Oscar Hengxuan Chi, Lu Lu, and Robin Nunkoo. 2019. Consumers acceptance of artificially intelligent (AI) device use in service delivery. *International Journal of Information Management* 49 (2019), 157–169. DOI: https://doi.org/10.1016/j.ijinfomgt.2019.03.008

Seunghee Han, Jennifer S. Lerner, and Dacher Keltner. 2007. Feelings and consumer decision making: The appraisal-tendency framework. *Journal of Consumer Psychology* 17, 3 (2007), 158–168. DOI: https://doi.org/10.1016/S1057-7408(07)70023-2

Peter A. Hancock, Deborah R. Billings, Kristin E. Schaefer, Jessie Y. C. Chen, Ewart J. de Visser, and Raja Parasuraman. 2011. A meta-analysis of factors affecting trust in human-robot interaction. *Human Factors* 53, 5 (2011), 517–527. DOI: https://doi.org/10.1177/0018720811417254

Peter A. Hancock, Theresa T. Kessler, Alexandra D. Kaplan, John C. Brill, and James L. Szalma. 2021. Evolving trust in robots: Specification through sequential and comparative meta-analyses. *Human Factors* 63, 7 (2021), 1196–1229. DOI: https://doi.org/10.1177/0018720820922080

Florian Heimerl, Steffen Lohmann, Simon Lange, and Thomas Ertl. 2014. Word cloud explorer: Text analytics based on word clouds. In *Proceedings of the IEEE 47th Hawaii International Conference on System Sciences (HICSS '14)*. IEEE, 1833–1842. DOI: https://doi.org/10.1109/HICSS.2014.231

Sebastian Hobert, Asbjørn Følstad, and Effie L.-C. Law. 2023. Chatbots for active learning: A case of phishing email identification. *International Journal of Human-Computer Studies* 179 (2023), 103108. DOI: https://doi.org/10.1016/j.ijhcs.2023.103108

Kevin Anthony Hoff and Masooda Bashir. 2015. Trust in automation: Integrating empirical evidence on factors that influence trust.*Human Factors* 57, 3 (2015), 407–434. DOI: https://doi.org/10.1177/0018720814547570

Carroll E. Izard. 1993. Four systems for emotion activation: Cognitive and noncognitive processes. *Psychological Review* 100, 1 (1993), 68–90. DOI: https://doi.org/10.1037/0033-295x.100.1.68

Sherrie Y. X. Komiak and Izak Benbasat. 2006. The effects of personalization and familiarity on trust and adoption of recommendation agents. *MIS Quarterly* 30, 4 (2006), 941–960. DOI: https://doi.org/10.2307/25148760

Knut Kvale, Eleonora Freddi, Stig Hodnebrog, Olav Alexander Sell, and Asbjørn Følstad. 2021. Understanding the user experience of customer service chatbots: What can we learn from customer satisfaction surveys? In *Proceedings of the CONVERSATIONS 2020*. Springer, 205–218. DOI: https://doi.org/10.1007/978-3-030-68288-0_14

Knut Kvale, Olav Alexander Sell, Stig Hodnebrog, and Asbjørn Følstad. 2020. Improving conversations: Lessons learnt from manual analysis of chatbot dialogues. In *Proceedings of the CONVERSATIONS 2019*. Springer, 187–200. DOI: https://doi.org/10.1007/978-3-030-39540-7_13

Nancy K. Lankton, D. Harrison McKnight, John Tripp. 2015. Technology, humanness, and trust: Rethinking trust in technology. *Journal of the Association for Information Systems* 16, 10 (2015), 880–918. DOI: https://doi.org/10.17705/1jais.00411

James Lappeman, Siddeeqah Marlie, Tamryn Johnson, and Sloane Poggenpoel. 2023. Trust and digital privacy: Willingness to disclose personal information to banking chatbot services. *Journal of Financial Services Marketing* 28 (2023), 337–357. DOI: https://doi.org/10.1057/s41264-022-00154-z

