

# *ContextBot*: Improving Response Consistency in Crowd-Powered Conversational Systems

---



Yao Ma



---

# *ContextBot*: Improving Response Consistency in Crowd-Powered Conversational Systems

---

THESIS

submitted in partial fulfillment of the  
requirements for the degree of

MASTER OF SCIENCE

in

COMPUTER SCIENCE

by

Yao Ma

born in Xi'an, ShaanXi Province, China



Web Information Systems Research Group  
Department of Computer Science  
Faculty EEMCS, Delft University of Technology  
Delft, the Netherlands  
<https://www.wis.ewi.tudelft.nl/>



---

# *ContextBot*: Improving Response Consistency in Crowd-Powered Conversational Systems

---

Author: Yao Ma  
Student id: 5222982  
Email: Y.MA-11@student.tudelft.nl

## Abstract

Crowd-powered conversational systems (CPCS) solicit the wisdom of crowds to quickly respond to on-demand users' needs. The very factors that make this a viable solution—such as the availability of diverse crowd workers on-demand—also lead to great challenges. The ever-changing pool of online workers powering conversations with individual users makes it particularly difficult to generate contextually consistent responses from a single user's standpoint. To tackle this, prior work has employed conversational facts extracted by workers to maintain a global memory, albeit with limited success. Leveraging systematic context in affective crowdsourcing tasks has remained unexplored. Through a controlled experiment, we explored if a conversational agent, dubbed *ContextBot*, can provide workers with the required context on the fly for successful completion of affective support tasks in CPCS, and explore the impact of *ContextBot* on the response quality of workers and their interaction experience. To this end, we recruited workers ( $N=351$ ) from the Prolific crowdsourcing platform and carried out a  $3 \times 3$  factorial between-subjects study. Experimental conditions varied based on (i) whether or not context was elicited and informed by *motivational interviewing* techniques (MI, non-MI, and chat history), and (ii) different conversational entry points for workers to produce responses (early, middle, and late). Our findings showed that workers who entered the conversation earliest were more likely to produce highly consistent responses after interacting with *ContextBot*. Better user experience from workers was expected after they interacted with *ContextBot* at a late entry. We found that interacting with *ContextBot* through task completion did not negatively impact workers' cognitive load.

---

Thesis Committee:

Chair: Prof. Dr. Ir. G.J.P.M. Houben, EEMCS Faculty, TU Delft  
University Supervisor: Dr. Ir. U. Gadiraju, EEMCS Faculty, TU Delft  
Daily Supervisor: Dr. T. Abbas, EEMCS Faculty, TU Delft  
Committee Member: Dr. M.L. Tielman, EEMCS Faculty, TU Delft

---

# Preface

This thesis project ends my two-year study for master of computer science in TU Delft. The eight-month reading and writing training improved my ability of conducting academic research. Most importantly, I was able to devote my enthusiasm into the academic work with a lot of discretion and logic thinking. I am glad that I found my enthusiasm in the field of human computer interaction, where the understanding of the complexity of human responses towards machine intelligence fascinated me.

I would like to give my sincerest thanks to my supervisor, Ujwal Gadiraju, who arranged weekly meetings to discuss the progress of the thesis and guided me to think like a real researcher. His encouragement and supervision helped me complete the research independently and confidently on time. In the meantime, I would also like to thank my daily supervisor, Tahir Abbas, who provided me with the most careful and patient guidance, from reading recommendation to writing feedback. Without his insightful and meticulous inputs, my research would not have been completed with sufficient quality. Finally, I would like to thank two committee members, Geert-Jan Houben and Myrthe Tielman, for taking time to discuss my work and join the thesis defense.

During this eight months of my paper writing, the world was experiencing another wave of COVID-19. In such case, the company of my friends was an indispensable motivator for me to keep working. Therefore, I would like to give the greatest thanks to those friends who helped me at those tough moments: Shreyan, Annam, Zhifan, Rebecca, Caroline, and Ojas. Their laughter and companionship lit up my winter and summer. At the same time, I also would like to thank families who were living far abroad. Their support and care were always my strongest backing.

In the end, I would like to thank myself for the courage and persistence that I did not give up even in the most difficult times in 2021. And I am grateful for myself that I have discovered the joy and enthusiasm of doing academic research. I'd like to tell myself, ahead this, a long way to go, and one day, the potential would be fully unleashed.

Yao Ma  
Budapest, Hungary  
August 10, 2022





---

# Contents

<b>Preface</b>	<b>iii</b>
<b>Contents</b>	<b>v</b>
<b>List of Figures</b>	<b>vii</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Problem Definition and Goal . . . . .	3
1.2 Research Questions . . . . .	4
1.3 Contributions . . . . .	5
1.4 Thesis Outline . . . . .	6
<b>2 Related Work</b>	<b>7</b>
2.1 Crowd-Powered Conversational Systems . . . . .	7
2.2 Context in Conversations . . . . .	11
2.3 Motivational Interviewing . . . . .	12
2.4 Conversational Interfaces in Crowdsourcing . . . . .	13
2.5 Summary . . . . .	13
<b>3 System Design and Implementation</b>	<b>15</b>
3.1 System Architecture . . . . .	15
3.2 Design of <i>ContextBot</i> . . . . .	15
3.3 System Implementation . . . . .	18
3.4 Summary . . . . .	21
<b>4 Evaluating ContextBot</b>	<b>23</b>
4.1 Goal and Hypotheses . . . . .	23
4.2 Study Design . . . . .	24
4.3 Participants . . . . .	27
4.4 Procedure . . . . .	27
4.5 Evaluation Metrics . . . . .	29

<b>5</b>	<b>Results</b>	<b>31</b>
5.1	Descriptive Statistics . . . . .	31
5.2	Hypothesis Tests . . . . .	33
5.3	Qualitative Analysis . . . . .	39
5.4	Exploratory Findings . . . . .	41
<b>6</b>	<b>Discussions</b>	<b>43</b>
6.1	Interpretation and Implications . . . . .	43
6.2	Limitations . . . . .	44
<b>7</b>	<b>Conclusions and Future Work</b>	<b>47</b>
7.1	Conclusions . . . . .	47
7.2	Future Work . . . . .	49
	<b>Bibliography</b>	<b>51</b>
<b>A</b>	<b>System Implementation Elements</b>	<b>63</b>
A.1	Entity Relation Diagram . . . . .	63
A.2	Dialog Data . . . . .	63
<b>B</b>	<b>Consent Form and Questionnaires</b>	<b>69</b>
B.1	Consent Form: MI Condition . . . . .	69
B.2	Consent Form: Non-MI Condition . . . . .	70
B.3	Consent Form: History Condition . . . . .	71
B.4	Post-Task Questionnaire Items . . . . .	72
B.5	Evaluation Questionnaire for Response Consistency . . . . .	73
B.6	Evaluation Questionnaire for Professionalism in Responses . . . . .	73

---

# List of Figures

1.1	Conversation flow between an end-user and crowd workers in CPCS. New workers constantly enter and exit the conversation. Their entry points (early, middle, and late) directly affect the amount of historical contexts that workers need to understand before coherently and effectively responding to the end-user.	2
1.2	Thesis Outline.	5
3.1	Overview of the system architecture.	16
3.2	Conversation flow of ContextBot.	18
3.3	Worker interface of the system. The interface to interact with <i>ContextBot</i> is placed next to the main task window where crowd workers can read the chat and respond to the user.	19
4.1	Overview of the task procedure.	27
4.2	The mood scale to select in the pre-task questionnaire.	28
5.1	The number of workers having different mood states.	33
5.2	Response consistency scores across interacted types and entry points by combining the MI and non-MI conditions. * = statistically different (interacted vs. not interacted vs. history).	36
5.3	Boxplots of UEQ-S scores (Fig (a), (b), * = statistically significant between interacted group and not interacted group) and NASA-TLX scores (Fig (c), (d), * = statistically significant among early, middle, and late groups in the History condition in Fig (c)) in terms of the interacted types across different conditions. Black points indicate the average value of this group.	39

## LIST OF FIGURES

---

5.4	Relations between context types and the satisfaction score. <i>bot_icon</i> is the group who only clicked the chatbot icon without further interactions with <i>ContextBot</i> ; <i>social</i> is the group who only interacted with social context; <i>ling</i> is the group who followed the interaction from the start to linguistic context; <i>seman</i> is the group who followed the interaction from the start to semantic context; <i>cogn</i> is the group who followed the interaction from the start to cognitive context; <i>all</i> is the group who finished the whole interaction with <i>ContextBot</i> . The average and median values of corresponding groups are indicated by black points and red lines, respectively. . . . .	40
5.5	Relations between context types and relevant variables (samples are selected from those who had interactive behaviour with <i>ContextBot</i> ). The average and median values of corresponding groups are indicated by black points and red lines, respectively. . . . .	41
A.1	Entity relation diagram for the database. “PK” means the primary key in the table while “FK” means the foreign key in the table. . . . .	63

# Chapter 1

---

## Introduction

Crowd-powered conversational systems (CPCS) leverage real-time human computation, allowing synchronized workers to collaboratively help the user with crowdsourcing tasks through conversations [47, 32]. An interactive two-way conversation with multiple workers who act as a single operator enables the user to receive more personalized and diverse assistance than traditional dialogue systems. Chorus is a text-based conversational agent where synchronized workers participate in the response generation and voting, assisting end-users with information retrieval tasks [47]. Evorus builds on Chorus by adding an automated module to select high-quality responses from workers [34]. Despite these filtering mechanisms to control responses, empirical results of Chorus on a small scale have revealed the potential challenge of maintaining response consistency across constantly changing workers [33]. The pool of online workers in current CPCS requires on-demand recruitment, which makes it intrinsically difficult for new workers to quickly understand all historical contexts in CPCS. Meanwhile, the time spent by workers on understanding the context can lead to delays in responses. It remains challenging to maintain the trade-off between response quality and latency. Figure 1.1 illustrates the conversation flow between an end-user and crowd workers in CPCS.

Previous research has adopted self-help approaches to maintain global worker memory by allowing workers to record and track context collaboratively. In addition to providing the chat history, a “fact board” with facts of current conversations selected or summarized by workers is updated [47]. However, this approach either increases the task burden of current workers beyond replying to users or the costs of recruiting additional workers to collect context. The memory curated by previous or other parallel workers mostly includes subjective assumptions about the important information needed in the current dialogue turn, which inhibits new workers from following the context in a systematic way.

Another concern is that for affective crowdsourcing where the emotional intelligence of crowds is applied to peer-to-peer mental health support [61], workers are involved in counselling conversations that demand higher global comprehension than information retrieval tasks. Although the idea of embedding paid crowd workers in an interactive user interface to deliver emotional support on demand is promising, existing crowd-powered systems about affective support focus more on the effectiveness of counselling strategies [62] and the availability of real-time recruitment [1], thus ignoring that inconsistent worker responses

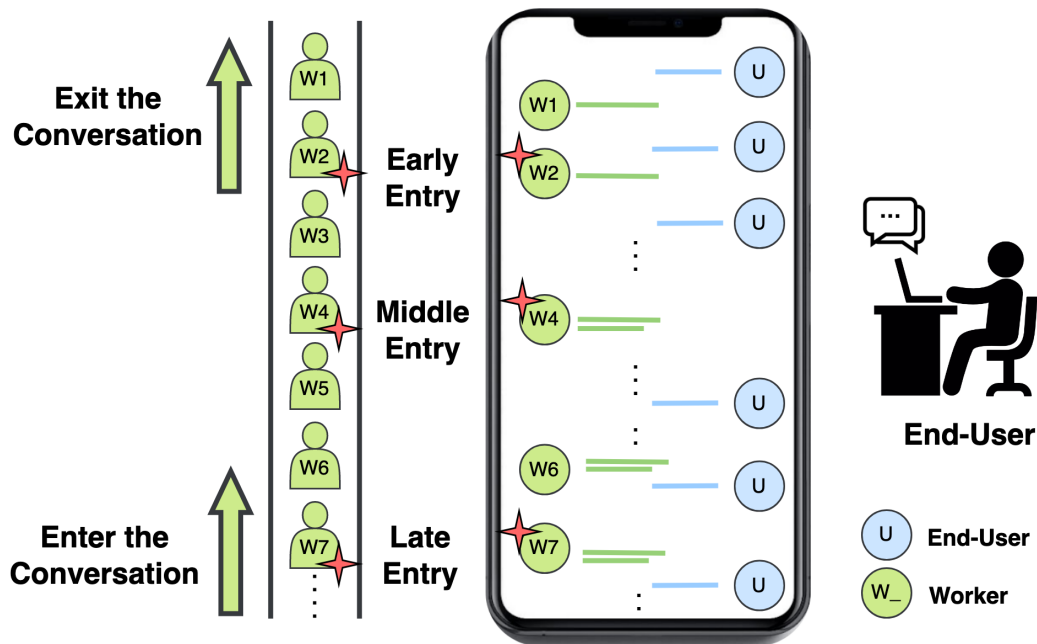


Figure 1.1: Conversation flow between an end-user and crowd workers in CPCS. New workers constantly enter and exit the conversation. Their entry points (early, middle, and late) directly affect the amount of historical contexts that workers need to understand before coherently and effectively responding to the end-user.

may fail to relate to users' feelings and problems mentioned in the chat history. This can be particularly harmful in online psychological interventions.

Meanwhile, the past decade has witnessed the development of human-computer interaction mediated by conversational interfaces for accessing information [99] and enhancing spoken dialogue. Driven by a chatbot, the conversational interface has been viewed as an effective tool that can be used to guide crowd workers to perform tasks in the workflow. As an alternative to traditional web interfaces, conversational interfaces increase the engagement of workers in crowdsourcing tasks without negatively impacting execution time and output qualities [54]. Researchers have also explored the efficacy of chatbots in training workers to complete therapeutic tasks [2], demonstrating the potential of conversational interfaces in taking on pragmatic roles. Therefore, we would like to design mechanisms to provide context using conversational interfaces and explore what effect can a conversational interface have on enhancing context consistency in crowd-powered chatbots. Inspired by [72], we developed a chatbot called *ContextBot* to help workers systematically track context about the current dialogue without recording it themselves.

## 1.1 Problem Definition and Goal

Our goal in this paper is to address response inconsistency in CPCS with a context-tracking tool mediated by a conversational interface. We employed an affective crowdsourcing task, which required workers to deliver emotional support by *Motivational Interviewing (MI)* [59]. To this end, workers used a client-centred conversation to stimulate users to change. To achieve this goal, we need to understand the challenges of leveraging a conversational interface to present context and guiding crowd workers to understand and use context while not burdening them.

First, context carries contextual information in the dialogue, and is a local configuration that both the speaker and the listener have a consensus on. We regard consistency as “the ability to generate non-conflicting responses and remember past information discussed in previous conversations” [48]. Inconsistent context may prevent the dialogue from proceeding more naturally and lead to the lack of important information. When we talk about contextual information, we hope not only to retain the content, task or structure information contained in the historical dialogue but also to understand the user characteristics reflected in the dialogue, such as the accurate inference of the user’s intentions and the resolution of ambiguity in associated attributes. Typically, in MI, the context of an ongoing conversation is comprised of the user’s disclosures to the therapist and the therapist’s inquiries. New workers who play the role of a therapy coach need to simultaneously grasp facts about the user, the therapist’s intentions, and other contextual factors. Even though the existing text-based dialogue systems have tried to present all of the chat histories [47] to new workers for browsing or combine worker-generated notes and aggregation methods [24], a systematic way of managing long-term conversational memory has not yet been studied. The challenge mainly lies in specifying a ‘step-by-step’ guideline to help crowd workers understand the perceived context systematically and answer the user’s request.

Another challenge is to reduce the negative impact of getting extra assistance about the dialogue context while workers respond to the user. CPCS are usually deployed in real-time environments, where the trade-off between response quality and response latency is an essential factor. In an affective crowdsourcing task, responses are expected to also convey a certain degree of counselling expertise, such as empathy, in addition to being contextually consistent with the chat history. Both parties in dialogue usually perform specific speech acts to pursue a dialogue goal, which could be potentially enhanced by providing workers with contextual guidance aligned with the consultation requirements. The dynamically changing worker pool also provides an opportunity to explore the behaviour of workers entering the dialogue at different points. A related challenge is to explore how the response quality and interaction experience are affected after interacting with the conversational interface for context under different contextual guidance and entry points.

In this paper, we investigate how to systematically provide context with a text-based conversational interface driven by a chatbot, *ContextBot*, to help new workers quickly grasp key information from the long chat history. The challenges needed to be addressed in order to achieve this goal are summarized as: (1) providing systematic context comprised of multi-dimensional information disclosed in an affective conversation; (2) exploring the impact of interacting with *ContextBot* on response quality and interaction experience from workers.

In the next section, we will refine the research questions to address these challenges.

### 1.2 Research Questions

Based on the background in the previous section, we would like to know how to use a text-based conversational interface, *ContextBot*, to provide context for crowd workers to complete affective support tasks in CPCS. We aim to apply the interface to affective support in online conversations. We set the focus of the paper to answer the main research question (RQ):

***RQ: How does ContextBot used for providing context in CPCS affect the response quality and interaction experience of workers?***

To answer the main question, two sub-questions are derived as follows:

***RQ1: How could the context be extracted and provided for crowd workers to understand the chat history?***

As mentioned earlier, the context involved in a counselling conversation not only includes the general context of persons, places, and events, but also includes multi-dimensional contexts such as the user's feelings, the therapist's intentions, and the goals of the therapeutic conversation. The context related to the user's current query should be decomposed into multiple factors, which should also be presented to workers in a logical and systematic way. The design of this part will address the challenges (1) by decomposing the context into multiple dimensions (2) and by mapping the conversational flow of the chatbot to provide the different dimensions of contexts.

***RQ2: How does interacting with ContextBot affect the workers' response quality (consistency and professionalism) and their interaction experience (user experience and cognitive load)?***

Driven by *RQ1*, contextual guidance aligned with the content of the current conversation could play an important role in shaping the consistency of responses in CPCS, for example, whether the guidance is adherent to MI-related techniques. To better simulate the crowdsourcing flow in real-time CPCS, we identify another contextual factor, entry points, to define the amount of the context for new workers to follow up when they enter the dialogue at an early, middle, or late entry. A between-subjects experiment will be expected to compare different conditions comprised of contextual guidance and entry points. Since our task is to provide affective support for the user, we are interested in whether *ContextBot* could also help improve the professionalism in responses. For the interaction experience, we identify two dependent variables, user experience and cognitive load to measure the effectiveness of *ContextBot*.