Bettina Laugwitz, Theo Held, and Martin Schrepp. 2008. Construction and evaluation of a user experience questionnaire. In *Proceedings of the 4th Symposium of the Workgroup Human-Computer Interaction and Usability Engineering of the Austrian Computer Society (USAB '08)*. Springer, 63–76. DOI: https://doi.org/10.1007/978-3-540-89350-9_6

Effie L.-C. Law, Asbjørn Følstad, and Nena van As. 2022. Effects of humanlikeness and conversational breakdown on trust in chatbots for customer service. In *Proceedings of the Nordic Human-Computer Interaction Conference* (NordiCHI '22). ACM, New York, NY, Article 56. DOI: https://doi.org/10.1145/3546155.3546665

Effie L.-C. Law, Nena van As, and Asbjørn Følstad. 2023. Effects of prior experience, gender, and age on trust in a banking chatbot with(out) breakdown and repair. In *Proceedings of the Human-Computer Interaction (*INTERACT '23), Springer, 277–296. DOI: https://doi.org/10.1007/978-3-031-42283-6_16

Richard S. Lazarus. (1991). Cognition and motivation in emotion. *American Psychologist* 46, 4 (1991), 352–367. DOI: https://doi.org/10.1037//0003-066x.46.4.352

Jennifer S. Lerner and Dacher Keltner. 2001. Fear, anger, and risk. *Journal of Personality and Social Psychology* 81, 1 (2001), 146–159. Retrieved from https://psycnet.apa.org/doi/10.1037/0022-3514.81.1.146

Chi-Hsun Li, Su-Fang Yeh, Tang-Jie Chang, Meng-Hsuan Tsai, Ken Chen, and Yung-Ju Chang. 2020. A conversation analysis of non-progress and coping strategies with a banking task-oriented chatbot. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI '20)*. ACM, New York, NY, Article 82. DOI: https://doi.org/10.1145/3313831.3376209

Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. Halueval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 6449–6464.

Ping Liu, Ya'nan Wang, Jiang'ning Hu, Lin'bo Qing, and Ke Zhao. 2023. Development and validation of a highly dynamic and reusable picture-based scale: A new affective measurement tool. *Frontiers in Psychology* 13 (2023), 1078691. DOI: https://doi.org/10.3389/fpsyg.2022.1078691

Yidu Lu and Nadine Sarter. 2019. Feedback on system or operator performance: Which is more useful for the timely detection of changes in reliability, trust calibration and appropriate automation usage?. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. SAGE Publications, 312–316. DOI: https://doi.org/10.1177/1071181319631345

Bei Luo, Raymond Y. K. Lau, Chunping Li, and Yain-Whar Si. 2022. A critical review of state-of-the-art chatbot designs and applications. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 12, 1 (2022), e1434. DOI: https://doi.org/10.1002/widm.1434

Sascha Mahlke, and Michael Minge. 2008. Consideration of multiple components of emotions in human-technology interaction. In *Affect and Emotion in Human-Computer Interaction: From Theory to Applications*. Springer, 51–62. DOI: https://doi.org/10.1007/978-3-540-85099-1_5

Roger C. Mayer, James H. Davis, and F. David Schoorman. 1995. An integrative model of organizational trust. *Academy of Management Review* 20, 3 (1995), 709–734. DOI: https://doi.org/10.2307/258792

D. Harrison Mcknight, Michelle Carter, Jason Bennett Thatcher, and Paul F. Clay. 2011. Trust in a specific technology: An investigation of its components and measures. *ACM Transactions on Management Information Systems (TMIS)* 2, 2 (2011), Article 12. DOI: https://doi.org/10.1145/1985347.1985353

Robert J. Moore, Sungeun An, and Guang-Jie Ren. 2022. The IBM natural conversation framework: A new paradigm for conversational UX design. *Human–Computer Interaction* 38 (2022), 168–193. DOI: https://doi.org/10.1080/07370024.2022.2081571

Camille Morvan, Alexander O'Connor. 2017. An Analysis of Leon Festinger's a Theory of Cognitive Dissonance. Macat Library. DOI: https://doi.org/10.4324/9781912282432

Nika Mozafari, Welf H. Weiger, and Maik Hammerschmidt. 2021. Trust me, I'm a bot – Repercussions of chatbot disclosure in different service frontline settings. *Journal of Service Management* 33, 2 (2021), 221–245. DOI: https://doi.org/10.1108/JOSM-10-2020-0380

Bennet B. Murdock Jr. 1962. The serial position effect of free recall.*Journal of Experimental Psychology* 64, 5 (1962), 482–488. DOI: https://doi.org/10.1037/h0045106

C. Daniel Myers, and Dustin Tingley. 2016. The influence of emotion on trust. *Political Analysis* 24, 4 (2016), 492–500. DOI: https://doi.org/10.1093/pan/mpw026

NENT. 2019. Guidelines for research ethics in science and technology. The Norwegian National Committee for Research Ethics in Science and Technology. Retrieved from https://www.forskningsetikk.no/en/guidelines/science-and-technology/guidelines-for-research-ethics-in-science-and-technology/

NESH. 2022. Guidelines for Research Ethics in the Social Sciences and the Humanities. The National Committee for Research Ethics in the Social Sciences and the Humanities. Retrieved from https://www.forskningsetikk.no/en/guidelines/social-sciences-humanities-law-and-theology/guidelines-for-research-ethics-in-the-social-sciences-humanities-law-and-theology/

Cecilie B. Nordheim, Asbjørn Følstad, Cato A. Bjørkli. 2019. An initial model of trust in chatbots for customer service—Findings from a questionnaire study. *Interacting with Computers* 31, 3 (2019), 317–335. DOI: https://doi.org/10.1093/iwc/iwz022

Asbjørn Følstad, Cameron Taylor. 2019. Conversational repair in chatbots for customer service: the effect of expressing uncertainty and suggesting alternatives. In *Proceedings of CONVERSATIONS 2019*. Springer, 201–214. DOI: https://doi.org/10.1007/978-3-030-39540-7_14

Charles E. Osgood. 1962. Studies on the generality of affective meaning systems.*American Psychologist* 17, 1 (1962), 10–28. Retrieved from https://psycnet.apa.org/doi/10.1037/h0045146

Amy Ostrom and Dawn Iacobucci. 1995. Consumer trade-offs and the evaluation of services. *Journal of Marketing* 59, 1 (1995), 17–28. DOI: https://doi.org/10.1177/002224299505900102

Robert Plutchik. 2001. The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American Scientist* 89, 4 (2001), 344–350.

Aleksandra Przegalinska, Leon Ciechanowski, Anna Stroz, Peter Gloor, and Grzegorz Mazurek. 2019. In bot we trust: A new methodology of chatbot performance measures. *Business Horizons* 62, 6 (2019), 785–797. DOI: https://doi.org/10.1016/j.bushor.2019.08.005

Lova Rajaobelina, Sandrine Prom Tep, Manon Arcand, and Line Ricard. 2021. Creepiness: Its antecedents and impact on loyalty when interacting with a chatbot. *Psychology & Marketing* 38, 12 (2021), 2339–2356. DOI: https://doi.org/10.1002/mar.21548

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: https://doi.org/10.1016/j.ijhcs.2021.102630

Alessandra Rossi, Kerstin Dautenhahn, Kheng Lee Koay, and Michael L. Walters. 2020. How social robots influence people's trust in critical situations. In *Proceedings of the 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN '20)*. IEEE, 1020–1025. DOI: https://doi.org/10.1109/RO-MAN47096.2020.9223471

Denise M. Rousseau, Sim B. Sitkin, Ronald S. Burt, and Colin Camerer. 1998. Not so different after all: A cross-discipline view of trust. *Academy of Management Review* 23, 3 (1998), 393–404.