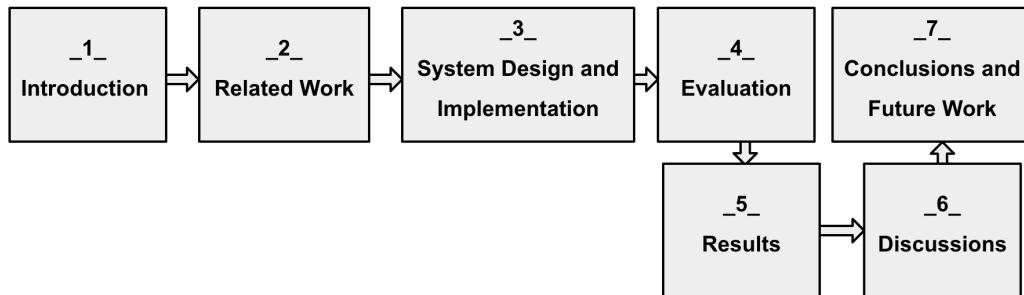


Figure 1.2: Thesis Outline.

### 1.3 Contributions

This project is the first attempt to apply a conversational interface as the context-tracking tool to facilitate an affective support task in CPCS. This presents several challenges from different fields, including contextual consistency, conversational interfaces, and affective crowdsourcing. Specifically, the contributions of this paper are as follows:

- We identified the challenges of presenting context to constantly changing crowd workers in CPCS. Given an affective support task of responding to the user by MI techniques, we leveraged a conversational interface, *ContextBot*, specialized for CPCS to show the possibility of dynamically supporting context in an ongoing counselling conversation.
- We conceptualized the way of providing the context of an MI-related counselling conversation in CPCS. Specifically, we decomposed context into four dimensions and organized a conversation flow to provide social context, linguistic context, semantic context, and cognitive context in sequence. The importance of this contribution is that it allows workers to follow the context in a systematic way, overcoming the weakness of new workers feeling overwhelmed by separate facts summarized by previous workers.
- We modified and adapted an MI-related counselling conversation to verify the effectiveness of contextual guidance aligned with counselling techniques in helping to improve response professionalism. Our findings demonstrated the complexity and necessity to introduce specialized context to affective support tasks using MI in CPCS.
- We conducted a between-subjects experiment to evaluate the response quality and interaction experience between crowd workers and *ContextBot* in different conditions of providing context. We defined two independent variables, contextual guidance and entry points, to comprehensively explore the benefits and risks of introducing a conversational interface to provide the context in CPCS from the perspectives of context content and real-time systems.

## **1.4 Thesis Outline**

This thesis consists of the following chapters, illustrated in Figure 1.2. In Chapter 2, we discuss the previous work related to crowd-powered conversational systems, the context in conversations, motivational interviewing, and conversational interfaces. Next, we introduce the design and implementation of our system in Chapter 3. In Chapter 4, the experimental design of how to evaluate our system is described, followed by the evaluation results in Chapter 5. Then, we further discuss the implications and limitations in Chapter 6. Finally, we conclude the thesis with conclusions and future work in Chapter 7.

## Chapter 2

---

# Related Work

In this chapter, we will first discuss the earlier research in the field of crowd-powered conversational systems in Section 2.1, where the real-time techniques, affective crowdsourcing, and approaches to maintain contextual consistency are often introduced to enrich relevant applications. Then, we will describe the concept of context in conversations from a linguistic perspective in Section 2.2. Next, we will include the practice of *motivational interviewing (MI)* in Section 2.3, followed by the related work about applications of conversational interfaces in crowdsourcing in Section 2.4. Finally, we will conclude how we build our system design upon the above research in Section 2.5.

### 2.1 Crowd-Powered Conversational Systems

Our research is closely related to solutions for providing context in crowd-powered conversational systems (CPCS), which allow workers to help users with online tasks through an interactive conversation.

The early development of chatbots was inspired by the Turing test in 1950 [87]. Recent studies have witnessed the rise of end-to-end architecture by employing sequence-to-sequence neural networks to help chatbots generate responses. Although automated solutions have been trying to simulate human-level conversations, a large number of intelligent systems are still facing huge challenges for long-term context tracking and random responses. In response to this challenge, hybrid systems have been introduced to integrate the power of crowds and the advantages of automated models to realize intelligible question answering [44, 71].

Besides integrating crowdsourcing into the automated modules, the “wisdom of crowds” can be solicited to organize collective intelligence in a more direct manner [62]. CPCS extend the prototyping concept of Wizard-of-Oz (WoZ) [53], with the ability to apply the wisdom of crowds to control different aspects of the dialogue system, forming an “assembly line” of dialogue system controllers [32]. Crowds are naturally adaptive to the dynamic dialogue flow and can organize utterances in a logical way. The coordination among distributed workers with different roles allows for more complex system control towards integrating automated modules. CPCS could overcome the difficulty of including social identities and

selecting random responses from only a limited database or corpora [95].

### 2.1.1 Real-Time Crowdsourcing for Chatbots

Real-time crowd-powered systems recruit workers on-demand and create an environment where crowd workers work synchronously to complete tasks. Early systems distribute batches of microtasks whose completion time ranges from hours to days. The real-time crowdsourcing (RTC) technique has been shown to reduce worker response time to seconds [8], such as VizWiz [10]. Legion [46] was first introduced as a continuous real-time crowdsourcing system to interact with users in a controlled way. Crowd workers were given the control to interact with a UI control task over ongoing tasks. Then the concept of interactive crowdsourcing was later applied to conversational interfaces [47, 35]. Interactive dialogue interfaces require real-time or near real-time practices to ensure smooth conversational sessions. Applications such as visual question answering [10], are still limited by the issue of latency. Current systems mostly focus on helping users with specific information needs and services. Chorus was firstly proposed and deployed in a real-world scenario to enable a group of synchronized workers to help users retrieve information in a conversational interface [47]. It assigned crowd workers to respond to users and vote on each other's responses. The optimal set of responses was then selected to the user. Evorus was built upon Chorus and further included a machine learning module to automatically select responses from workers [34], thereby reducing real-time latency and improving output quality. Guardian took task-oriented input from a web-based API and then assigned tasks to workers to interpret the parameters for generating responses [32]. InstructableCrowd [31] was also designed as a task-oriented agent to guide users to manage tasks through their mobile phones. The conversational interface provided a way for synchronized workers to interact with users and transform their needs to IF-THEN rules. Alternatives such as Facebook M used proficiently trained employees working in shifts to interact with users [39].

Despite the success of recruiting workers on-demand, the empirical results of Chorus on a small scale have revealed the potential challenges of maintaining worker consensus in real-time CPCS [33]. In addition to directly providing the previous chat history, a "Working Memory" assisted in tracking context by using facts of current conversations selected or summarized by previous workers [47]. However, this approach increased the additional task burden for workers beyond replying to users. We aim to address the challenge of enabling a better balance between context tracking in real-time settings and response latency from workers.

### 2.1.2 Chatbots in Mental Health

The growing need to address mental health issues has promoted the development of psychotherapy. However, due to the time and space limitations of traditional face-to-face communications with the therapist, chatbots of supportive purposes have been designed to provide users with more timely help. In the early development, Eliza was first designed in a text-based manner to resemble the behaviour of a Rogerian psychotherapist [91]. Parry then extended Eliza by simulating a person with paranoid schizophrenia, thus being the

first chatbot to introduce judgemental attitudes towards conversations in development [15]. Recent surveys of chatbots in health have systematically reviewed the use of chatbots in mental health and chronic diseases [45, 77, 60]. Chatbots can be effectively perceived as not just forced labour [94] but as companions who provide users with self-management and support. If necessary, common counselling techniques, such as cognitive behavioural therapy (CBT) [19, 52, 20, 36], motivational interviewing (MI) [50, 20], and self-compassion therapy [20], would be incorporated into the design of chatbots to achieve cognitive relief or some level of psychotherapy.

WoeBot was a fully automated mental health application designed with CBT [19]. The interactions between the user and the chatbot were marked by general questions about the user's emotions and context, aiming to assess and respond to the user's anxiety and depression levels. Results after a two-week trial have demonstrated that participants who used the service experienced less depressive symptoms than the control group. Similarly, the study with Shim found that users who adhered to communicating with the chatbot under the guidance of CBT reported reduced levels of stress and increased mental well-being after the trial [52]. Wysa adopted a wide range of counselling techniques to help depressed users [36]. The study revealed a positive relationship between user engagement and health improvement. Tess was adapted as an integrative psychological chatbot to provide customized therapies depending on the user's emotions and concerns [20]. It gained significant improvements over depression and anxiety among college students. Beyond textual chatbots, Philip et al. used virtual embodied conversational agents (ECA) to identify the user's symptoms of major depressive disorder based on specific diagnosis criteria [70]. Lucas et al. demonstrated that a virtual agent could better help elicit traumatic symptoms from users with a verbal semi-structured interview [51]. Studies in delivering psychotherapy through chatbots generally track and evaluate participants' mental health empirically, and have achieved improvements in terms of user experience, depression levels, self-disclosure, etc., which supports the notion that chatbots could be used as an alternative to providing moderate counselling for users who seek help at any time.

Yet, evidence has shown that a fully automated text-based chatbot based on CBT principles may provide repetitive, short and unnatural responses [19]. Also, it has been argued that parsing and reflecting on the client's stories described with unique idiosyncratic words is essential for creating effective psychotherapy [68]. The idea of incorporating human intelligence into supportive chatbots is expected to address the challenge of generating more nuanced responses and expressing better empathy for users who need more personalized emotional support.

Applications of affective crowdsourcing [61] solicit the collective emotional intelligence and implement the concept of peer-to-peer mental health support [66], which has been demonstrated in systems such as in-person and online support forums [5]. The web-based system, Panoply [62, 63], trained crowd workers to perform cognitive reconstructing and generate new appraisals for users who seek emotional help. KokoBot [64] further extended the idea and acted as a computer agent to guide users by predicting appropriate responses from an existing crowdsourced peer support corpus. However, such systems required users to log in to the system or get help from peers through agents [43]. Direct and continuous counselling interactions between users and chatbots remain blank. Crowd of Oz (CoZ), by

contrast, enabled the real-time two-way interaction between a stressed user and synchronous crowds by employing real-time, synchronous crowdsourcing (RTC) techniques [1]. However, previous studies of delivering affective support have focused more on the effectiveness and real-time features. So far, no attempts have been made to apply context-tracking tools to crowd-powered systems for working memory maintenance in real-time affective conversations. We built our study upon the work of [47, 33], addressing the challenge of uncovering important context from the lengthy chat history while not burdening workers in an affective support task.

### 2.1.3 Contextual Consistency in Chatbots

A human-like dialogue system is expected to deliver consistent responses in terms of the given persona, conversational styles, and context related to the previous chat history [30]. For persona consistency, automated dialogue systems model encode persona-specific statements [97, 49] as vectors to be fed into the network-based learning architecture, such as sequence-to-sequence models [83]. A novel two-stage framework was proposed to generate diverse and persona-consistent responses [80]. In addition to adopting additional dialogue attribute to encode the latent variables extracted from responses, it introduced a persona-consistency checking module to correct and rewrite the responses inconsistent with the given persona. Li et al. generated responses with a consistent persona conditioned on specific speakers' personalities such as gender and hobbies [49]. It represented each speaker with the user embedding and targeted the content consistency but could not handle unknown speakers. For stylistic consistency, a common strategy is to train stylistic parameters by taking advantage of transfer learning or domain adaptation. Zhang et al. decomposed the procedure of generating personalized responses as a two-phase approach, namely initialization and adaptation [96]. Wang et al. adapted a decoding algorithm with specific language styles and demonstrated the effectiveness of generating stylistically consistent responses based on an open-domain corpus [90].

However, contextual consistency with respect to the dialogue history is yet to be explored. Context is usually regarded as a sequence of past dialogue exchanges of any length and then encoded as vectors to be fed into the learning architecture. Despite the simplicity of embedding context as vectors from a computational perspective, current automated models focus more on modelling a single dimension of context. Sordoni et al. addressed the challenge of generating responses that are sensitive to the linguistic context [82]. They encoded the past utterances as contextual information to hidden continuous representations by conditional language models. However, they did not consider the word order within the message and context utterances. Sato et al. extended the linguistic context to conversational situations by including user profiles and timestamps of the utterances [76]. They proposed neural conversational models to inject given situations that could help generate situation-aware responses. Yet, the models cannot comprehensively link the current utterance with multi-dimensional contexts, such as the historical text, semantic facts, speaker intentions, and dialogue goals. Moreover, both participants in dialogue usually perform specific actions to pursue a communicative goal while existing research in automated dialogue systems ignores the contextual information that drives the goal [56].

Even in CPCS, consensus among different workers on specific questions is hard to achieve. Crowd-powered systems require crowds to share and maintain a globally consistent memory. To accomplish this goal, a fact board interface with important facts noted down by a group of previous workers was updated in a relative long-term interaction session as the conversation continued [48]. The facts were extracted from the previous conversation turns, including the description of the current task (e.g. ask for a restaurant) and the properties of the users themselves (e.g. the current user is vegan). Only 10 facts got maintained at a time and removed when more important facts were returned. Moreover, in an affective support task where the context necessary for new workers to follow up was not limited to identified facts, such a fact board could fail to capture the user’s intention of asking for help and the counselling techniques required from workers. The summarized facts curated by previous workers mostly included subjective assumptions about the current dialogue, which also inhibited new workers from following the context in a systematic way. We decide to guide workers to explore the conversation by our contextual guidance, where a systematic flow of helping understand the context is provided.

## 2.2 Context in Conversations

The concept of “context” was originally used in linguistics to refer to accompanying text [23], and was later extended to refer to the situation in which the discourse events and actions take place. We examine the role of context in coordinating communicative behaviour from a linguistic perspective, in which Bunt believed that the understanding of the context is relevant to five factors [13], namely the linguistic, semantic, cognitive, social and physical context. The linguistic context is defined as the surrounding utterances that have been said in the previous conversational turns. The semantic context is derived from the underlying goal of the communicative task, which means the understanding of the specific facts and the meanings are also involved. For the cognitive context, “current participants’ beliefs, intentions, and other attitudes” are reflected to perform the ongoing communicative task [13]. Specifically, Grosz and Sidner proposed the concept of the attentional state as the “information about the objects, properties, relations, and discourse intentions that are most salient at any given point” [27], which further implied that speakers’ intentions of uttering a specific discourse could not be ignored when presenting the cognitive context. The social context comprises the type of the dialogue (i.e., information-seeking dialogue; debating dialogue; ...) and the social roles of participants (i.e., employer-employee; customer-server; ...). The physical context refers to the spatiotemporal characteristics of the dialogue where the non-verbal behaviour or circumstances of the interaction could be emphasized. Different context dimensions contribute to the pragmatic knowledge [88] about the current situations and mutual understanding for both the speaker and the addressee [14] as the dialogue proceeds. The discourses are thus built progressively when the initial information keeps being updated by the context [26].

The mental model [38] also emphasizes the importance of cognitive and semantic dimensions in interpreting and constructing communication events [88]. Although the dialogue takes place at the social level from a macro perspective, the cognitive interaction

process of the interlocutors cannot be ignored. The mental model has been widely used to explain the cognitive bases that participants need to understand and build together while engaging in conversations. Besides, another purpose between the participants is to explore what the dialogue is about in a semantic way, where interlocutors constantly refer to the things discussed in the ongoing dialogue. Therefore, we intend to draw on the multi-dimensional context in linguistics to systematically present important information about the current conversation.

### 2.3 Motivational Interviewing

*Motivational Interviewing (MI)* is a client-centred, collaborative conversation about change and is widely used in behaviour change [12, 65, 75] and psychotherapy [2, 84]. The spiritual core that constitutes MI consists of four aspects: partnership, acceptance, compassion and evocation (PACE) [59]. Miller and Rollnick described four basic processes for using MI: engaging, focusing, evoking and planning [59]. Each stage requires a different amount and type of context to be focused on and elicited in the conversation. During the engaging process, the therapist develops a safe interactive environment where the patient can seek support. Micro-skills include open questions, affirmations, reflective listening and summarises (OARS) are proposed to strengthen links and facilitate the deep understanding of patients. In focusing, the therapist attempts to work with the patient to establish the central goal of the conversation, develop a focus and definition of the main challenge and help identify why the change is applicable based on internal values. In evoking, the therapist uses a variety of questions to explore the deep reasons for the change. Different types of change talk are involved, such as desire, ability, reasons and need (DARN) [6, 59]. To further strengthen the change talk, responses using elaborate, affirm, reflect and summarise (EARS) are advocated for the therapist. In planning, the therapist helps the patient decide the specific direction for change and develop an associated action plan.

MI was first applied to various behaviour change applications including physical activity change [67, 79]. Then it gained popularity in treating mental health problems, such as stress and depression. Previous research has shown that MI could be integrated into chatbots to deliver psychological interventions [50] but more contextualized responses were required [69]. However, these applications were proposed from the user perspective and the counsellor's responses were often pre-defined in the system.

In client-centred psychotherapy like MI, the information disclosed by the user to the therapist and the inquiry clues taken by the therapist constitute the context in the ongoing conversation, where different stages might be involved with different types of treatment. Skilled questioning [37] is a questioning technique that demands the therapist to recall and integrate the previous context, while simultaneously planning for the new questions. The combination of context in the previous and current conversations is indispensable for further questioning. Some common approaches used in counselling such as 5W1H [28] do not follow the context so the user may feel overwhelmed or stressed to answer aggressive questions. Therefore, we adopted MI-adherent guidance in our crowd-powered system to emphasize contextual understanding while workers respond to users with empathy.



## 2.4 Conversational Interfaces in Crowdsourcing

Conversational interfaces facilitate more natural interactions between individuals with the technology. Advances in dialogue systems have promoted the interest in conversational interfaces [98]. Users engage these third-party chatbots in dialogues to accomplish complex tasks, such as collaborative information-seeking [7, 89], task management [85], time scheduling [16], and group discussions [41, 42]. However, these applications are mediated by chatbots for main tasks, ignoring the potential of exploiting the conversational interface as an assistant tool for crowd-powered systems.

Earlier research has adopted text-based conversational interfaces to help crowd workers in various tasks such as image annotation and information finding [54], yielding comparable results and positive satisfaction from workers compared to traditional web interfaces. Furthermore, the impact of conversational styles on the output quality, user engagement, and cognitive load was investigated in conversational interfaces. The results revealed that with a more enthusiastic conversational style, workers generated better output and experienced less cognitive load in more difficult microtasks [73]. Besides traditional tasks, Trainbot [2] adopted conversational interfaces to train workers on MI to provide emotional support for stressed users. Workers reported less stress and performed better in answering quizzes when taking the test. Building upon [54, 2], we decide to introduce the conversational interface as a tool for providing context and investigated how the interface affected the output quality and interaction experience in the microtask of delivering support in an MI-centered dialogue.

## 2.5 Summary

Existing research on crowdsourcing has demonstrated the potential of delivering affective support through CPCS. However, effective approaches to providing context for on-demand workers have not yet been explored. To step further toward the field of integrating conversational interfaces into CPCS for systematically tracking context, we will build our system design upon the following aspects:

- **Crowd-powered conversational systems (CPCS):** We will design our worker interface for completing the affective support task by simulating the interactive interface in real-time CPCS, where a two-way ongoing conversation will be presented to the workers.
- **Context in conversations:** To address the contextual consistency issue in CPCS, we will conceptualize the context systematically based on Section 2.2, where multiple dimensions of information from the chat history will be extracted and provided in a customized conversation flow.
- **Motivational Interviewing (MI):** We will root our experiment in a therapeutic conversation using MI as the treatment method. Contextual guidance adherent to specific MI stages will be designed to enhance a better contextual understanding of the ongoing conversation.

## 2. RELATED WORK

---

- **Conversational interfaces in crowdsourcing:** We will design a conversational interface for workers to interact with a text-based chatbot. The conversational interface will be adopted as an assistant when workers need help to understand the previous context before responding to the user.

## Chapter 3

---

# System Design and Implementation

This chapter will discuss the system design and implementation in detail. We first provide the system architecture in Section 3.1 and then mainly explain the design of the conversational interface *ContextBot* in Section 3.2, including the context dimensions and conversational flow. After that, we describe the implementation of each element in the system in Section 3.3. Finally, we conclude our main design and implementation in Section 3.4.

### 3.1 System Architecture

We describe the architecture of our system supporting the conversations between an end-user and crowd workers, and the conversation flow between *ContextBot* and crowd workers. Figure 3.1 shows the main modules of the system. The system is comprised of one main **Task Window** where *ContextBot* is included, the backend programming application interfaces (**Task Executioner** and **Conversation Executioner**) and a relational database (**PostgreSQL Database**).

Crowd workers are required to help users with information needs in the **Task Window** which simulates the worker interface in Chorus [47]. The interaction between the end-user and workers is controlled by the **Task Executioner**. *ContextBot* is available as a conversational interface within the **Task Window** for the further query of the context. The **Conversation Executioner** is responsible for generating prompts for the chatbot and analyzing responses from workers. Responses and actions from workers related to interacting with the system will be saved to the **PostgreSQL Database**.

### 3.2 Design of *ContextBot*

To address *RQ1*, we aim to design a conversational interface that can deliver the context from multiple dimensions in a systematic way. In this section, we will first define the context dimensions used in *ContextBot* and how the context will be extracted based on a therapeutic conversation. Next, we will map the above context dimensions to a customized conversation flow, which will be adopted by *ContextBot* to prompt the context in a systematic way.

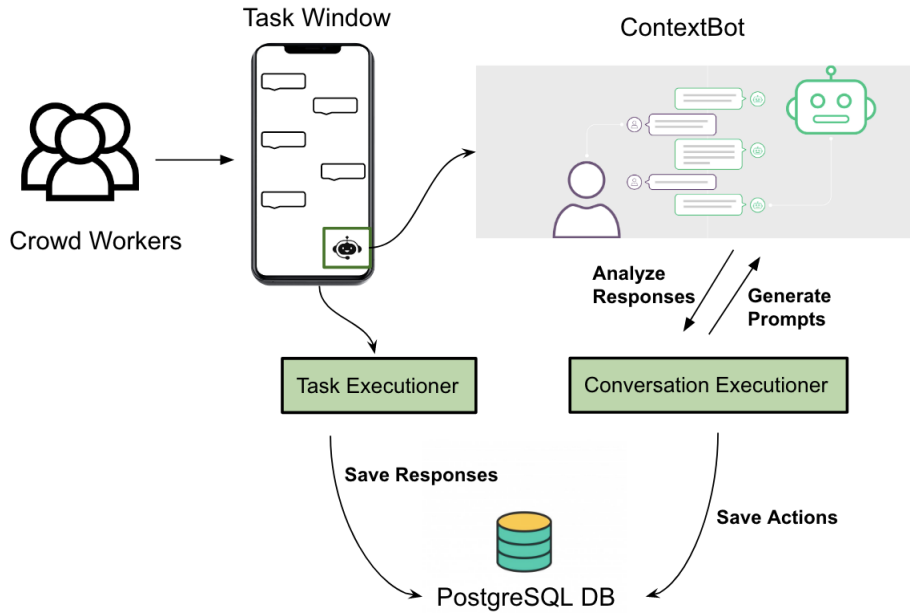


Figure 3.1: Overview of the system architecture.

### 3.2.1 Conceptualizing Context

Different context dimensions are related to each other and together constitute the factors for comprehending the current sentence. The dynamic process of MI also puts forward different discourse focusing on the context in each stage. Based on Section 2.2, we consider four dimensions of context mentioned in earlier work [13]: social context, linguistic context, semantic context, and cognitive context. Physical context involves the non-verbal behaviour of the dialogue, which is not applicable to our text-based conversational interface.

*Social Context* provides the type of dialogue (e.g., information-seeking dialogue) and the roles (e.g., employer-employee) that participants need to play from a global perspective, which is helpful for eliminating a new worker’s uncertainty about the current topic. Therefore, it is expected to be put in the first step of interaction to give global goals about the current conversation (e.g., engage a user; evoke a user’s desire to change).

*Linguistic Context* refers to the surrounding utterances that have been said in the previous conversational turns. Uncertain references in the current sentence can be elucidated by directly giving the previous sentences which contain events, places, and pronouns that have been referred to. We will manually inspect the previous chat history and extract the linguistic context. Automated solutions such as co-reference resolution could be employed to find relevant linguistic expressions related to the current entities [93]. However, this is not the current focus of our research.

*Semantic Context* is known to be “specific facts in the domain of discourse; the current state of the underlying task” [13]. It is derived from the underlying goal of the communicative task where specific facts and meanings are involved. We assume that the facts

contained in the semantic context can guide the worker to formulate a big picture of the current state of the ongoing chat, thus facilitating consistent discussion on change-focused talk. Specific facts until the current utterance would be summarized to help workers grasp what psychological issues or the situation of the user have been discussed in a faster way than reading the entire chat history. Three types of summary have been discussed in MI. (1) A linking summary, used to establish linking between the current utterance with what has been previously said. (2) A transitional summary, used to shift the topic focus of the dialog. (3) A collecting summary, used to combine relevant situations that are mentioned by the patient. We aim to make sure that the new worker who enters the dialogue at any time could have a basic understanding of the chat history and the user's problems. We thus provide a collecting summary of the previous turns for every new worker. We will manually summarize the corresponding number of facts based on the position of the current utterance in the chat. Text summarization can also be performed automatically by advanced natural language processing tools [22]. However, this is not the current focus of our research.

*Cognitive Context* reflects the intention and attentional state of participants toward the current utterance. Due to the subjectivity and elusiveness of cognitive states, different crowd workers may have different impressions of the same piece of dialogue. They tend to perceive the information from one's own given point, which makes it more difficult to maintain the consistency delivered to the same user, especially when they do not know the intentions of previous workers. For example, some workers may respond with an intention of being emotional while some workers may have the intention of presenting a practical solution. Therefore, we will explicitly summarize the intentions of the three roles involved in the current dialogue, the user, the previous workers, and the current worker. The user's intention is related to the type of help that he/she tried to seek in specific MI stages, which can also help the current worker understand why the current query was uttered by the user. The previous workers' intention provides a consistent standard for the current worker to refer to. The indication of the intention of the current worker then guides the worker to think in a consistent and desired way.

*MI Techniques* are suggested by *ContextBot* after all contexts are provided to the worker, aiming to help the worker respond in a more professional and empathetic manner. We will modify the templates previously used in [69, 2] and add instructions relevant to the corresponding MI stage.

### 3.2.2 Mapping Conversation Flow

We aim to provide contextual information in a systematic way, overcoming the weakness in traditional crowd-powered systems where separate facts are presented. To decompose the cognitive burden of workers understanding a large number of contexts at once, we use a linear way to sequentially provide four context dimensions. Figure 3.2 displays the conversation flow between *ContextBot* and the crowd worker.

① Once the worker clicks the *ContextBot* icon for help, the greeting from *ContextBot* is first displayed to ask if he/she would like to continue the dialogue. ② After receiving the acceptance, *ContextBot* drives the conversation by first introducing the global social context, including what role the worker should generally act as and the corresponding requirement

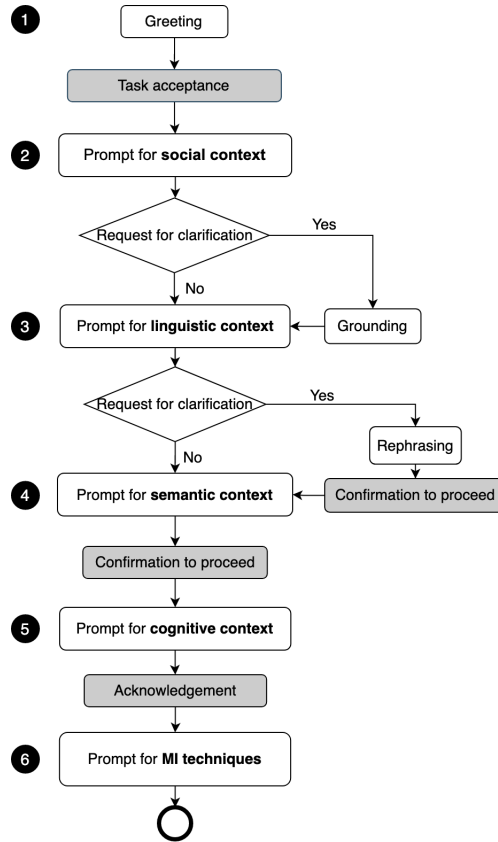


Figure 3.2: Conversation flow of ContextBot.

for the specific dialogue stage (i.e., engaging, focusing, evoking, planning). As the conversation is often viewed as a joint activity that requires turn-taking and grounding [56], we design quick reply buttons for workers to request further clarification from the bot. If the request is triggered, *ContextBot* will rephrase the context again and initiate the next prompt to carry on the dialogue [86]. ③ Next, *ContextBot* prompts the guidance of exploring linguistic context. The same request option is provided for the worker to ask for another turn of rephrasing. ④ After the confirmation, the worker is prompted to read the summarized semantic context. ⑤ Then, *ContextBot* guides the worker to explore each participant's (i.e., the user, the previous workers, the current worker) intentions as the cognitive context. ⑥ The final prompt for MI techniques comes after the acknowledgement of the current information from the worker. Adapted to the specific MI stage that the worker is working on, different micro-skills popularly implemented in MI will be shown to the worker.

### 3.3 System Implementation

We built our system as a web application using Flask [25]. The actual conversational interfaces in the **Task Window** were written with JavaScript using libraries such as jQuery [9]

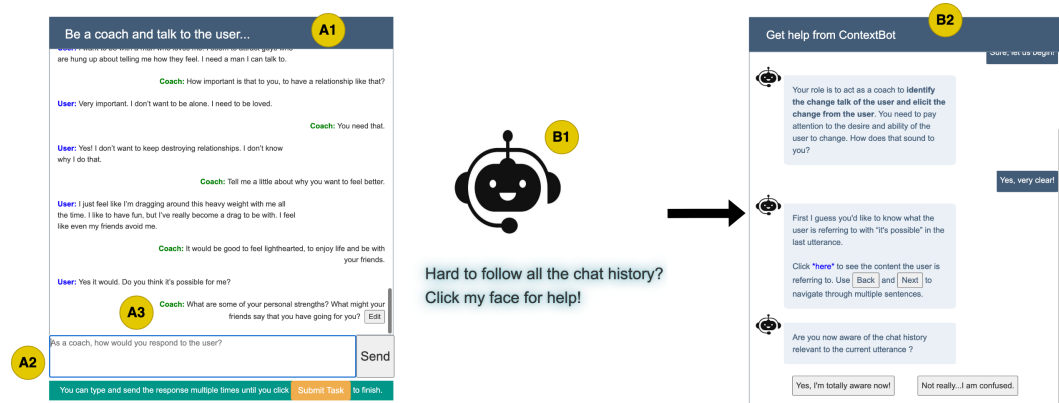


Figure 3.3: Worker interface of the system. The interface to interact with *ContextBot* is placed next to the main task window where crowd workers can read the chat and respond to the user.

to implement specific interactive behaviours. When crowd workers use the system, interactions will be recorded in a PostgreSQL database [18]. We deployed the system on Heroku [58] for the final production mode.

### 3.3.1 Worker Interface

As shown in Figure 3.3, we designed and implemented two conversational interfaces in the **Task Window**, one serving as a live chat window (A1) with a complete chat history to date was presented to new workers, who could scroll up and down to read the previous chat before writing responses in the input box (A2). To simulate a real chat, workers were also allowed to edit and refine their responses in an open-ended manner as shown in (A3). Next to the live chat window, *ContextBot* was displayed as a chatbot icon (B1). Workers could freely choose whether to click the icon to enter a conversation with *ContextBot* in (B2).

The interaction with the user was processed by the **Task Executioner**. Since the need for the quantity and content of context was affected by the number of conversational turns workers participated in, we restricted each worker to only respond in a single turn. We believe that if workers could provide consistent responses in a single turn, their ability to participate in multiple turns in CPCS would not be compromised.

For the interaction with *ContextBot*, another conversational interface will be activated when the chatbot icon (B1) is clicked. We leveraged the techniques used in most messenger applications [4], where a two-way interactive conversation was shown on the screen. Since our focus was to use the chatbot to provide context, we disabled the free text inputs from workers and only enabled quick-reply buttons for workers. Another concern was that workers could quickly go through the context by clicking buttons, which also helped reduce the latency caused by interacting with the chatbot in real-time CPCS. Inspired by the work of [73], we implemented two quick-reply choices with different conversation styles for every turn that required input from workers. For example, as shown in Figure 3.3, after *Con-*

*textBot* requested clarification about the current linguistic context, choices where one with high involvement “Yes, I’m totally aware now!” and another one with high considerateness “Not really...I am confused.” were shown to the worker at the same time. Then the answer would be processed by the backend API where **Conversation Executioner** controlled the pattern matching and analysis of quick-reply choices from workers. New prompts for the next conversational turn would be then generated based on the conversation flow shown in Section 3.2.2.

#### 3.3.2 Database

We adopted a PostgreSQL relational database to log the workers’ behaviour. The complete entity relationship diagram can be found in the appendix A.1. We briefly introduce each entity as follows.

**Worker:** The “Worker” table uses `id` as the primary key to uniquely identify all fields. It also stores the prolific ID of each worker as `prolific_id` and the timestamp of creating each record as `time_stamp`.

**Worker\_behavior:** It defines all possible clicking behaviours of the participant on the system page, including clicking quick-reply buttons, clicking the chatbot icon and clicking texts. It uses `worker_id` as a foreign key to link the table “Worker”.

**Message:** It stores the content of the messages sent by workers as `worker_utterance`, the message status as `msg_status`, which could be “Added” or “Deleted”, and the timestamp of creating one message as `time_stamp`. It further uses `worker_id` as a foreign key to link the table “Worker”.

**Time\_spent:** Each row records the execution time of a worker for a specific task condition stage. The execution time could be calculated by the minus between `end_time` and `start_time`. It further uses `worker_id` as a foreign key to link the table “Worker”.

#### 3.3.3 Data Preparation

We employed an affective support task which required workers to deliver emotional support by *Motivational Interviewing (MI)* [59], a client-centred conversation enhancing users to change. We selected and modified a complete case<sup>1</sup> as our study data which covered the four stages of MI by describing an interaction about how a therapist helped a depressed user from understanding her problems to develop a plan for change (Appendix A.2). Specifically, we shortened the conversation and focused on only one problem of the user. We chose this case because it contained the full four stages of MI. The rich explanations on the use of OARS techniques in the original case also helped us form MI prompts and extract relevant context.

---

<sup>1</sup><https://www.guilford.com/add/miller2/julia.pdf>



### 3.3.4 The platform for Crowdsourcing Tasks

We used Prolific.co platform to distribute our task to crowd workers. To connect the platform with our web application, we appended parameters to the end of our URL to record the worker IDs. Specifically, we stored the parameter `PROLIFIC_ID` in the “Worker” table as `prolific_id`. When a worker clicks the URL of our application through Prolific.co, the parameter `PROLIFIC_ID` will be automatically replaced by the actual prolific ID as an argument. Later on, the value of the ID will be extracted for processing the behaviour of this worker.

## 3.4 Summary

This chapter described the design and implementation details of our system that supports communication between the end-user and crowd workers. Specifically, our system consists of the **Task Window**, the backend programming application interfaces (**Task Executioner** and **Conversation Executioner**) and a relational database (**PostgreSQL Database**). We first discussed the context dimensions involved in the content provided by *ContextBot* and visualized the conversation flow of the interaction between *ContextBot* and workers. Next, we gave the implementation details of how the worker interface connected with the backend programming application interfaces. We stored relevant worker inputs into a relational database and connected the entire web application to a crowdsourcing platform for recruiting crowd workers.



## Chapter 4

---

# Evaluating ContextBot

In this chapter, the experimental goal and hypotheses will be first described in Section 4.1. Then we discuss the combination of 3(**entry points**)\*3(**contextual guidance**) experimental conditions and the mapping of context based on our dialogue in Section 4.2. In Section 4.3, we describe the participant recruitment strategy and then the procedure of participating in the whole study will be described in Section 4.4. Finally, we explain the dependent variables that we will use to test the hypotheses and evaluate the system in Section 4.5.

### 4.1 Goal and Hypotheses

To address *RQ2*, we propose the following hypotheses.

**Hypothesis 1 (H1):** *Interacting with ContextBot yields more consistent responses as compared to other conditions, across varying entry points and contextual guidance.*

**Hypothesis 2 (H2):** *When using ContextBot with MI-adherent guidance, crowd workers respond with more professional responses than those who interact with general guidance.*

**Hypothesis 3 (H3):** *When workers enter the dialogue late, interacting with ContextBot yields a better user experience compared to other conditions.*

**Hypothesis 4 (H4):** *Interacting with ContextBot does not significantly increase the cognitive load of workers as compared to the condition of only providing chat history.*

The goal of our experiments is to test the above hypotheses by exploring the impact of contextual factors in helping build such an assistant interface of providing context, specifically the impact of different entry points and types of contextual guidance on response quality and interaction experience. We assume that different entry points will result in workers being exposed to different amounts of context and stages of the MI counselling conversation, which will affect the quality of their responses and perception of *ContextBot*. As for the contextual guidance, it is expected that workers' understanding of the nature of the conversation will determine the professionalism in their responses and focus on the contextual information.