James A. Russell. 1978. Evidence of convergent validity on the dimensions of affect. *Journal of Personality and Social Psychology* 36, 10 (1978), 1152. Retrieved from https://psycnet.apa.org/doi/10.1037/0022-3514.36.10.1152

James A. Russell and Lisa Feldman Barrett. 1999. Core affect, prototypical emotional episodes, and other things called emotion: dissecting the elephant. *Journal of Personality and Social Psychology* 76, 5 (1999), 805–819. DOI: https://doi.org/10.1037//0022-3514.76.5.805

James A. Russell and Albert Mehrabian. 1977. Evidence for a three-factor theory of emotions. *Journal of Research in Personality* 11, 3 (1977), 273–294. DOI: https://doi.org/10.1016/0092-6566(77)90037-X

Clara Sainz-de-Baranda Andujar, Laura Gutiérrez-Martín, José Ángel Miranda-Calero, Marian Blanco-Ruiz, and Celia López-Ongil. 2022. Gender biases in the training methods of affective computing: Redesign and validation of the Self-Assessment Manikin in measuring emotions via audiovisual clips.*Frontiers in Psychology* 13 (2022), 955530. DOI: https://doi.org/10.3389/fpsyg.2022.955530

Sean Sands, Carla Ferraro, Colin Campbell, and Hsiu-Yuan Tsao. 2021. Managing the human–chatbot divide: how service scripts influence service experience. *Journal of Service Management* 32, 2 (2021), 246–264. DOI: https://doi.org/10.1108/JOSM-06-2019-0203

Emanuel A. Schegloff. 1987. Some sources of misunderstanding in talk-in-interaction. *Linguistics* 25 (1987), 201–218. DOI: https://doi.org/10.1515/ling.1987.25.1.201

Emanuel A. Schegloff, Gail Jefferson, and Harvey Sacks. 1977. The preference for self-correction in the organization of repair in conversation. *Language* 53, 2 (1977), 361–382. DOI: https://doi.org/10.2307/413107

Klaus R. Scherer. 2005. What are emotions? And how can they be measured?. *Social Science Information* 44, 4 (2005), 695–729. DOI: https://doi.org/10.1177/0539018405058216

Craig A. Smith and Phoebe C. Ellsworth. 1985. Patterns of cognitive appraisal in emotion. *Journal of Personality and Social Psychology* 48, 4 (1985), 813–838. Retrieved from https://psycnet.apa.org/doi/10.1037/0022-3514.48.4.813

Andreas Sonderegger, Klaus Heyden, Alain Chavaillaz, and Juergen Sauer. 2016. AniSAM & AniAvatar: Animated visualizations of affective states. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, 4828–4837. DOI: https://doi.org/10.1145/2858036.2858365

Statista. 2022. Consumer Satisfaction with Chatbot Customer Service in the United States as of June 2022. Statistics Brief. Statista. Retrieved from https://www.statista.com/statistics/657148/united-states-consumer-satisfaction-with-chatbot-service/

Sruti Subramanian, Katrien De Moor, Markus Fiedler, Kamil Koniuch, and Lucjan Janowski. 2023. Towards enhancing ecological validity in user studies: a systematic review of guidelines and implications for QoE research. *Quality and User Experience* 8, 1 (2023), 1–17. DOI: https://doi.org/10.1007/s41233-023-00059-2

Mark Taylor, Anne-Laure Thieullent, Simon Bachelet, Gagandeep Gadri, Scott Turton, Luca Cito, Steffen Elsasser, Darshan Shankavaram, Sudhir Rokade, Gagandeep Gadri, Scott Turton, Allan Frank, Shannon Warner, and Partha Panda. 2019. *Smart Talk: How Organizations and Consumers are Embracing Voice and Chat Assistants*. Technical Report. Capgemini. Retrieved from https://www.capgemini.com/wp-content/uploads/2019/09/Report_Conversational-Interfaces-1.pdf