## 4. EVALUATING CONTEXTBOT

Entry Point	MI Stage	User's Utterance
Early	Beginning of engaging (4th turn)	Yes! It is so bad.
Middle	Half way of focusing (25th turn)	Yes it would. Do you think it's possible for me?
Late	End of planning (33rd turn)	I might just take a walk or see my friends. But like I said, it seems like they don't want to be around me so much anymore because I bring them down with me. Do you have some suggestions on what I should do?

Table 4.1: Choice of entry points corresponding to the user's utterance in each condition.

### 4.2 Study Design

We designed and implemented a conversational interface, *ContextBot*, to provide contextual information about the dialogue. To simulate CPCS, we set up another conversational interface to allow workers to respond to the user while a complete chat history to date was presented. The task goal for the worker was to join an online conversation as a therapy coach to provide a depressed user with consistent responses based on the chat history. New workers who entered the current dialogue could scroll up and down the chat to follow the history. As discussed before, we selected a counselling dialogue using MI as the treatment to verify the role of the system in delivering affective support. The study was authorized by the Ethics Committee of the Eindhoven University of Technology.


#### 4.2.1 Experimental Conditions

A 3(entry points)  $\times$  3(contextual guidance) between-subjects design was implemented to study the impact of entry points (early, middle, and late) and contextual guidance (MI-adherent guidance, general guidance, and no guidance) on response quality and interaction experience with *ContextBot*.

**Entry Points** The dynamic process of MI puts forward different requirements on the discourse focus of each stage, and the intervention from a later stage also requires more context. We considered three entry points for new workers: early, middle, and late. The later the workers entered the system, the more chat history they had to read. The difference was also reflected in the linguistic context referred by the current utterance and the amount of semantic context. We divided three different dialogue entry points according to the content involved in the dialogue and the stage it was in, corresponding to engaging, evoking, and planning respectively. For each stage, we selected the current user utterance having more ambiguous meanings as the experimental object because replying to such a sentence required more context from the history and the understanding of the MI stage, which also helped explore the roles played by different dimensions of context. Table 4.1 shows the user's current utterance selected for the three stages. The total number of conversational turns in our dialogue is 37. We gave the specific number of turns used in our conversation to mark each stage.

**Contextual Guidance** We aim to understand how *ContextBot*, with or without MI-adherent guidance, affects workers' behaviour. We created two versions of *ContextBot* (with MI-adherent guidance, with general non-MI guidance) and one version without *ContextBot* but only the chat history. Specifically, the chat history version without *ContextBot* was

used as a baseline condition in our experiment. The displayed context and amounts of contextual information shown by *ContextBot* under MI and non-MI conditions were the same. Differences would only occur in the specific content provided by social context and cognitive context directly related to corresponding MI stages (Table 4.5). We chose these two contexts because social context played the role of setting global conversational goals for workers, while cognitive context set local goals related to the current sentence. Specifically, cognitive context concretized the worker’s speech act in the pursuit of the conversation goal. In addition, we adapted the psychotherapy techniques to MI techniques in the MI condition. In the non-MI condition, only general techniques about how to respond empathetically were provided for workers. In the control condition, there was only one chat history window and no *ContextBot*.

To conclude, the tasks for three conditions with different contextual guidance include: (1) respond after receiving MI-adherent guidance from *ContextBot*; (2) respond after receiving general non-MI guidance from *ContextBot*; (3) respond naturally after reading the chat history. For conditions (1) and (2), after the worker enters the ongoing dialogue, *ContextBot* follows the conversation flow mentioned in Chapter 3 to guide the worker explore the corresponding context in terms of the current state of the MI stage. For condition (3), a complete chat history will be presented in the live chat window (Figure 3.3 ).

#### 4.2.2 Extracting Context

Exact instructions for workers to interact with *ContextBot* are based on the extraction of context at the specific dialogue stage. Table 4.2, Table 4.3, Table 4.4, and Table 4.5 respectively summarize the specific representations of different contextual content extracted from our study data.

For social context, we summarized the MI goals related to the MI stage of the current utterance, thus allowing workers to understand the MI focus of the current dialogue from a global perspective. For non-MI experimental conditions, workers would not be aware of the conversation that they were involved in was an MI-related counselling conversation. General prompts about using consistent responses would be provided for these workers.

For linguistic context, we invited another two experienced human-computer interaction (HCI) experts to jointly determine the relevant linguistic context that should be extracted from the selected current utterance. The criteria include, for each expert, (i) finding an unspecified noun or pronoun in the current utterance; (ii) and identifying sentences related to the unspecified word in the preceding text. Finally, the author of this project and two experts reached an agreement on the identified sentences as linguistic context.

For semantic context, we differentiated the amount of context shown by the point of the entry. Workers who entered earlier would be provided with less related sentences to the current utterance as the semantic context. For the early, middle and late entry points, the number of semantic context was 2, 4, and 6, respectively. Similar to the procedure of determining linguistic context, the author of this project and two experts collectively discussed past semantic information relevant to the current utterance, including but not limited to the user’s mental state, the type of help the user was seeking, objective facts about the user, factual events that the user had told the counsellor, etc.

#### 4. EVALUATING CONTEXTBOT

For cognitive context, we proposed intentions for the three roles involved in the current conversation, the user, the previous coach, and the current coach. For the user, we summarized the type of help asked by the user based on the progress of the conversation. For the previous coach, which was actually played by previous workers who entered the conversation earlier, their intention was about how to use relevant MI techniques to solve the user’s problems. Since it was difficult to summarize the intention that appeared in every previous conversation turn when the dialogue was too long, we only summarized the intention that was closest to the current user’s intention. For the current coach, we highlighted the MI techniques that they were supposed to use for the current MI stage, such as affirmations and reflective listening. For the non-MI experimental conditions, workers were not provided with any intentions related to the MI techniques, only with instructions to help them respond as consistently as possible.

Contextual Guidance	Entry Points	Social Context
<b>MI</b>	<b>Early</b>	Listen to and <b>engage</b> the user
	<b>Middle</b>	Identify the change talk of the user and <b>elicit</b> the change from the user
	<b>Late</b>	Develop a specific change <b>plan</b> for the user to implement
<b>Non-MI</b>	<b>Early</b>	Respond the user according to the historical context in this session
	<b>Middle</b>	
	<b>Late</b>	

Table 4.2: Social context summarized for different experimental conditions.

Contextual Guidance	Entry Points	Linguistic Context
<b>MI/Non-MI</b>	<b>Early</b>	“Kind of like a pattern..”
	<b>Middle</b>	“I don’t want to be alone. I need to be loved..”
	<b>Late</b>	“I feel like even my friends avoid me..”

Table 4.3: Linguistic context selected for different experimental conditions.

Contextual Guidance	Entry Points	Semantic Context
<b>MI/Non-MI</b>	<b>Early</b>	1. The user has broken up with her boyfriend. 2. She feels upset and confused by what’s happened to her past awful relationships.
	<b>Middle</b>	1, 2 3. She is depressed and wants to feel happy again. 4. She thinks she is a burden to her friends and feels like her friends avoid her.
	<b>Late</b>	1, 2, 3, 4 5. She is a resourceful person. She moved to the current place from Ireland all on her own and set up a new life for herself. 6. She has told the coach in the past how she lift her mood, such as going to see a funny and romantic movie, getting out of her apartment for a walk, and seeing friends.

Table 4.4: Semantic context summarized for different experimental conditions.

Contextual Guidance	Entry Points	Cognitive Context
MI	Early	User's intention: seek help and reasons for her low/negative emotions and bad relationships; Previous coach's intention: provide empathetic and reflective responses in terms of the user's problems; Your intention: continue improving the engagement of this dialog by <b>affirming and reflecting</b> what the user has said in an empathetic way;
	Middle	User's intention: need and desire to not destroy relationships and change her relationship patterns with friends; Previous coach's intention: figure out and focus on the depression problem of the user by asking relevant open questions, demonstrating empathy, and reflecting and confirming the user's concerns; Your intention: (1) start with <b>affirming and reflecting</b> the user's descriptions; (2) propose open questions to elicit the user's <b>ability</b> of change;
	Late	User's intention: find ways to lift her mood; Previous coach's intention: improve the user's mood by asking relevant open-ended questions, demonstrating empathy, and reflecting and confirming the user's concerns in order to elicit the user's feelings and preferences; Your intention: (1) start with <b>affirming and reflecting</b> the user's descriptions; (2) <b>summarize</b> the activities that the user has mentioned to lift her mood; (3) propose open questions to <b>ask further alternative options</b> that the user could use to relieve her tension;
Non-MI	Early	User's intention: same as above in each stage
	Middle	Previous coach's intention: provide consistent responses;
	Late	Your intention: continue providing <b>consistent</b> responses.

Table 4.5: Cognitive context summarized for different experimental conditions.



Figure 4.1: Overview of the task procedure.

### 4.3 Participants

We recruited 351 participants (~35 for each condition) via the Prolific.co platform. We only considered participants whose first language was English and those who came from the US or the UK. To limit the bias caused by work environments [21], participants can only join the study if they were using a laptop. We estimated £7.54/h for the task. Each worker was finally paid £8.63 per hour on average, which was considered good according to Prolific.co. After removing 13 workers who failed the attention check questions and 15 workers who interacted with *ContextBot* only after responding to the user, we finally had 323 unique workers (75.5% female, 23.5% male, 0.9% unknown gender). Their average age was 38.6 years old (SD=13.4).

### 4.4 Procedure

Figure 4.1 illustrates the task procedure for each worker participating in our study.

(1) **Pre-Task Questionnaire.** Participants will be first asked to select their mood out of nine scales before proceeding. The mood is designed based on the “Pick-A-Mood” instrument [17] which could help measure nine distinct mood states quickly. Figure 4.2 shows the mood characters used in our pre-task questionnaire.

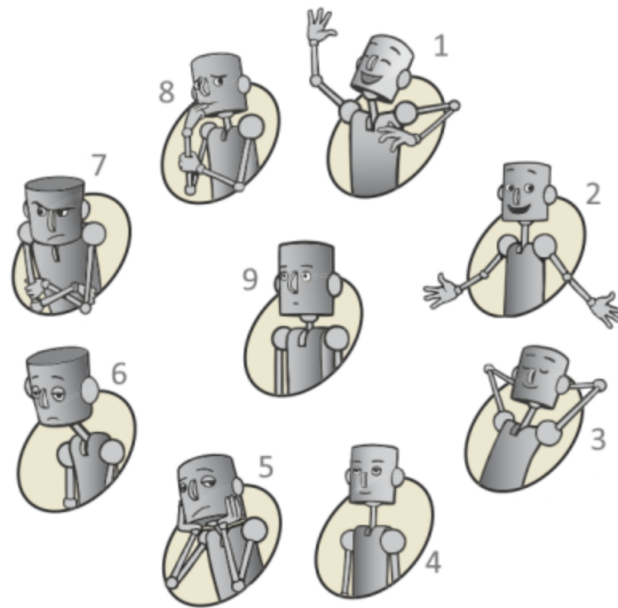


Figure 4.2: The mood scale to select in the pre-task questionnaire.

**(2) Introduction.** Next, we inform workers about the purpose of the study and how their data would be used. Once they give their consent, they are randomly assigned to one of the nine experimental conditions. Specific instructions on how to interact with the system are then provided to introduce the participants to the main components of the system.

**(3) Main Task.** The main task starts with clicking the “Start Task” button. The task goal is to give an appropriate response to the depressed user in the chat history window. Workers can freely choose whether to get assistance from *ContextBot* (if available). Each worker is allowed to respond to one latest user’s utterance in an open-ended manner. The end of the task is marked by clicking the “Submit Task” button after messages were sent.

**(4) Post-Task Questionnaire.** Workers are redirected to a post-task questionnaire after finishing the task. We ask workers to fill in a short User Experience Questionnaire (UEQ-S) to understand their perceived pragmatic and hedonic quality of the system design [78]. Next, workers are asked to report the cognitive load by completing the NASA Task Load Index (NASA-TLX) [29], which contains six items (mental demand, physical demand, temporal demand, self-performance, effort, and frustration). We further include several questions for measuring the user perceptions of general chatbots and design choices of *ContextBot* on a 7-point Likert scale. Finally, workers are allowed to give a satisfaction score on a 10-point Likert scale and leave comments about *ContextBot* and the system design.

**(5) Completion.** After completing the post-task questionnaire successfully, the workers are thanked for their participation and redirected to Prolific to receive their reward.



## 4.5 Evaluation Metrics

Throughout the experiment, we collected both quantitative and qualitative data to evaluate the following metrics.

**Response Consistency.** Automated dialogue systems usually consider whether the generated responses are consistent with the facts describing the speaker’s role, by regarding the consistency as a natural language inference (NLI) problem by calculating the inference relation scores of responses with the facts for each given persona [57, 81]. While the inference score is often used as a metric to compare with baseline models, human evaluation is still heavily used as an adjunct in judging context consistency [40]. Random samples are selected and provided for humans to assign labels. Since we value the consistency between the response and the chat history where multiple dimensions of context are involved, existing NLI frameworks which are mostly trained on fact-based datasets are ill-suited to our measurement [92]. Therefore, we recruited crowd workers from Prolific.co to rate the consistency of generated responses. To decide the number of responses to sample, we observed that the number of workers interacting with *ContextBot* was around 15 for each condition. The number of those who did not interact with *ContextBot* was almost twice of those who interacted. Considering the balance of responses generated after interaction and without interaction, all responses from those who interacted with *ContextBot* and 15 responses from those who did not were randomly sampled under each entry point from the MI and non-MI conditions, respectively. Also, we randomly sampled 15 responses from the history condition under each entry point. Each response was rated by three unique crowd workers. The consistency was measured on a 7-point Likert scale (1: *Highly inconsistent*, 7: *Highly consistent*) while considering the criteria in Appendix B.5.

**Professionalism in Responses.** We explored whether workers have followed the contextual guidance and counselling techniques applicable to the MI stage. We recruited two qualified psychologists on Fiverr to rate 5 sampled responses from those who fully interacted with *ContextBot* in MI and non-MI conditions, respectively. Each response was rated based on a 7-point Likert scale (1: *Highly unprofessional*, 7: *Highly professional*) while considering the criteria in Appendix B.6.

**User Experience.** We measured eight constructs by using the UEQ-S where each item was scored on a 7-point Likert scale (Appendix B.4). The eight constructs were divided into four items measuring the pragmatic level of the system and four items measuring the hedonic level. Based on prior work, we expected workers to obtain a better user experience after interacting with *ContextBot*, especially when they had to read a long chat history to follow the context.

**Cognitive Load.** To evaluate the perceived cognitive load for each task, workers were asked to answer six NASA-TLX questions where each was in 5-point increments, ranging from 0 to 100 (Appendix B.4). The lower the score, the lower the worker perceived the cognitive task load. Compared to only reading the chat history, although the interaction with *ContextBot* required some effort, it was not expected to significantly increase the cognitive load of finishing the whole task.



## Chapter 5

---

# Results

This chapter discusses the experimental results obtained from the evaluation. We divide the discussion into four parts: first, we perform descriptive analyses of the collected user data in Section 5.1, and then we perform statistical tests to test the hypotheses from the previous chapter in Section 5.2, and we give the qualitative results from the questionnaires in Section 5.3. Finally, we provide some exploratory insights from the data in Section 5.4.

### 5.1 Descriptive Statistics

After the following data preprocessing steps, Table 5.1 shows the number of records for each experimental condition across  $3(\text{entry points}) \times 3(\text{contextual guidance})$ . The total number of records is 323.

(1) Deleted the data that only existed in the post-task questionnaire but did not exist in the database due to the unstable database connection. Records of 18 workers were deleted.

(2) Deleted rows of unrecorded worker responses due to the unstable database connections. Records of 3 workers were deleted.

(3) Deleted a record that has been submitted twice to the database, and kept the time of the last submission as the final submission record.

(4) Removed workers who failed either of the two attention check questions. Records of 13 workers were deleted. The vacancies were recruited again by Prolific.

(5) Removed workers who lied about actual interactions, such as they have interacted with *ContextBot*, but reported that they did not in the post-task questionnaire. Records of 9 workers were deleted.

(6) Modified records that did not match the reported interaction behaviour. Some workers interacted after responding to the user but reported the actual interaction in the post-task questionnaire. We believed that such interactions did not reflect the true impact of interacting with *ContextBot* on the completion of the main task. We thus revised the interaction status as not interacted.

The number of workers participating in each condition is approximately equal to 35, which was estimated by G\*Power before the experiment. We also observed that for the

## 5. RESULTS

---

“MI+Early” condition, the number of workers who interacted with *ContextBot* was higher compared to other conditions.

Experimental Condition	Interacted	Count	Total
MI + Early	✓	17	36
	✗	19	
MI + Middle	✓	13	33
	✗	20	
MI + Late	✓	14	38
	✗	24	
Non_MI + Early	✓	14	36
	✗	22	
Non_MI + Middle	✓	10	32
	✗	22	
Non_MI + Late	✓	16	36
	✗	20	
History + Early	/	39	39
History + Middle	/	37	37
History + Late	/	36	36

Table 5.1: The number of records (by worker’s ID) across three types of contextual guidance and three entry points. The column “Interacted” indicates whether workers have interacted with *ContextBot* before responding to the user.

### 5.1.1 Pre-Task Questionnaire

We divided nine mood states into three groups: pleasant (1-4), unpleasant (5-8), and neutral (9). Figure 5.1 shows the number of workers in three groups where each group further comprises workers who have interacted or not interacted with *ContextBot* (three history conditions are not taken into account). Compared to the unpleasant group, more workers who were in a pleasant mood state interacted with *ContextBot*. The result might imply that further research could look into the relationship between the pre-task mood and the intention of interacting with *ContextBot*.

### 5.1.2 Execution Time

We calculated the execution time based on the interaction states under different experimental conditions in Table 5.2. For MI and non-MI conditions, workers who interacted with *ContextBot* generally spent more time on the main task regardless of the entry point. Meanwhile, the standard deviation for the interacted groups was also higher than the not interacted groups. We observed an increasing trend in the execution time when workers entered the dialogue later.

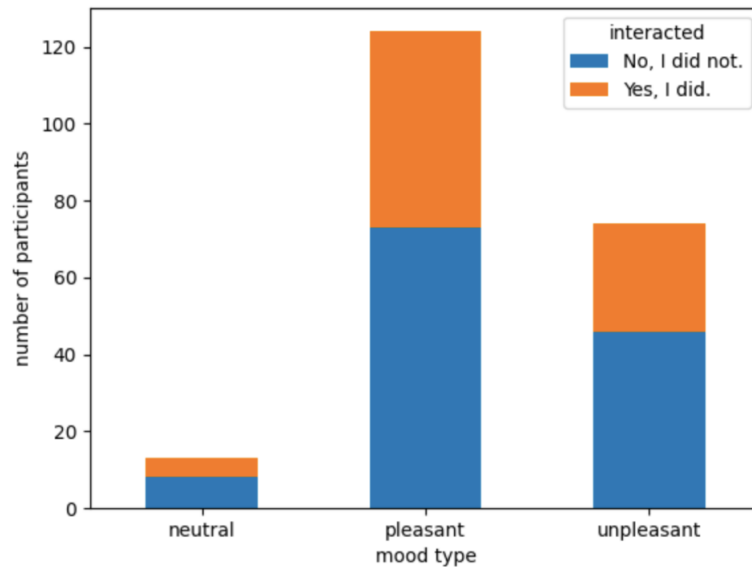


Figure 5.1: The number of workers having different mood states.

For three entry points under each contextual guidance, we adopted a Kruskal-Wallis test [55] to verify whether there was a significant difference among entry points. We noticed that the execution time was significantly different among entry points in the MI (interacted:  $p = .008$ , not interacted:  $p = .008$ ) and history ( $p = .000$ ) conditions. However, the difference was not significant in the non-MI condition. We inferred that when only general guidance instead of the MI-adherent guidance was provided for workers, workers no longer considered coming up with MI-related replies based on different stages of historical conversations. They just focused on the consistency of conversations of different lengths, where we still observed an increasing trend of the time spent on lengthy conversations for reading and following the context.

### 5.1.3 Workers' Responses

The average length of the responses was 26.56 words for each worker. Table 5.3 shows the example responses in each experimental condition.

## 5.2 Hypothesis Tests

### 5.2.1 Response Consistency

To understand the impact of interaction types (interacted with *ContextBot*, not interacted with *ContextBot*) on response consistency in all groups with *ContextBot* available, we first compared response consistency between the interacted groups in MI and non-MI conditions under three entry points. The results of the two-tailed independent

## 5. RESULTS

Contextual Guidance	Interacted	Entry Points			Sig.
		Early	Middle	Late	
MI	✓	252.24 ± 182.53	335.20 ± 223.40	421.61 ± 170.48	<b>.008</b>
	✗	108.92 ± 55.72	150.60 ± 101.29	191.31 ± 93.08	<b>.008</b>
Non-MI	✓	217.32 ± 87.59	262.62 ± 143.84	316.11 ± 205.98	.549
	✗	106.80 ± 60.02	166.10 ± 111.78	133.54 ± 87.64	.121
History	/	123.89 ± 116.49	136.09 ± 106.01	221.58 ± 126.62	<b>.000</b>

Table 5.2: Execution time (*mean ± std*, in seconds) of different conditions across interacted status. ✓ stands for “interacted with *ContextBot*” while ✗ is for “did not interact with *ContextBot*”. Bold numbers indicate the difference among entry points for the corresponding contextual guidance is significant.

Entry Point	User’s Utterance	Response(MI)	Response(Non-MI)	Response(History)
Early	Yes! It is so bad.	I appreciate that you are sharing your emotions so honestly with me. I know that breaking up with your boyfriend has made you sad, and that must be really difficult.	Tell me the three qualities that are important to you in a man. What is it you are looking for?	Do you have any insight why this pattern is occurring?
Middle	Yes it would. Do you think it’s possible for me?	Yes. I appreciate you speaking so openly. In what ways would you benefit from spending time with your friends and family?	Of course it is possible. I see you are fearful of the same things repeating but you’ll never know if you don’t give yourself a chance	I can hear that this is something you would like to be able to do
Late	I might just take a walk or see my friends. But like I said, it seems like they don’t want to be around me so much anymore because I bring them down with me. Do you have some suggestions on what I should do?	Going for a walk sounds like a good idea, maybe listening to music will lift your mood too. You said you feel like your friends don’t want to be around you, is there anything they’ve said or done that makes you think this?	I would recommend embarking on a new hobby, or looking into joining a walking group in your area. This would be a great chance to meet some new people and expand your social circle.	what in particular makes you feel like it seems they don’t want to be around you?

Table 5.3: Example responses from workers who interacted with *ContextBot* towards the user’s utterance.

T-test showed that the contextual guidance did not significantly affect the consistency at any entry point (*early*,  $p=.490$ ; *middle*,  $p=.219$ ; *late*,  $p=.745$ ). Next, we computed response consistency across three entry points by combining the MI and non-MI conditions, as shown in Figure 5.2. Without considering the specific contextual content of the interaction and the stage where the interaction finally ended, we observed that workers who entered the early dialogue after the interaction produced more consistent responses compared to not interacted and history groups. The difference was significant ( $p<.001$ ) under the one-way ANOVA test at  $\alpha=.05$  level. A posthoc analysis with the Tukey–Kramer test showed

that the interacted group and the not interacted group were significantly different ( $p < .001$ ) under  $\alpha = .05$  level while the interacted group and the history group did not show the significant difference ( $p = .054$ ). When we extracted only the samples that completed the entire interaction flow from all the samples having interactive behaviours, we found that the difference was still significant ( $p = .002$ ) for the early entry point.

However, although we expected that workers who entered the conversation at a late point would grasp the long context faster and produce more consistent responses after interacting with *ContextBot*, the difference was not significant for the late entry point. To explore the potential reasons why we found partial support for **H1**, we examined other factors related to the consistency scores, such as the percentage of workers using different types of context during the interaction, their familiarity with chatbots, their satisfaction with the system, etc., reported by workers in the post-task questionnaire. As shown in Table 5.4, two subgroups (lower than the average consistency score vs. higher than the average consistency score) were extracted from samples who had interactive behaviours in the MI and non-MI conditions, respectively.

We examined the difference between each subgroup and the history group under MI and non-MI conditions by using the Mann-Whitney test and calculating the Hedges'  $g$  effect size  $g$ . We found that for the MI condition, the group with lower consistency ( $N=8$ ,  $p=.001$ ,  $g=1.284$ ) and higher consistency ( $N=5$ ,  $p=.020$ ,  $g=1.612$ ) both spent longer time than the history group ( $N=36$ ). Interestingly, the group with lower consistency scores reported higher UEQ ( $p=.012$ ,  $g=1.048$ ) and hedonic scores ( $p=.011$ ,  $g=1.080$ ) than the history group in the MI condition. For the non-MI condition, we also observed significant differences in the UEQ ( $p=.015$ ,  $g=1.050$ ) and pragmatic scores ( $p=.005$ ,  $g=1.074$ ) between the lower consistency group ( $N=8$ ) and the history group ( $N=36$ ). Moreover, workers from both MI and non-MI conditions with above-average consistency scores had higher percentages of finishing the whole interaction. The results suggest that the perceived user experience and completeness of interacting with *ContextBot* could potentially relate to the consistency level of responses.

### 5.2.2 Trade-Offs Between Quality and Execution Time

We measured the time between loading the main task page and the submission of a task from the worker as execution time (in seconds). The average time required to complete the task is shown in Table 5.5. As expected, when workers entered a conversation later, they tended to spend more time on the main task. The response time in the history condition was also comparable to that of existing real-time CPCS [47], where a worker spent 103.4s on replying. Next, we examined whether interacting with *ContextBot* produced highly consistent responses at the expense of more time. Since we used a 7-point Likert scale to evaluate consistency, we considered responses with a score higher than 4 as highly consistent responses. We divided the samples that interacted with *ContextBot* under each entry point and contextual guidance into two groups, with consistency scores higher than 4 and less than or equal to 4. We compared the time differences of high consistency groups between MI and history conditions, and non-MI and history conditions, respectively. The Mann-Whitney tests showed the difference was significant for all entry points between MI and

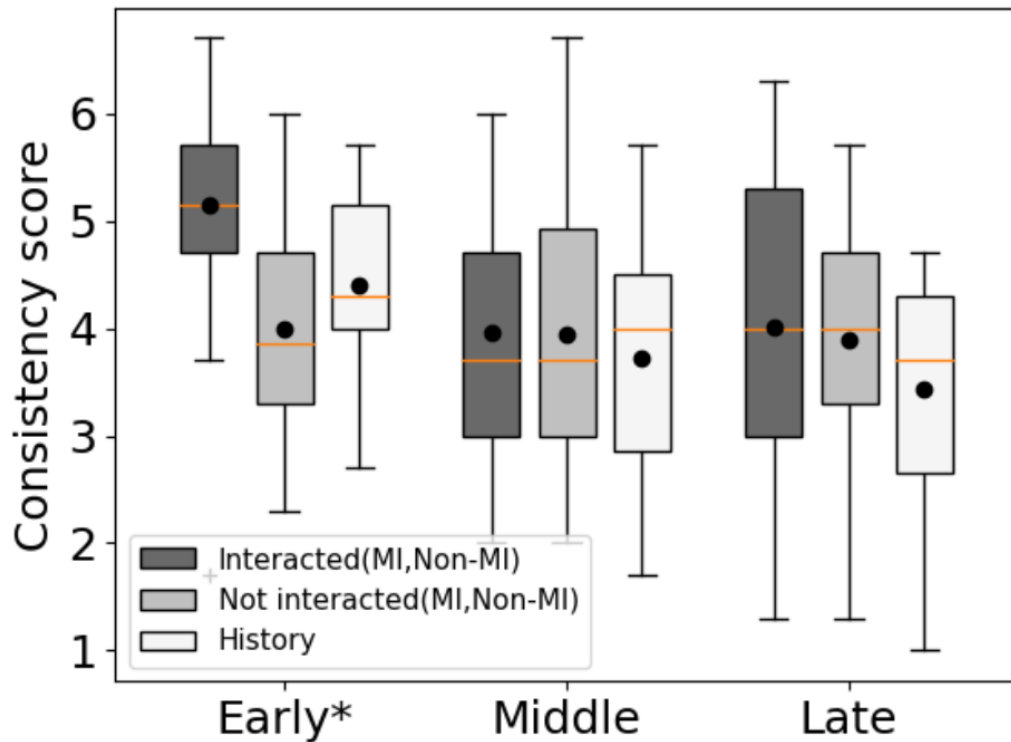


Figure 5.2: Response consistency scores across interacted types and entry points by combining the MI and non-MI conditions. \* = statistically different (interacted vs. not interacted vs. history).

history conditions (*early*,  $p=.003$ ; *middle*,  $p=.041$ ; *late*,  $p=.004$ ) under  $\alpha=.05$  level. For the comparison between non-MI and history conditions, we only compared the early and late entry points since the sample size was too small ( $N=3$ ) for the middle entry. The difference was significant for both groups (*early*,  $p=.045$ ; *late*,  $p=.030$ ). Our results indicate a trade-off between response consistency and execution time. It is feasible to introduce *ContextBot* to CPCS in real-time at the cost of response delays, provided that mitigation techniques could be employed to reduce the annoyance caused by waiting [3].

In terms of the consistency level, we found that for the early and middle conditions, workers who were able to provide highly consistent responses spent less time on average. In contrast, when the dialogue became longer, generating highly consistent responses required more time. The trend implies that for the conversation of short to medium length, not interacting with *ContextBot* not only impairs response consistency but may also result in a longer time to read the chat and come up with replies.

### 5.2.3 Professionalism in Responses

Table 5.6 shows the average scores for sampled responses assessed by two professional psychologists. Since hiring psychologists was expensive, we limited our sample size in



Related Variables	MI		Non-MI		History, N=36
	Lower, N=8	Higher, N=5	Lower, N=8	Higher, N=6	
Social Context (%)	100	100	100	100	
Linguistic Context (%)	50	50	71.4	66.7	
Semantic Context (%)	87.5	100	85.7	66.7	
Cognitive Context (%)	87.5	75	57.1	66.7	
MI Techniques (%)	50	100	57.1	66.7	
All Guidance (%)	37.5	50	57.1	66.7	
Execution Time (in seconds)*	<b>381.33</b>	<b>446.96</b>	253.02	434.29	232.98
UEQ*	<b>5.52</b>	5.10	<b>5.52</b>	5.23	4.72
Pragmatic Score*	5.72	5.80	<b>6.25</b>	5.54	5.15
Hedonic Score*	<b>5.31</b>	4.40	4.78	4.92	4.29
Task Load	37.75	39.07	36.90	41.33	38.61
Familiarity	3.25	4.60	4.00	4.00	
Frequency	2.29	3.60	2.50	3.20	
AskForHelp	2.57	4.00	3.43	3.40	
Perceived Explanation	5.00	4.20	4.88	4.60	
Context Flow	5.13	4.00	4.88	4.40	
Control Feeling	4.13	4.00	3.88	4.20	
Satisfaction Score	7.88	8.00	8.25	6.83	7.00

Table 5.4: Comparison of consistency scores and other related variables for the late entry point. In “Lower” group, each sample has the consistency score lower than the average of the current condition while in “Higher” group, samples have scores higher than the average. Variables with (%) indicate the percentage of interacting with these variables in the current group. Variables with “\*” mean that the bold number in the current group is significantly different from the history (control) group in the posthoc test.

each condition to  $N=5$ . Workers following MI-adherent guidance consistently produced more professional responses across three entry points as shown in Table 4.1. These results support **H2** but a larger scale validation with more samples is required in the future to evaluate the effectiveness of MI-adherent guidance on improving the professionalism of responses. Especially for the early entry point, the higher average score and comparable standard deviation indicated that MI-adherent guidance might increase workers’ level of responding to the user with more empathy and relevance.

#### 5.2.4 Perceived Interaction Experience with ContextBot

**User experience.** The user experience scores obtained by averaging the eight items of UEQ-S are shown in Figure 5.3. We found that the choice of whether or not to interact with *ContextBot* significantly affected the user experience perceived by workers under different conditions. Workers who interacted with *ContextBot* reported higher scores than those who did not interact, regardless of being provided with MI-adherent or general guidance (Figure 5.3 (a)). Based on the data distributions, we performed a two-tailed independent T-test

## 5. RESULTS

Contextual Guidance	Consistency	Entry Points		
		Early	Middle	Late
MI	High	267.35 ± 186.38	252.21 ± 105.07	428.98 ± 208.47
	Low	*	406.32 ± 268.82	387.37 ± 119.18
Non-MI	High	208.81 ± 100.04	178.20 ± 38.28	434.29 ± 261.72
	Low	238.58 ± 34.66	298.80 ± 156.73	253.02 ± 116.65
History	High	120.45 ± 94.54	140.50 ± 49.84	136.74 ± 53.88
	Low	180.85 ± 245.19	114.95 ± 47.35	225.26 ± 120.08

Table 5.5: Execution time (*mean ± std*, in seconds) of different conditions across consistency levels. Responses with consistency scores higher than 4 are regarded as “High” level of consistency. \* indicates no samples in this group.

	MI	Non-MI
	<i>Mean ± Std</i>	<i>Mean ± Std</i>
Early	5.60 ± 1.24	4.40 ± 1.24
Middle	5.60 ± 0.58	5.30 ± 0.98
Late	5.10 ± 1.32	4.50 ± 0.95
Overall	5.43 ± 1.12	4.73 ± 1.14

Table 5.6: Professionalism in responses (*Mean ± Std*, measured by a 7-point Likert scale) for groups with and without contextual guidance across three entry points.

and Mann-Whitney test respectively to test the difference in MI condition (interacted ( $M=5.01$ ,  $SD=0.82$ ), not interacted ( $M=4.60$ ,  $SD=0.94$ ),  $p=.020$ ) and non-MI condition (interacted ( $M=5.12$ ,  $SD=0.95$ ), not interacted ( $M=4.62$ ,  $SD=0.88$ ),  $p=.017$ ) under  $\alpha=.05$  level. When workers entered into long conversations with longer chat history (Figure 5.3 (b)), workers who interacted with *ContextBot* experienced a better user experience ( $M=5.34$ ,  $SD=0.66$ ), which was significantly different ( $p=.011$ ) by the Mann-Whitney test. Interestingly, we also observed that the above results were consistent at the pragmatic level when we divided the user experience items into four pragmatic and four hedonic items. This suggests that *ContextBot* can be practically used with contextual guidance to help workers and we found support for **H3**.

**Cognitive load.** Figure 5.3 (c) and (d) show the effect of different conditions on the cognitive load perceived by workers. Workers at different entry points did not report significantly different cognitive load when *ContextBot* was available. When only the chat history was provided, the Kruskal-Wallis test indicated that there was a significant difference ( $p=.010$ ) among three entry points under  $\alpha=.05$  level. A posthoc analysis with Dunn’s test using the Bonferroni correction further showed that workers who entered the conversation in the late state ( $M=37.27$ ,  $SD=10.27$ ) reported more cognitive load ( $p=.007$ ) than those who entered the conversation at medium length ( $M=29.27$ ,  $SD=9.42$ ). In terms of entering stages,

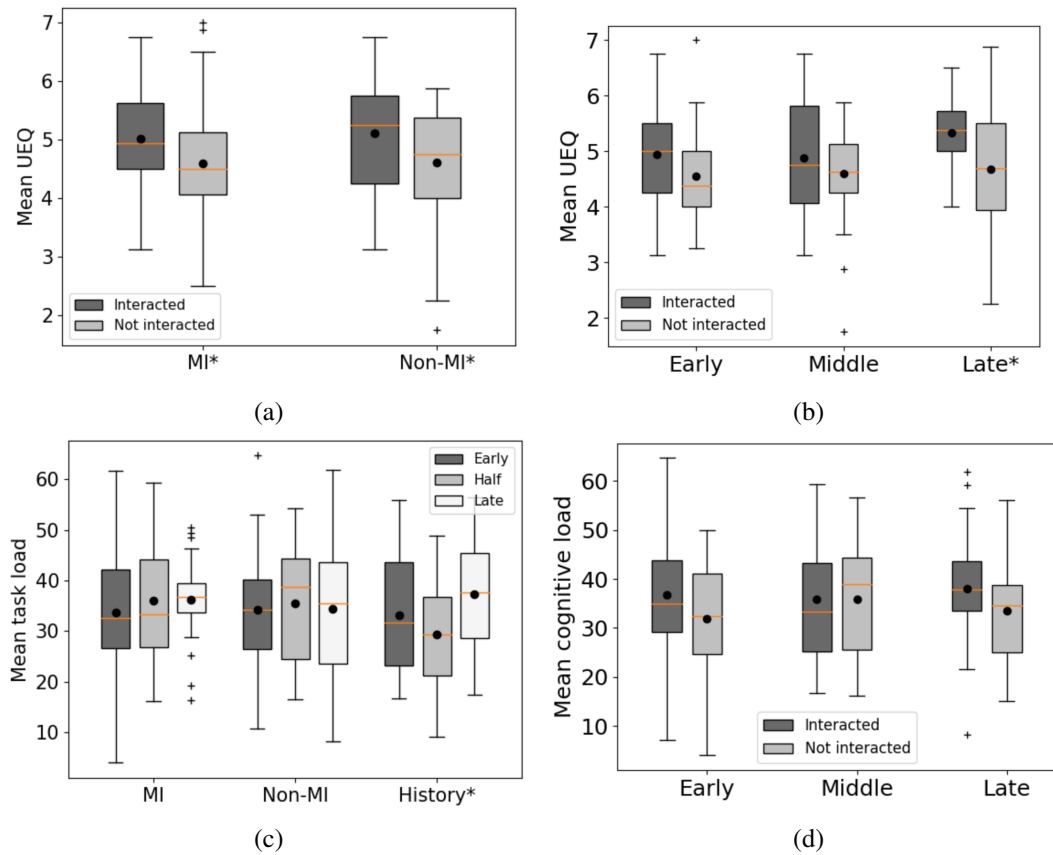


Figure 5.3: Boxplots of UEQ-S scores (Fig (a), (b), \* = statistically significant between interacted group and not interacted group) and NASA-TLX scores (Fig (c), (d), \* = statistically significant among early, middle, and late groups in the History condition in Fig (c)) in terms of the interacted types across different conditions. Black points indicate the average value of this group.

workers who interacted with *ContextBot* showed a higher cognitive load on average compared to those who did not, but the difference was not significant at  $\alpha=.05$  level using a Mann-Whitney test. These results support **H4**, reflecting that increasing interaction with *ContextBot* does not significantly increase the perceived cognitive load of workers.