Margot J. van der Goot, Laura Hafkamp, and Zoë Dankfort. 2020. Customer service chatbots: A qualitative interview study into the communication journey of customers. In *Proceedings of the CONVERSATIONS 2020 International Workshop on Chatbot Research and Design*. Springer, 190–204. DOI: https://doi.org/10.1007/978-3-030-68288-0_13

Cuicui Wang, Yiyang Li, Weizhong Fu, and Jia Jin. 2023. Whether to trust chatbots: Applying the event-related approach to understand consumers' emotional experiences in interactions with chatbots in e-commerce. *Journal of Retailing and Consumer Services* 73 (2023), 103325. DOI: https://doi.org/10.1016/j.jretconser.2023.103325

David Watson, Lee Anna Clark, and Auke Tellegen. 1988. Development and validation of brief measures of positive and negative affect: The PANAS scales.*Journal of Personality and Social Psychology* 54, 6 (1988), 1063–1070. DOI: https://doi.org/10.1037//0022-3514.54.6.1063

Cynthia Webster and D. S. Sundaram. 1998. Service consumption criticality in failure recovery. *Journal of Business Research* 41, 2 (1998), 153–159. DOI: https://doi.org/10.1016/S0148-2963(97)00004-0

Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, Courtney Biles, Sasha Brown, Zac Kenton, Will Hawkins, Tom Stepleton, Abeba Birhane, Lisa Anne Hendricks, Laura Rimell, William Isaac, Julia Haas, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2022. Taxonomy of risks posed by language models. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*. ACM, New York, NY, 214–229. DOI: https://doi.org/10.1145/3531146.3533088

Joel Wester, Tim Schrills, Henning Pohl, and Niels van Berkel. 2024. "As an AI language model, I cannot": Investigating LLM Denials of User Requests. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. ACM, New York, NY, Article 979. DOI: https://doi.org/10.1145/3613904.3642135

Douglas A. Wiegmann, Aaron Rich, and Hui Zhang. 2001. Automated diagnostic aids: The effects of aid reliability on users' trust and reliance. *Theoretical Issues in Ergonomics Science* 2, 4 (2001), 352–367. DOI: https://doi.org/10.1080/14639220110110306

Magdalena Wischnewski, Nicole Krämer, and Emmanuel Müller. 2023. Measuring and understanding trust calibrations for automated systems: A survey of the state-of-the-art and future directions. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '23)*. ACM, New York, NY, Article 755. DOI: https://doi.org/10.1145/3544548.3581197

Jie Xu and Enid Montague. 2015. Affect and trust in technology in teams: The effect of incidental affect and integral affect. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. SAGE Publications, 205–209. DOI: https://doi.org/10.1177/1541931215591042

Anbang Xu, Zhe Liu, Yufan Guo, Vibha Sinha, Rama Akkiraju. 2017. A new chatbot for customer service on social media. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems – CHI '17*. ACM, 3506–3510. DOI: https://doi.org/10.1145/3025453.30254

Juliana J. Y. Zhang, Asbjørn Følstad, and Cato A. Bjørkli. 2023. Organizational factors affecting successful implementation of chatbots for customer service. *Journal of Internet Commerce* 22, 1 (2023), 122–156. DOI: https://doi.org/10.1080/15332861.2021.1966723

Appendices

A List of Tasks

Task	Task description	Task	Task description
T1	Find out the terms and conditions for a savings agreement	T31	Reach out because you are unable to pay off your loan
T2	Find out your options to get insurance for your boat	T32	Stop your annuity
T3	Reach out because your card was taken by an ATM (cash machine)	T33	Reach out because you suspect your credit card has been used for fraud
T4	Find out how long it takes to make a money transfer to someone living in a foreign country	T34	Find out how much it costs to use your credit card abroad
T5	Find out whether contactless payments are safe	T35	Find out whether you can receive bank statements by email
T6	Reach out because you've forgotten your online bank password and need to login	T36	Find out what the repayment period is for a personal loan
T7	Find out the age requirement for children's cards	T37	Close your savings agreement fund
T8	Fix a problem with an invoice	T38	Find out where you can find your credit card invoices
T9	Find out the cost of a savings account for your child	T39	Find out where you can check if your payment was sent
T10	Find out if there is a tax advantage in having a shared savings account	T40	Find out the tax on stock dividends
T11	Check an unknown transaction in your own bank account	T41	Make a payment into your pension fund
T12	Find out how much it costs to change bank	T42	Find out whether you can use your smartwatch to login to your online bank
T13	Find the closest ATM (cash machine)	T43	Schedule a meeting with your advisor
T14	Find out whether the bank offers loans for solar panels	T44	Reach out to get support in buying a house for the first time
T15	Find out how to schedule automatic payments from your account	T45	Insure your car for travel abroad
T16	Find out whether children can use an online bank	T46	Find out how to set up an equity fund agreement
T17	Find out the terms and conditions for an annuity	T47	Find out the advantages and disadvantages of opening a shared savings account
T18	Get a price indication for a specific type of account	T48	Find out whether you can pay your bills during the weekend
T19	Cancel a double payment you made by mistake	T49	Move a loan from a different bank
T20	Open a bank account for your child	T50	Block your debit card
T21	Find out the ATM (cash machine) withdrawal limit in a foreign country	T51	Find the cost of a standard home insurance package
T22	Get a new card because your current one expires soon	T52	Find out how many stock savings accounts you are allowed to have
T23	Get a card for the person you're sharing your account with	T53	Find out what happens to your annuity payments in case you die
T24	Insure your pet	T54	Find out how to repay your credit card debt
T25	Find out how to login to your online bank	T55	Reach out because you've accidentally paid the wrong person
T26	Reach out because you're getting a 404 error when you try to login to your online bank	T56	Find out your maximum mortgage limit
T27	Find out the annual cost for children's cards	T57	Find out the effective and nominal interest rates for loans
T28	Find out how to buy stocks for the first time	T58	Find out what a disability insurance entails
T29	Find out whether you can pay using a chequebook	T59	Fix an error on your mortgage invoice
T30	Receive advice about saving money	T60	Find out the interest rate for green car loans

B Pre-Interaction Variables

Table B1 shows the descriptive statistics of the five pre-interaction variables—`complexity`, `personalisation`, `risk`, `trust` and `useChatbot`. Table B2 shows the inferential statistics for pairwise comparisons among the three tasks of HC. There were significant differences for the attribute Complexity and Personalisation between Task A and B and between Task A and C. In contrast, as shown in Table B3, for Task X, Y and Z, only a few significant differences. For brevity's sake, only p values are presented, leaving out other statistics.

Table B1. Means (SD) of Pre-Interaction Ratings of the Three Attributes for the Three Tasks in Each of the Eight Experimental Conditions