### 5.3 Qualitative Analysis

We followed an inductive thematic analysis approach to analyse the open-ended comments from 20 workers who have interacted with *ContextBot* [11]. To this end, we carried out the thematic analysis process. Through iterative deliberations, we identified and reviewed themes from the codes to capture important narratives in relation to our system design. The worker IDs in the following example excerpts are randomized for anonymous review.

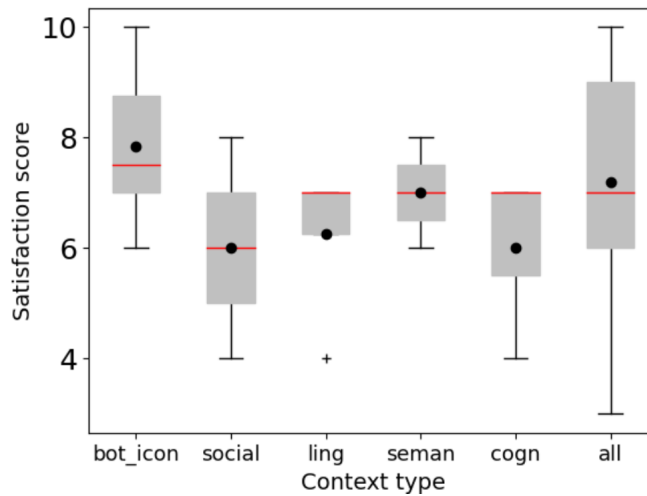


Figure 5.4: Relations between context types and the satisfaction score. *bot\_icon* is the group who only clicked the chatbot icon without further interactions with *ContextBot*; *social* is the group who only interacted with social context; *ling* is the group who followed the interaction from the start to linguistic context; *seman* is the group who followed the interaction from the start to semantic context; *cogn* is the group who followed the interaction from the start to cognitive context; *all* is the group who finished the whole interaction with *ContextBot*. The average and median values of corresponding groups are indicated by black points and red lines, respectively.

**Positive Experience with ContextBot.** 55% of workers reported positive experiences with *ContextBot* as being helpful, easy, clear, interesting, and supportive.

**W3:** I thought it was a very clear and easy system to understand and use.

**W1:** I found the suggestion of what kind tone to adopt helpful and was glad there were a couple of example phrases.

**W19:** It gave me different options if I needed more help.

**Effectiveness of Guidance.** 15% of workers felt that *ContextBot* provided them with limited information, resulting in a lack of effective guidance to help them. Part of the complaints came from questions about the design of *ContextBot* and part of the complaints were about the quantity and quality of content presented by *ContextBot*. The expectations about *ContextBot* might differ among workers, highlighting the importance of giving clear instructions on helping them perceive such a conversational interface as an assistive tool rather than an extra task.

**W9:** The options were a bit too limited and formulaic.

**W18:** It was just stating facts, it didn't feel like it was actually supporting or helping if I didn't know what to say.

**W20:** It doesn't look appealing.

**Confidence in Affective Support Tasks.** 25% of workers acknowledged the difficulty of being a coach without enough training. Similar concerns could be addressed by pre-training workers with online exercises [2].

**W7:** I can't help but think it may be more ethical/responsible if the replies/counsel was coming from someone with enough experience to know how to phrase, reflect and support the user appropriately without the use of ContextBot.

**W11:** Not ever having been in the position of a “coach” it was still quite difficult to know how to respond.

**Low Latency and Collective Identity in CPCS.** Two workers have mentioned the challenge of deploying *ContextBot* in CPCS, which lies in reducing the response latency and maintaining collective identity across workers [33].

**W2:** However, if this exchange between counsellor and depressed user was real time, I wonder how professional it would be to disturb the flow with having to look up, read and digest the information given by the chatbot? In some cases this pause could be misinterpreted by the user.

**W6:** it was quite difficult to pretend i [sic] was the same person as previous people.

## 5.4 Exploratory Findings

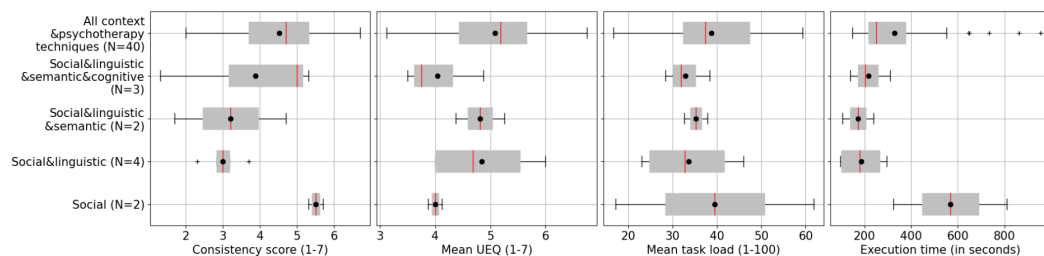


Figure 5.5: Relations between context types and relevant variables (samples are selected from those who had interactive behaviour with *ContextBot*). The average and median values of corresponding groups are indicated by black points and red lines, respectively.

Our results on response consistency, execution time, UEQ, and cognitive load illustrated the potential benefits of interacting with *ContextBot*. Figure 5.5 presents the relations between four variables and the interacted context type. We observed that not every worker completed the whole interaction with *ContextBot*. We classified the workers based on the number of contexts that they finished interacting, including the ones who only interacted with social context ( $N = 2$ ), the ones who interacted with both social and linguistic context ( $N = 4$ ), the ones who interacted with social, linguistic, and semantic context ( $N = 2$ ), the ones who interacted with social, linguistic, semantic and cognitive context ( $N = 3$ ), and the ones who interacted with all context dimensions and MI techniques ( $N = 40$ ). Constrained

## 5. RESULTS

---

by the sample size though, we found that workers who read all context on average, could systematically produce more consistent responses, had higher user experience, comparable cognitive load, and less execution time.

Furthermore, Figure 5.4 shows the relations between context types and the satisfaction score towards the system design from workers. Despite the fact that we did not observe a clear effect of different context dimensions on the satisfaction score, the satisfaction scores above 8 were only found in the worker samples that completed all interactions. Interestingly, the average reported satisfaction score was the highest for those workers who only clicked the chatbot icon but did not proceed with obtaining more context information. We reasoned that the workers' curiosity about using the chatbot might be taken into account of their satisfaction with our system design.

## Chapter 6

---

# Discussions

In this chapter, we review the findings from the results and link them to the relevant literature in Section 6.1. Next, we discuss the research limitations of our study in Section 6.2.

### 6.1 Interpretation and Implications

Our results suggest that *ContextBot* can serve as an effective tool for providing context in CPCS in affective support tasks. The setting of MI-adherent contextual guidance in this paper initially realized the future prospect of [69], providing a strategy for generating responses consistent with MI to avoid naive reflections towards the user's concerns. The task of contextual understanding was embodied as the perception of specific contextual factors and MI-adherent templates which could help workers generate custom responses. In fact, the comprehension of context is indispensable in many counselling techniques, such as skilled questioning where the therapist needs to recall and integrate the previous context while planning for further questioning [37].

Our findings of the response quality corroborated results from the work of [54] where the difference in the output quality was not significant as a result of using conversational interfaces. In addition, our results also confirmed the benefits of leveraging conversational interfaces for improving workers' interaction experience without negatively increasing the task burden. Similar to other studies using conversational interfaces to enhance the design diversity in crowdsourcing, we believe that a properly designed conversational interface will help workers better grasp contexts in such affective support tasks. At the same time, the entry points we considered suggest that interacting with short conversations and contexts helps yield more consistent responses. When the chat becomes long, higher completion of the context flow and lower perceived entertainment may be helpful for high consistency in MI-centered dialogue but large-scale studies are needed to validate the significance. Aligned with our expectations, workers are more likely to follow the chatbot's guidance if they can view *ContextBot* as an assistant with informative benefits, rather than an entertaining feature. In terms of the execution time, it does not necessarily take longer to generate highly consistent responses when workers interact with *ContextBot*. However, results in Table 5.5 have revealed that more time may be required when workers enter the dialogue

late. The original intention of designing *ContextBot* is to balance low latency and high-quality results in real-time CPCS. Given the variance, our results of execution time for early and middle entry points are similar to that reported in [47], where for the retrieval task it took an individual worker 103.4s to reply. However, the execution time in the late entry reveals the risk of deploying our system in the real world. While the average duration of a conversation was 11.21 minutes in Chorus, we cannot determine whether the conversation would be too long when workers reach the late entry point defined in our paper. A possible solution for this would be using a brief MI [74] as an online intervention instead. Despite the trade-off between response consistency and execution time, introducing *ContextBot* to long conversations could help improve the user experience of workers without burdening them with significantly increased cognitive load. It is therefore applicable to use *ContextBot* in CPCS where techniques of mitigating waiting time or asynchronous models are present.

Our takeaway from adopting conversational interfaces to provide the context in CPCS is that the attractiveness of the interface design, the trade-off between simplicity (amount of context) and functionality (quality of context) of the interface, and task types are all essential factors to consider when creating supportive dialogue interfaces for therapeutic applications. When the task requires specific guidance (e.g., MI-adherent techniques), workers may expect professional advice and context at the same time. The conversation flow should take into account both stated facts and effective guidance where the amount of utterances presented by the chatbot is controlled appropriately.

### 6.2 Limitations

Although our results on response quality, execution time, UEQ, and cognitive load illustrated the potential benefits of interacting with *ContextBot*, our limitation is that these results are affected by the interaction states of context types and still need to be validated with larger datasets. In this study, we only presented a way of mapping multi-dimensional context to a linear conversation flow, which put high demands on the interaction intention of workers towards the system. If the worker failed to continue in the process of linear interaction, they would take the risk of not knowing other important contexts relevant to the user concerns or previous workers' intentions. An alternative approach is to choose a side-by-side or open-choice design to provide different context dimensions. Yet, mechanisms to retain workers to explore as many contexts as possible are still required. From the perspective of contextual content, although the focus of this study is not on how to extract personalized context from the conversation, the manual reviewing and summarising procedures adopted in our experiment will not be applicable in larger-scale studies. A combination of automated methods and human intelligence might be an alternative. While our system only focused on a single-turn interaction between the end-user and crowd workers, in a real-time CPCS, workers can participate in multiple conversation turns and choose to exit the dialogue at any time. It remains challenging for *ContextBot* to provide dynamically updating context and identify the end of the conversation for an individual worker [33].

Moreover, contextual factors involved in counselling conversations vary with the specific techniques (MI, CBT, etc.) employed during the therapy. Although we took advantage



of the different contextual contexts involved in the four-stage counselling conversation exemplified by MI, such contextual guidance does not necessarily apply to other types of counselling conversations. When the context related to specific treatment skills is difficult to be specified, the summarization of the context will become more ambiguous, which is challenging even for the technology of automatic text summarization. A better paradigm for embedding contextual understanding into similar context-tracking tools still remains an open topic. Another potential limitation is that we only considered one type of crowdsourcing task, the affective support task. In most CPCS where information retrieval tasks are commonly found, contextual guidance related to the task could be more explicit and homogeneous in different stages of the conversation.

Likewise, the definition of entry points was determined according to the content of a pre-existing corpus selected in our experiment. The manual approach to deciding on the context type and content was time-consuming and hard to generalize to other dialogues. It is expected that research on conversation structures could answer the challenge of dividing appropriate dialogue stages from a linguistic perspective.



## Chapter 7

---

# Conclusions and Future Work

This chapter concludes the main findings of the project and specifically answers the two research questions raised at the beginning of the project in Section 7.1, and finally we present some possible directions for future work in Section 7.2.

### 7.1 Conclusions

Online peer-to-peer psychological support has the advantages of instantaneity and easy availability that cannot be easily replaced by traditional counselling. CPCS provide a chance to solicit emotional intelligence from crowds to help users in real-time. However, it is difficult to maintain global worker memory, and new workers may give inconsistent responses as a result of not being able to quickly identify and understand the context of the chat history. In this paper, we introduced a conversational interface, *ContextBot*, to provide workers with systematic context and guidance, and explored the impact of this interface on response quality and interaction experience. We verified the effectiveness of *ContextBot* in MI-centered dialogue and found that with short chat histories, interacting with *ContextBot* improved response consistency, but with longer chat histories, response consistency may still be affected by contextual completion and perceived entertainment level. In addition, workers who interacted with *ContextBot* did not feel the increased cognitive load, but showed better user experience across different contextual guidance, and also when faced with a longer chat history.

To answer our main *RQ* “**How does *ContextBot* used for providing context in CPCS affect the response quality and interaction experience of workers?**” in Section 1.2, specifically we summarized our findings based on *RQ1* and *RQ2* as follows:

***RQ1: How could the context be extracted and provided for crowd workers to understand the chat history?***

We did a literature review on the linguistic perspective of context in conversations. Inspired by the work of [13] where five factors were proposed to describe the context, we adopted four dimensions to conceptualize the important information involved in the history of dialogue: social context, linguistic context, semantic context, and cognitive context. The

social context explains the global goal of the conversation. The linguistic context gives the utterances in the chat history that contain the referential nouns involved in the current utterance. The semantic context includes important facts relevant to the user and the topic. The cognitive context is represented by the intentions of all speakers participating in the conversation. We further introduced contextual guidance tailored to the affective support task, where context aligned with the specific counselling stage was presented to workers. To provide workers with the context, we designed a linear conversation flow to present social, linguistic, semantic, cognitive context, and MI techniques sequentially to help workers grasp the contexts systematically, overcoming the weakness of traditional CPCS where only separate facts were recorded by previous workers. We managed to embed deep contextual understanding into multi-dimensional contexts and enable workers to follow the contexts in a systematic way.

***RQ2: How does interacting with ContextBot affect the workers' response quality (consistency and professionalism) and their interaction experience (user experience and cognitive load)?***

We performed a  $3 \times 3$  between-subjects design to evaluate the effect of **contextual guidance** and **entry points** on the response quality and interaction experience. Three types of contextual guidance were used in our experiment, namely MI-adherent guidance, general guidance, and no guidance. MI-adherent guidance differed from general guidance in that customized social and cognitive contexts based on one of the four stages of MI were provided for workers accordingly, while with general guidance workers were only required to respond consistently to the current utterance. In addition, workers could enter a crowd-powered system at different stages of the conversation. We manually identified three entry points based on the number of conversation turns in our experiment: early, middle, and late. In summary, we found that:

- Contextual guidance did not significantly affect response consistency, but workers who entered the conversation earliest tended to generate highly consistent responses after interacting with *ContextBot*. This lay in contrast to those who did not interact with *ContextBot*.
- There was a trade-off between response consistency and execution time by using *ContextBot*. When workers interacted with *ContextBot*, it did not take more time to generate highly consistent responses in conversations of short to medium length.
- MI-adherent guidance can help yield more professional responses that are compliant with counselling requirements.
- Better user experience levels from workers were associated with interacting with *ContextBot*. Workers who entered the system late reported significantly increased user experience after the interaction.
- Regardless of the entry points, the interaction with *ContextBot* did not negatively affect cognitive load.

## 7.2 Future Work

Based on this project, we propose the following directions for future work to explore.

**Experiment with more task types.** Our work concretized and provided context for the affective support task that only used MI as the main counselling technique. A potential future direction would be to explore affective conversations which adopt several counselling techniques in combination or information-oriented conversations that require inquiry skills. A general paradigm for the abstraction of contextual information in different task types is expected in the future.

**Automation of multi-dimensional context extraction.** Although context extraction under the human reviewing process can cover important chat histories in a more detailed and accurate manner, automated context extraction is still required in long conversations. A possible future direction would be to use keyword detection techniques to automate linguistic context extraction and then use text summarization techniques to automate social, semantic, and cognitive context extraction.

**Improvement of context-tracking interfaces.** We noticed that about 2/3 of the workers chose not to use *ContextBot* when it was available, depending on the attractiveness of the design and the instructive details of the task. We propose that future researchers could look into how to improve workers' intentions to use dialogue assistant tools to obtain context or to incorporate the traditional "fact board" to create more engaging and effective worker memory maintenance tools.

**Integration with real-time techniques.** Our work simulated the scenario of providing context in a single-turn manner in CPCS without controlling the response delay explicitly. However, in current real-time CPCS, existing real-time techniques are dedicated to maintaining a pool of workers, so that online users can receive responses within a few seconds or minutes. It is expected to include real-time techniques to test the prototype of our system to verify the feasibility of introducing an auxiliary interface for memory maintenance on a larger scale.



---

## Bibliography

- [1] Tahir Abbas, Vassilis-Javed Khan, Ujwal Gadiraju, Emilia I. Barakova, and Panos Markopoulos. Crowd of oz: A crowd-powered social robotics system for stress management. *Sensors*, 20(2):569, 2020. doi: 10.3390/s20020569. URL <https://doi.org/10.3390/s20020569>.
- [2] Tahir Abbas, Vassilis-Javed Khan, Ujwal Gadiraju, and Panos Markopoulos. Trainbot: A conversational interface to train crowd workers for delivering on-demand therapy. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 8, pages 3–12, 2020.
- [3] Tahir Abbas, Ujwal Gadiraju, Vassilis-Javed Khan, and Panos Markopoulos. Making time fly: Using fillers to improve perceived latency in crowd-powered conversational systems. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 9, pages 2–14, 2021.
- [4] Nahdatul Akma Ahmad, Mohamad Hafiz Che, Azaliza Zainal, Muhammad Fairuz Abd Rauf, and Zuraidy Adnan. Review of chatbots design techniques. *International Journal of Computer Applications*, 181(8):7–10, 2018.
- [5] Kathina Ali, Louise Farrer, Amelia Gulliver, and Kathleen M Griffiths. Online peer-to-peer support for young people with mental health problems: a systematic review. *JMIR mental health*, 2(2):e4418, 2015.
- [6] Paul C Amrhein, William R Miller, Carolina E Yahne, Michael Palmer, and Laura Fulcher. Client commitment language during motivational interviewing predicts drug use outcomes. *Journal of consulting and clinical psychology*, 71(5):862, 2003.
- [7] Sandeep Avula, Gordon Chadwick, Jaime Arguello, and Robert Capra. Searchbots: User engagement with chatbots during collaborative search. In Chirag Shah, Nicholas J. Belkin, Katriina Byström, Jeff Huang, and Falk Scholer, editors, *Proceedings of the 2018 Conference on Human Information Interaction and Retrieval, CHIIR 2018, New Brunswick, NJ, USA, March 11-15, 2018*, pages 52–61. ACM, 2018. doi: 10.1145/3176349.3176380. URL <https://doi.org/10.1145/3176349.3176380>.

- [8] Michael S. Bernstein, Joel Brandt, Robert C. Miller, and David R. Karger. Crowds in two seconds: enabling realtime crowd-powered interfaces. In Jeffrey S. Pierce, Maneesh Agrawala, and Scott R. Klemmer, editors, *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology, Santa Barbara, CA, USA, October 16-19, 2011*, pages 33–42. ACM, 2011. doi: 10.1145/2047196.2047201. URL <https://doi.org/10.1145/2047196.2047201>.
- [9] Bear Bibeault, Aurelio De Rosa, and Yehuda Katz. *jQuery in Action*. Simon and Schuster, 2015.
- [10] Jeffrey P. Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C. Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samuel White, and Tom Yeh. Vizviz: nearly real-time answers to visual questions. In Ken Perlin, Mary Czerwinski, and Rob Miller, editors, *Proceedings of the 23rd Annual ACM Symposium on User Interface Software and Technology, New York, NY, USA, October 3-6, 2010*, pages 333–342. ACM, 2010. doi: 10.1145/1866029.1866080. URL <https://doi.org/10.1145/1866029.1866080>.
- [11] Virginia Braun and Victoria Clarke. *Thematic analysis*. American Psychological Association, 2012.
- [12] Christine Bundy. Changing behaviour: using motivational interviewing techniques. *Journal of the royal society of medicine*, 97(Suppl 44):43, 2004.
- [13] Harry Bunt. Context and dialogue control. *Think Quarterly*, 3(1):19–31, 1994.
- [14] Herbert H. Clark and Catherine R. Marshall. Definite knowledge and mutual knowledge. In Aravind K. Joshi, Bonnie L. Webber, and Ivan A. Sag, editors, *Elements of Discourse Understanding*, pages 10–63. Cambridge, UK: Cambridge University Press, 1981.
- [15] Kenneth Mark Colby. Modeling a paranoid mind. *Behavioral and Brain Sciences*, 4(4):515–534, 1981.
- [16] Justin Cranshaw, Emad Elwany, Todd Newman, Rafal Kocielnik, Bowen Yu, Sandeep Soni, Jaime Teevan, and Andrés Monroy-Hernández. Calendar.help: Designing a workflow-based scheduling agent with humans in the loop. In Gloria Mark, Susan R. Fussell, Cliff Lampe, m. c. schraefel, Juan Pablo Hourcade, Caroline Appert, and Daniel Wigdor, editors, *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, Denver, CO, USA, May 06-11, 2017*, pages 2382–2393. ACM, 2017. doi: 10.1145/3025453.3025780. URL <https://doi.org/10.1145/3025453.3025780>.
- [17] Pieter MA Desmet, Martijn H Vastenburg, and Natalia Romero. Mood measurement with pick-a-mood: review of current methods and design of a pictorial self-report scale. *Journal of Design Research*, 14(3):241–279, 2016.



- 
- [18] Joshua D Drake and John C Worsley. *Practical PostgreSQL*. ” O’Reilly Media, Inc.”, 2002.
- [19] Kathleen Kara Fitzpatrick, Alison Darcy, and Molly Vierhile. Delivering Cognitive Behavior Therapy to Young Adults With Symptoms of Depression and Anxiety Using a Fully Automated Conversational Agent (Woebot): A Randomized Controlled Trial. *JMIR Mental Health*, 4(2):e19, June 2017. ISSN 2368-7959. doi: 10.2196/mental.7785. URL <http://mental.jmir.org/2017/2/e19/>.
- [20] Russell Fulmer, Angela Joerin, Breanna Gentile, Lysanne Lakerink, Michiel Rauws, et al. Using psychological artificial intelligence (tess) to relieve symptoms of depression and anxiety: randomized controlled trial. *JMIR mental health*, 5(4):e9782, 2018.
- [21] Ujwal Gadiraju, Alessandro Checco, Neha Gupta, and Gianluca Demartini. Modus operandi of crowd workers: The invisible role of microtask work environments. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 1(3):49:1–49:29, 2017. doi: 10.1145/3130914. URL <https://doi.org/10.1145/3130914>.
- [22] Mahak Gambhir and Vishal Gupta. Recent automatic text summarization techniques: a survey. *Artificial Intelligence Review*, 47(1):1–66, 2017.
- [23] Mohsen Ghadessy. *Text and context in functional linguistics*, volume 169. John Benjamins Publishing, 1999.
- [24] SR Gouravajhala, YOUXUAN Jiang, Preetraj Kaur, Jarir Chaar, and Walter S Lasecki. Finding mnemo: Hybrid intelligence memory in a crowd-powered dialog system. In *Collective Intelligence Conference (CI 2018)*. Zurich, Switzerland, 2018.
- [25] Miguel Grinberg. *Flask web development: developing web applications with python*. ” O’Reilly Media, Inc.”, 2018.
- [26] Jeroen Groenendijk and Martin Stokhof. Context and information in dynamic semantics. *Law Policy - LAW POLICY*, pages 458–486, 01 1988. doi: 10.1016/B978-0-12-238050-1.50027-5.
- [27] Barbara J Grosz and Candace L Sidner. Attention, intentions, and the structure of discourse. *Computational linguistics*, 12(3):175–204, 1986.
- [28] Sangdo Han, Kyusong Lee, Donghyeon Lee, and Gary Geunbae Lee. Counseling dialog system with 5w1h extraction. In *Proceedings of the SIGDIAL 2013 Conference, The 14th Annual Meeting of the Special Interest Group on Discourse and Dialogue, 22-24 August 2013, SUPELEC, Metz, France*, pages 349–353. The Association for Computer Linguistics, 2013. URL <https://aclanthology.org/W13-4054/>.
- [29] Sandra G Hart and Lowell E Staveland. Development of nasa-tlx (task load index): Results of empirical and theoretical research. In *Advances in psychology*, volume 52, pages 139–183. Elsevier, 1988.

- [30] Minlie Huang, Xiaoyan Zhu, and Jianfeng Gao. Challenges in building intelligent open-domain dialog systems. *ACM Trans. Inf. Syst.*, 38(3):21:1–21:32, 2020. doi: 10.1145/3383123. URL <https://doi.org/10.1145/3383123>.
- [31] Ting-Hao K. Huang, Amos Azaria, Oscar J. Romero, and Jeffrey P. Bigham. Instructablecrowd: Creating IF-THEN rules for smartphones via conversations with the crowd. *Hum. Comput.*, 6:113–146, 2019. doi: 10.15346/hc.v6i1.7. URL <https://doi.org/10.15346/hc.v6i1.7>.
- [32] Ting-Hao (Kenneth) Huang, Walter S. Lasecki, and Jeffrey P. Bigham. Guardian: A crowd-powered spoken dialog system for web apis. In Elizabeth Gerber and Panos Ipeirotis, editors, *Proceedings of the Third AAI Conference on Human Computation and Crowdsourcing, HCOMP 2015, November 8-11, 2015, San Diego, California, USA*, pages 62–71. AAAI Press, 2015. URL <http://www.aaai.org/ocs/index.php/HCOMP/HCOMP15/paper/view/11599>.
- [33] Ting-Hao Kenneth Huang, Walter S. Lasecki, Amos Azaria, and Jeffrey P. Bigham. ”is there anything else I can help you with?” challenges in deploying an on-demand crowd-powered conversational agent. In Arpita Ghosh and Matthew Lease, editors, *Proceedings of the Fourth AAI Conference on Human Computation and Crowdsourcing, HCOMP 2016, 30 October - 3 November, 2016, Austin, Texas, USA*, pages 79–88. AAAI Press, 2016. URL <http://aaai.org/ocs/index.php/HCOMP/HCOMP16/paper/view/14050>.
- [34] Ting-Hao Kenneth Huang, Joseph Chee Chang, and Jeffrey P. Bigham. Evorus: A crowd-powered conversational assistant built to automate itself over time. In Regan L. Mandryk, Mark Hancock, Mark Perry, and Anna L. Cox, editors, *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, CHI 2018, Montreal, QC, Canada, April 21-26, 2018*, page 295. ACM, 2018. doi: 10.1145/3173574.3173869. URL <https://doi.org/10.1145/3173574.3173869>.
- [35] Ting-Hao (Kenneth) Huang, Joseph Chee Chang, and Jeffrey P. Bigham. Evorus: A Crowd-powered Conversational Assistant Built to Automate Itself Over Time. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–13, Montreal QC Canada, April 2018. ACM. ISBN 978-1-4503-5620-6. doi: 10.1145/3173574.3173869. URL <https://dl.acm.org/doi/10.1145/3173574.3173869>.
- [36] Becky Inkster, Shubhankar Sarda, Vinod Subramanian, et al. An empathy-driven, conversational artificial intelligence agent (wysa) for digital mental well-being: real-world data evaluation mixed-methods study. *JMIR mHealth and uHealth*, 6(11):e12106, 2018.
- [37] Ian Andrew James, Rachel Morse, and Alan Howarth. The science and art of asking questions in cognitive therapy. *Behavioural and Cognitive Psychotherapy*, 38(1):83–93, 2010.

- [38] P. N. Johnson-Laird. *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness*. Harvard University Press, USA, 1986. ISBN 0674568826.
- [39] Hee Ju Kang and Seung In Kim. Evaluation on the usability of chatbot intelligent messenger mobile services-focusing on google (allo) and facebook (m messenger). *Journal of the Korea Convergence Society*, 8(9):271–276, 2017.
- [40] Hyunwoo Kim, Byeongchang Kim, and Gunhee Kim. Will I sound like me? improving persona consistency in dialogues through pragmatic self-consciousness. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 904–916. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.65. URL <https://doi.org/10.18653/v1/2020.emnlp-main.65>.
- [41] Soomin Kim, Jinsu Eun, Changhoon Oh, Bongwon Suh, and Joonhwan Lee. Bot in the bunch: Facilitating group chat discussion by improving efficiency and participation with a chatbot. In Regina Bernhaupt, Florian 'Floyd' Mueller, David Verweij, Josh Andres, Joanna McGrenere, Andy Cockburn, Ignacio Avellino, Alix Goguey, Pernille Bjøn, Shengdong Zhao, Briane Paul Samson, and Rafal Kocielnik, editors, *CHI '20: CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, April 25-30, 2020*, pages 1–13. ACM, 2020. doi: 10.1145/3313831.3376785. URL <https://doi.org/10.1145/3313831.3376785>.
- [42] Soomin Kim, Jinsu Eun, Joseph Seering, and Joonhwan Lee. Moderator chatbot for deliberative discussion: Effects of discussion structure and discussant facilitation. *Proc. ACM Hum. Comput. Interact.*, 5(CSCW1):1–26, 2021. doi: 10.1145/3449161. URL <https://doi.org/10.1145/3449161>.
- [43] Rachel Kornfield, David C Mohr, Rachel Ranney, Emily G Lattie, Jonah Meyerhoff, Joseph J Williams, and Madhu Reddy. Involving crowdworkers with lived experience in content-development for push-based digital mental health tools: Lessons learned from crowdsourcing mental health messages. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW1):1–30, 2022.
- [44] Ben Krause, Marco Damonte, Mihai Dobre, Daniel Duma, Joachim Fainberg, Federico Fancellu, Emmanuel Kahembwe, Jianpeng Cheng, and Bonnie L. Webber. Edina: Building an open domain socialbot with self-dialogues. *CoRR*, abs/1709.09816, 2017. URL <http://arxiv.org/abs/1709.09816>.
- [45] Liliana Laranjo, Adam G. Dunn, Huong Ly Tong, Ahmet Baki Kocaballi, Jessica A. Chen, Rabia Bashir, Didi Surian, Blanca Gallego, Farah Magrabi, Annie Y. S. Lau, and Enrico W. Coiera. Conversational agents in healthcare: a systematic review. *J. Am. Medical Informatics Assoc.*, 25(9):1248–1258, 2018. doi: 10.1093/jamia/ocy072. URL <https://doi.org/10.1093/jamia/ocy072>.

- [46] Walter S. Lasecki, Kyle I. Murray, Samuel White, Robert C. Miller, and Jeffrey P. Bigham. Real-time crowd control of existing interfaces. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology - UIST '11*, page 23, Santa Barbara, California, USA, 2011. ACM Press. ISBN 978-1-4503-0716-1. doi: 10.1145/2047196.2047200. URL <http://dl.acm.org/citation.cfm?id=2047196.2047200>.
- [47] Walter S. Lasecki, Rachel Wesley, Jeffrey Nichols, Anand Kulkarni, James F. Allen, and Jeffrey P. Bigham. Chorus: a crowd-powered conversational assistant. In Shahram Izadi, Aaron J. Quigley, Ivan Poupyrev, and Takeo Igarashi, editors, *The 26th Annual ACM Symposium on User Interface Software and Technology, UIST'13, St. Andrews, United Kingdom, October 8-11, 2013*, pages 151–162. ACM, 2013. doi: 10.1145/2501988.2502057. URL <https://doi.org/10.1145/2501988.2502057>.
- [48] Walter Stephen Lasecki and Jeffrey Philip Bigham. Automated support for collective memory of conversational interactions. In *Human Computation and Crowdsourcing: Works in Progress and Demonstration Abstracts, An Adjunct to the Proceedings of the First AAAI Conference on Human Computation and Crowdsourcing, November 7-9, 2013, Palm Springs, CA, USA*, volume WS-13-18 of AAAI Technical Report. AAAI, 2013. URL <http://www.aaai.org/ocs/index.php/HCOMP/HCOMP13/paper/view/7502>.
- [49] Jiwei Li, Michel Galley, Chris Brockett, Georgios P. Spithourakis, Jianfeng Gao, and William B. Dolan. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics, 2016. doi: 10.18653/v1/p16-1094. URL <https://doi.org/10.18653/v1/p16-1094>.
- [50] Christine L. Lisetti, Reza Amini, Ugan Yasavur, and Naphtali Rishe. I can help you change! an empathic virtual agent delivers behavior change health interventions. *ACM Trans. Manag. Inf. Syst.*, 4(4):19:1–19:28, 2013. doi: 10.1145/2544103. URL <https://doi.org/10.1145/2544103>.
- [51] Gale M Lucas, Albert Rizzo, Jonathan Gratch, Stefan Scherer, Giota Stratou, Jill Boberg, and Louis-Philippe Morency. Reporting mental health symptoms: breaking down barriers to care with virtual human interviewers. *Frontiers in Robotics and AI*, 4:51, 2017.
- [52] Kien Hoa Ly, Ann-Marie Ly, and Gerhard Andersson. A fully automated conversational agent for promoting mental well-being: A pilot rct using mixed methods. *Internet interventions*, 10:39–46, 2017.
- [53] David Maulsby, Saul Greenberg, and Richard Mander. Prototyping an intelligent agent through wizard of oz. In Bert Arnold, Gerrit C. van der Veer, and Ted N. White, editors, *Human-Computer Interaction, INTERACT '93, IFIP TC13 International Conference on Human-Computer Interaction, 24-29 April 1993, Amsterdam*,

- The Netherlands, jointly organised with ACM Conference on Human Aspects in Computing Systems CHI'93*, pages 277–284. ACM, 1993. doi: 10.1145/169059.169215. URL <https://doi.org/10.1145/169059.169215>.
- [54] Panagiotis Mavridis, Owen Huang, Sihang Qiu, Ujwal Gadiraju, and Alessandro Bozzon. Chatterbox: Conversational interfaces for microtask crowdsourcing. In George Angelos Papadopoulos, George Samaras, Stephan Weibelzahl, Dietmar Jan-nach, and Olga C. Santos, editors, *Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization, UMAP 2019, Larnaca, Cyprus, June 9-12, 2019*, pages 243–251. ACM, 2019. doi: 10.1145/3320435.3320439. URL <https://doi.org/10.1145/3320435.3320439>.
- [55] Patrick E McKight and Julius Najab. Kruskal-wallis test. *The corsini encyclopedia of psychology*, pages 1–1, 2010.
- [56] Michael McTear, Zoraida Callejas, and David Griol. *The Conversational Interface: Talking to Smart Devices*. Springer Publishing Company, Incorporated, 1st edition, 2016. ISBN 3319329650.
- [57] Mohsen Mesgar, Edwin Simpson, and Iryna Gurevych. Improving factual consistency between a response and persona facts. In Paola Merlo, Jörg Tiedemann, and Reut Tsarfaty, editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 549–562. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.eacl-main.44. URL <https://doi.org/10.18653/v1/2021.eacl-main.44>.
- [58] Neil Middleton and Richard Schneeman. *Heroku: up and running: effortless application deployment and scaling*. ” O’Reilly Media, Inc.”, 2013.
- [59] William R Miller and Stephen Rollnick. *Motivational interviewing: Helping people change*. Guilford press, 2012.
- [60] Madison Milne-Ives, Caroline de Cock, Ernest Lim, Melissa Harper Shehadeh, Nick de Pennington, Guy Mole, Eduardo Normando, Edward Meinert, et al. The effectiveness of artificial intelligence conversational agents in health care: systematic review. *Journal of medical Internet research*, 22(10):e20346, 2020.
- [61] Robert Morris. Crowdsourcing workshop: the emergence of affective crowdsourcing. In *Proceedings of the 2011 annual conference extended abstracts on Human factors in computing systems*. Citeseer, 2011.
- [62] Robert R. Morris and Rosalind Picard. Crowdsourcing Collective Emotional Intelligence. *arXiv:1204.3481 [cs]*, April 2012. URL <http://arxiv.org/abs/1204.3481>.