		HC				LC			
	Task	**Group 1 (bd0)**	**Group 2 (bd1)**	**Group 3 (bd2)**	**Group 4 (bd3)**	**Group 5 (bd0)**	**Group 6 (bd1)**	**Group 7 (bd2)**	**Group 8 (bd3)**
Complex	A/X	2.77(1.31)	3.00(1.3)	3.03 (1.3)	3.36(1.84)	3.00(1.37)	2.91(1.63)	2.88(1.36)	3.12(1.45)
	B/Y	4.45(1.55)	3.76(1.56)	4.13(1.36)	4.17(1.8)	3.4(1.25)	3.63(1.77)	3.16(1.39)	3.7(1.72)
	C/Z	3.84(1.68)	4.06(1.64)	4.37(1.54)	4.39(1.52)	3.3(1.54)	3.25(1.5)	2.91(1.42)	3.64(1.45)
Personal	A/X	3.9(1.96)	3.33(1.87)	4.1(1.67)	3.5(1.98)	3.73(1.76)	4.59(1.98)	4.13(2.06)	3.61(2.03)
	B/Y	5.48(1.59)	4.85(1.86)	5.27(1.26)	4.75(1.96)	3.33(1.65)	3.22(1.6)	4.13(1.95)	4.06(2.08)
	C/Z	4.97(1.68)	5.07(1.72)	4.83(1.72)	3.00(1.44)	3.03(1.62)	3.56(1.95)	3.64(1.73)	4.16(1.97)
Risk	A/X	4.16(1.97)	4.53(1.74)	4.67(1.77)	3.2(1.96)	3.03(1.81)	3.78(2.12)	3.45(2.17)	5.13(1.77)
	B/Y	5.13(1.77)	5.4(1.48)	5.69(1.7)	3.43(1.57)	3.19(1.69)	4.22(1.72)	3.36(1.88)	4.06(1.86)
	C/Z	4.06(1.86)	4.43(1.68)	4.64(1.93)	3.23(1.41)	2.78(1.34)	3.28(1.61)	3.67(1.61)	4.94(1.67)
Trust	A/X	4.94(1.67)	5.30(1.53)	4.97(1.63)	5.0(1.85)	4.33(1.88)	4.56(1.93)	5.31(1.53)	5.27(1.89)
	B/Y	5.84(1.24)	5.94(1.17)	6.03(0.93)	5.64(1.55)	4.73(1.72)	4.91(1.69)	5.66(1.38)	5.21(1.78)
	C/Z	5.16(1.77)	5.36(1.52)	5.47(1.36)	5.14(1.71)	4.1(1.77)	4.22(1.91)	5.16(1.63)	5.03(1.72)
UseChatbot	A/X	4.23(1.84)	4.55(1.89)	4.2(2.33)	4.08(2.05)	4.57(1.83)	4.06(2.15)	4.81(1.98)	4.58(2.03)
	B/Y	2.87(1.65)	3.7(1.98)	2.8(1.83)	2.86(1.93)	4.33(1.92)	4.34(2.15)	4.25(1.69)	4.48(1.96)
	C/Z	3.48(1.69)	4.39(1.78)	3.2(1.65)	2.92(1.65)	4.93(1.76)	5.0(1.88)	4.91(1.75)	4.3(1.94)

Table B2. Results of within-Group Pairwise Comparisons for the Attributes of the Three HC Tasks (A, B and C)

		High Criticality			
	Task	**Group 1**	**Group 2**	**Group 3**	**Group 4**
Complex	A-B	$p < .001$	n.s.	$p = .001$	n.s.
	A-C	$p = .007$	$p = .005$	$p < .001$	$p = .006$
	B-C	n.s.	n.s.	n.s.	n.s.
Personal	A-B	$p < .001$	$p < .001$	$p = .002$	$p < .001$
	A-C	n.s.	$p = .006$	$p = .04$	$p = .009$
	B-C	n.s.	n.s.	n.s.	n.s.
Risk	A-B	n.s.	n.s.	$p = .029$	$p = .003$
	A-C	$p = .05$	n.s.	n.s.	n.s.
	B-C	$p = .022$	n.s.	$p = .032$	$p = .015$
Trust	A-B	$p = .005$	n.s.	$p = .002$	n.s.
	A-C	n.s.	n.s.	n.s.	n.s.
	B-C	n.s.	n.s.	n.s.	n.s.
UseChatbot	A-B	$p = .002$	n.s.	$p = .006$	$p = .005$
	A-C	n.s	n.s.	n.s.	$p = .024$
	B-C	n.s	n.s.	n.s.	n.s.

Table B3. Results of within-Group Pairwise Comparisons for the Attributes of the Three LC Tasks (X, Y and Z)

		Low Criticality			
	Task	**Group 5**	**Group 6**	**Group 7**	**Group 8**
Complex	X-Y	n.s.	n.s.	n.s.	n.s.
	X-Z	n.s.	n.s.	n.s.	n.s.
	Y-Z	n.s.	n.s.	n.s.	n.s.
Personal	X-Y	n.s.	p = .016	n.s.	n.s.
	X-Z	n.s.	p = .004	n.s.	n.s.
	Y-Z	n.s.	n.s.	n.s.	n.s.
Risk	X-Y	n.s.	n.s.	n.s.	n.s.
	X-Z	n.s.	n.s.	n.s.	n.s.
	Y-Z	n.s.	n.s.	p = .01	n.s.
Trust	X-Y	n.s.	n.s.	n.s.	n.s.
	X-Z	n.s.	n.s.	n.s.	n.s.
	Y-Z	n.s.	n.s.	n.s.	n.s.
UseChatbot	X-Y	n.s.	n.s.	n.s.	n.s.
	X-Z	n.s.	p = .045	n.s.	n.s.
	Y-Z	n.s.	n.s.	n.s.	n.s.

Table B4. One-Way ANOVA among the Eight Experimental Conditions for the Five Pre-Interaction Task-Related Variables

Variable	**F(7,249)**	**p**	**η2**
Complexity	3.03	.004	0.08
Personalisation	5.85	<.001	0.14
Risk	10.31	<.001	0.23
Trust	3.76	<.001	0.10
UseChatbot	5.59	<.001	0.14

Table B5. One-Way ANOVA between HC and LC Groups for the Five Pre-Interaction Task-Related Variables

Variable	**F(1,255)**	**p**	**η2**
Complexity	3.03	.004	0.06
Personalisation	5.85	<.001	0.11
Risk	10.31	<.001	0.20
Trust	3.76	=.001	0.04
UseChatbot	5.59	<.001	0.10

Tables B4 and B5 show the results of one-way ANOVAs to investigate differences between the eight experimental conditions (Table B4) and the high and low-criticality groups (Table B5) for the pre-interaction task-related variables.

Table B6 provides an overview of scores for pre-interaction emotion measures in the high and low-criticality groups.

Table B6. Mean (SD) of the Pre-Interaction Emotion Measures (Baseline) for HC and LC Groups

	HC	LC
Valence	6.62(1.71)	6.61(1.64)
Activation	4.06(2.25)	4.17(2.23)
Control	8.38(1.85)	6.6(1.77)

C Participant Expressions of Emotion after Low-criticality Tasks

In this appendix, the Figures C1–C3 correspond to those for HC in Section 5.3.5, illustrating the words participants used to express their emotions after completing the respective tasks. Similar patterns can be observed between the two groups of Task Criticality. Note that only words that are common to at least two of the tasks under comparison are listed. One consistent observation is that the variety of words is higher in the case of the third task (Z3) than the other tasks. It could be that participants had experienced different emotions after exposure to different cases.

	annoyed	irritated	disappointed	frustrated	confused	happy	calm	neutral
Bd Task1	5	2	4	4	2	2	4	2
Bd Task2	10	4	1	5	0	0	0	1
Bd Task3	5	2	4	4	2	2	4	2

Fig. C1. Word-clouds and word-counts for tasks of LC with breakdown (Bd) at Task 1 (Left), Task 2 (Middle), and Task 3 (Right).

	happy	calm	content	relaxed	pleased	satisfied
Before (base)	6	5	8	6	0	0
After (task 2)	4	2	3	3	2	6
After (task 3)	5	4	4	4	2	2

Fig. C2. For the breakdown BdTask1 of LC, word-clouds and word-counts for the emotions measured before (Left: Baseline), and after (Middle: Task 2 with no breakdown and Left: Task 3 with no breakdown).

	content	calm	relaxed	chilled	happy	interested	agitated
Before (task 1)	8	3	3	2	2	2	1
After (task 3)	4	4	3	1	6	1	1

Fig. C3. For the breakdown BdTask2 of LC, word-clouds and word-counts for the emotions measured before (Left: Task 1 no breakdown) and after (Right: Task 3 no breakdown).

Received 22 September 2023; revised 30 July 2024; accepted 31 July 2024