- [63] Robert R Morris, Stephen M Schueller, and Rosalind W Picard. Efficacy of a Web-Based, Crowdsourced Peer-To-Peer Cognitive Reappraisal Platform for Depression: Randomized Controlled Trial. *Journal of Medical Internet Research*, 17(3):e72, March 2015. ISSN 1438-8871. doi: 10.2196/jmir.4167. URL <http://www.jmir.org/2015/3/e72/>.
- [64] Robert R Morris, Kareem Kouddous, Rohan Kshirsagar, and Stephen M Schueller. Towards an artificially empathic conversational agent for mental health applications: system design and user perceptions. *Journal of medical Internet research*, 20(6):e10148, 2018.
- [65] Katie Morton, Mark Beauchamp, Anna Prothero, Lauren Joyce, Laura Saunders, Sarah Spencer-Bowdage, Bernadette Dancy, and Charles Pedlar. The effectiveness of motivational interviewing for health behaviour change in primary care settings: a systematic review. *Health psychology review*, 9(2):205–223, 2015.
- [66] John A Naslund, Kelly A Aschbrenner, Lisa A Marsch, and Stephen J Bartels. The future of mental health care: peer-to-peer support and social media. *Epidemiology and psychiatric sciences*, 25(2):113–122, 2016.
- [67] Stefan Olafsson, Teresa O’Leary, and Timothy W. Bickmore. Coerced change-talk with conversational agents promotes confidence in behavior change. In Oscar Mayora, Stefano Forti, Jochen Meyer, and Lena Mamykina, editors, *Proceedings of the 13th EAI International Conference on Pervasive Computing Technologies for Healthcare, PervasiveHealth 2019, Trento, Italy, 20-23 May 2019*, pages 31–40. ACM, 2019. doi: 10.1145/3329189.3329202. URL <https://doi.org/10.1145/3329189.3329202>.
- [68] Christine A Padesky. Socratic questioning: Changing minds or guiding discovery. In *A keynote address delivered at the European Congress of Behavioural and Cognitive Therapies, London*, volume 24, 1993.
- [69] SoHyun Park, Jeewon Choi, Sungwoo Lee, Changhoon Oh, Changdai Kim, Soohyun La, Joonhwan Lee, and Bongwon Suh. Designing a chatbot for a brief motivational interview on stress management: Qualitative case study. *Journal of medical Internet research*, 21(4):e12231, 2019.
- [70] Pierre Philip, Jean-Arthur Micoulaud-Franchi, Patricia Sagaspe, Etienne De Sevin, Jérôme Olive, Stéphanie Bioulac, and Alain Sauteraud. Virtual human as a new diagnostic tool, a proof of concept study in the field of major depressive disorders. *Scientific reports*, 7(1):1–7, 2017.
- [71] Jan Pichl, Petr Marek, Jakub Konrad, Martin Matulık, Hoang Long Nguyen, and Jan Sedivy. Alquist: The alexa prize socialbot. *CoRR*, abs/1804.06705, 2018. URL <http://arxiv.org/abs/1804.06705>.

- [72] Sihang Qiu, Ujwal Gadiraju, and Alessandro Bozzon. Improving worker engagement through conversational microtask crowdsourcing. In Regina Bernhaupt, Florian 'Floyd' Mueller, David Verweij, Josh Andres, Joanna McGrenere, Andy Cockburn, Ignacio Avellino, Alix Goguey, Pernille Bjøn, Shengdong Zhao, Briane Paul Samson, and Rafal Kocielnik, editors, *CHI '20: CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, April 25-30, 2020*, pages 1–12. ACM, 2020. doi: 10.1145/3313831.3376403. URL <https://doi.org/10.1145/3313831.3376403>.
- [73] Sihang Qiu, Ujwal Gadiraju, and Alessandro Bozzon. Estimating conversational styles in conversational microtask crowdsourcing. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW1):1–23, 2020.
- [74] Stephen Rollnick, Nick Heather, and Alison Bell. Negotiating behaviour change in medical settings: the development of brief motivational interviewing. *Journal of mental health*, 1(1):25–37, 1992.
- [75] Sune Rubak, Anelli Sandbæk, Torsten Lauritzen, Knut Borch-Johnsen, and Bo Christensen. General practitioners trained in motivational interviewing can positively affect the attitude to behaviour change in people with type 2 diabetes: One year follow-up of an rct, addition denmark. *Scandinavian journal of primary health care*, 27(3): 172–179, 2009.
- [76] Shoetsu Sato, Naoki Yoshinaga, Masashi Toyoda, and Masaru Kitsuregawa. Modeling situations in neural chat bots. In Allyson Ettinger, Spandana Gella, Matthieu Labeau, Cecilia Ovesdotter Alm, Marine Carpuat, and Mark Dredze, editors, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Student Research Workshop*, pages 120–127. Association for Computational Linguistics, 2017. doi: 10.18653/v1/P17-3020. URL <https://doi.org/10.18653/v1/P17-3020>.
- [77] Theresa Schachner, Roman Keller, Florian von Wangenheim, et al. Artificial intelligence-based conversational agents for chronic conditions: systematic literature review. *Journal of medical Internet research*, 22(9):e20701, 2020.
- [78] Martin Schrepp, Andreas Hinderks, and Jörg Thomaschewski. Design and evaluation of a short version of the user experience questionnaire (ueq-s). *International Journal of Interactive Multimedia and Artificial Intelligence*, 4 (6), 103-108., 2017.
- [79] Daniel Schulman, Timothy W. Bickmore, and Candace L. Sidner. An intelligent conversational agent for promoting long-term health behavior change using motivational interviewing. In *AI and Health Communication, Papers from the 2011 AAAI Spring Symposium, Technical Report SS-11-01, Stanford, California, USA, March 21-23, 2011*. AAAI, 2011. URL <http://www.aaai.org/ocs/index.php/SSS/SSS11/paper/view/2401>.

- [80] Tianyuan Shi and Yongduan Song. A novel two-stage generation framework for promoting the persona-consistency and diversity of responses in neural dialog systems. *IEEE Transactions on Neural Networks and Learning Systems*, 2021.
- [81] Haoyu Song, Wei-Nan Zhang, Jingwen Hu, and Ting Liu. Generating persona consistent dialogues by exploiting natural language inference. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8878–8885. AAAI Press, 2020. URL <https://ojs.aaai.org/index.php/AAAI/article/view/6417>.
- [82] Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. A neural network approach to context-sensitive generation of conversational responses. In Rada Mihalcea, Joyce Yue Chai, and Anoop Sarkar, editors, *NAACL HLT 2015, The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado, USA, May 31 - June 5, 2015*, pages 196–205. The Association for Computational Linguistics, 2015. doi: 10.3115/v1/n15-1020. URL <https://doi.org/10.3115/v1/n15-1020>.
- [83] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. In Zoubin Ghahramani, Max Welling, Corinna Cortes, Neil D. Lawrence, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 3104–3112, 2014. URL <https://proceedings.neurips.cc/paper/2014/hash/a14ac55a4f27472c5d894ec1c3c743d2-Abstract.html>.
- [84] Benjamin Tolchin, Gaston Baslet, Steve Martino, Joji Suzuki, Hal Blumenfeld, Lawrence J Hirsch, Hamada Altalib, and Barbara A Dworetzky. Motivational interviewing techniques to improve psychotherapy adherence and outcomes for patients with psychogenic nonepileptic seizures. *The Journal of neuropsychiatry and clinical neurosciences*, 32(2):125–131, 2020.
- [85] Carlos Toxtli, Andrés Monroy-Hernández, and Justin Cranshaw. Understanding chatbot-mediated task management. In Regan L. Mandryk, Mark Hancock, Mark Perry, and Anna L. Cox, editors, *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, CHI 2018, Montreal, QC, Canada, April 21-26, 2018*, page 58. ACM, 2018. doi: 10.1145/3173574.3173632. URL <https://doi.org/10.1145/3173574.3173632>.
- [86] David R Traum. Computational models of grounding in collaborative systems. In *Psychological Models of Communication in Collaborative Systems-Papers from the AAAI Fall Symposium*, pages 124–131, 1999.



- 
- [87] Alan M. Turing. Computing machinery and intelligence. In Margaret A. Boden, editor, *The Philosophy of Artificial Intelligence*, Oxford readings in philosophy, pages 40–66. Oxford University Press, 1990.
- [88] Teun A Van Dijk. Comments on context and conversation. *Discourse and contemporary social change*, 54:281–316, 2007.
- [89] Alexandra Vtyurina, Denis Savenkov, Eugene Agichtein, and Charles L. A. Clarke. Exploring conversational search with humans, assistants, and wizards. In Gloria Mark, Susan R. Fussell, Cliff Lampe, m. c. schraefel, Juan Pablo Hourcade, Caroline Appert, and Daniel Wigdor, editors, *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, Denver, CO, USA, May 06-11, 2017, Extended Abstracts*, pages 2187–2193. ACM, 2017. doi: 10.1145/3027063.3053175. URL <https://doi.org/10.1145/3027063.3053175>.
- [90] Di Wang, Nebojsa Jojic, Chris Brockett, and Eric Nyberg. Steering output style and topic in neural response generation. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel, editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 2140–2150. Association for Computational Linguistics, 2017. doi: 10.18653/v1/d17-1228. URL <https://doi.org/10.18653/v1/d17-1228>.
- [91] Joseph Weizenbaum. Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45, 1966.
- [92] Sean Welleck, Jason Weston, Arthur Szlam, and Kyunghyun Cho. Dialogue natural language inference. In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 3731–3741. Association for Computational Linguistics, 2019. doi: 10.18653/v1/p19-1363. URL <https://doi.org/10.18653/v1/p19-1363>.
- [93] Xiaofeng Yang, Jian Su, Jun Lang, Chew Lim Tan, Ting Liu, and Sheng Li. An entity-mention model for coreference resolution with inductive logic programming. In *Proceedings of Acl-08: Hlt*, pages 843–851, 2008.
- [94] Akihiro Yorita, Simon Egerton, Jodi Oakman, Carina Chan, and Naoyuki Kubota. A robot assisted stress management framework: Using conversation to measure occupational stress. In *IEEE International Conference on Systems, Man, and Cybernetics, SMC 2018, Miyazaki, Japan, October 7-10, 2018*, pages 3761–3767. IEEE, 2018. doi: 10.1109/SMC.2018.00637. URL <https://doi.org/10.1109/SMC.2018.00637>.
- [95] Akihiro Yorita, Simon Egerton, Jodi Oakman, Carina Chan, and Naoyuki Kubota. Self-adapting chatbot personalities for better peer support. In *2019 IEEE International*

## BIBLIOGRAPHY

---

- Conference on Systems, Man and Cybernetics, SMC 2019, Bari, Italy, October 6-9, 2019*, pages 4094–4100. IEEE, 2019. doi: 10.1109/SMC.2019.8914583. URL <https://doi.org/10.1109/SMC.2019.8914583>.
- [96] Wei-Nan Zhang, Qingfu Zhu, Yifa Wang, Yanyan Zhao, and Ting Liu. Neural personalized response generation as domain adaptation. *World Wide Web*, 22(4):1427–1446, 2019. doi: 10.1007/s11280-018-0598-6. URL <https://doi.org/10.1007/s11280-018-0598-6>.
- [97] Peixiang Zhong, Chen Zhang, Hao Wang, Yong Liu, and Chunyan Miao. Towards persona-based empathetic conversational models. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 6556–6566. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.531. URL <https://doi.org/10.18653/v1/2020.emnlp-main.531>.
- [98] Victor W. Zue. Conversational interfaces: advances and challenges. In George Kokkinakis, Nikos Fakotakis, and Evangelos Dermatas, editors, *Fifth European Conference on Speech Communication and Technology, EUROSPEECH 1997, Rhodes, Greece, September 22-25, 1997*. ISCA, 1997. URL [http://www.isca-speech.org/archive/eurospeech\\_1997/e97\\_KN09.html](http://www.isca-speech.org/archive/eurospeech_1997/e97_KN09.html).
- [99] Victor W. Zue and James R. Glass. Conversational interfaces: advances and challenges. *Proc. IEEE*, 88(8):1166–1180, 2000. doi: 10.1109/5.880078. URL <https://doi.org/10.1109/5.880078>.

## Appendix A

# System Implementation Elements

### A.1 Entity Relation Diagram

Figure A.1 shows how the various entities in the database implemented in this project are related.

### A.2 Dialog Data

This is the dialogue used in our study, which is adapted from the original conversation<sup>1</sup> between a depressed user and a therapist. We name the depressed user as “User” and the

<sup>1</sup><https://www.guilford.com/add/miller2/julia.pdf>

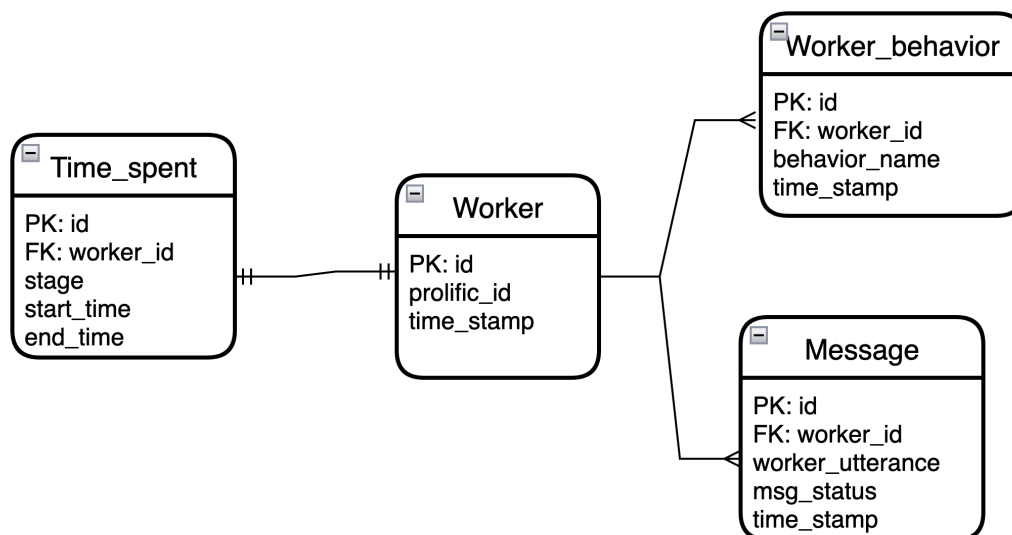


Figure A.1: Entity relation diagram for the database. “PK” means the primary key in the table while “FK” means the foreign key in the table.

## A. SYSTEM IMPLEMENTATION ELEMENTS

---

therapist as “Bot” to match our experimental purpose. Three bold sentences are chosen as the target utterances for early, middle, and late entry points, respectively.

USER: I feel like I am falling apart. I don't have any energy. I don't know what is going on with me.

BOT: You are feeling upset and confused. One confusing thing, then, is you don't understand why you're doing what you're doing. What's been happening?

USER: I just broke up with my boyfriend. I mean we've been living together and I thought he loved me, but he's just so distant. He won't talk to me, and I think maybe he's seeing someone else. Anyhow, he told me I'm crazy.

BOT: This blowup wasn't the first time that's happened to you.

USER: Men just drive me crazy. This is the third guy I've lived with, and they all kind of ended in the same way. I just seem to fall in love with the wrong guys.

BOT: Kind of like a pattern that's repeating itself.

USER: **Yes! It is so sad.**

BOT: These are really strong feelings. It's pretty painful that this has happened once again.

USER: I can't sleep. I can't think. I'm a mess at work. I was waiting on a customer this week and just started crying for no reason. I think I'm losing it.

BOT: That frightens you, not understanding what is going on with you.

USER: It's just so discouraging! I was so happy with Ray when we were first together. There's this soft teddy bear inside his tough exterior, and that's the man I loved, but then he wouldn't open up to me anymore.

BOT: And that annoyed you.

USER: Yes! It's such a waste for him to stay locked up inside there, and I was lonely even though we were living together. Anyhow, he's gone now. He moved out. It's over.

BOT: All right. What you've told me so far is that these really strong feelings, and how you've reacted, are scaring you a little. You're not sure what's going on. You're having trouble sleeping and concentrating at work. You really blew up at Ray and felt out of control, and you're also wondering about what seems like a painful pattern that repeats itself in your relationships with men. You don't have much energy, feel lonely, and sometimes you just break out crying for no apparent reason. But you especially wonder what's happening.

USER: Yes. Do you think you can help me?

BOT: Yes, I do. This all feels pretty strange to you, even coming here, but I've worked with women before who have had concerns like this, and I believe I can help. If our work together were really successful from your perspective, what would be different?

USER: I guess I wouldn't feel so bad all the time.

BOT: So one thing you'd like to change is how you're feeling. Tell me a little more about that.

USER: I just feel upset and I'm crying a lot. I'm not sleeping and I feel worn out, run down.

BOT: OK, you'd like to get your emotional life settled down some, to be able to sleep better and have more energy. What else?

USER: Yes. I also want to know why I keep screwing up all my relationships. What's wrong with me?

BOT: That's another thing that upsets you—not knowing why you feel so bad and why these things happen to you.

USER: Yes, can you help me figure these out?

BOT: I do have an idea that puts some of the pieces of the puzzle together. If it's alright with you, we can talk about that next.

USER: Sure. What is your idea?

BOT: You're really struggling to understand what is happening to you, and I wonder if I might ask you what you already knew about depression.

USER: I guess it's like when you feel really sad and down, maybe don't have energy to do anything. Do you think that's what I have?

BOT: Well, if it's all right, let me describe some of what people experience with depression, and you can tell me what parts of this may fit for you.

USER: Yes, OK.

BOT: Depression is really a set of different symptoms, and you don't need to have all of them. It's like when people catch a cold; they experience it in different ways. Depression is like that—a set of symptoms that might or might not be present. Does that make sense?

USER: Yes, what are the symptoms?

BOT: One of them, as you said, is in a sad mood, feeling down, crying.

USER: That sounds like me. I'm not having much fun lately.

BOT: Another change is in sleeping patterns. Some people have trouble sleeping. Some lose their weight. Have you experienced any of those?

USER: I'm certainly not sleeping well. But my weight did not change.

BOT: OK. As I said, different people have different symptoms. It sounds like you have quite a few of them. Have I been clear? What else can I tell you about depression?

USER: That sounds alright.

BOT: The next thing is how you would like things to be different. You know how you don't want to feel. How do you want to feel?

USER: Normal, I guess. Happy. To have energy to do things again. When I broke up with Ray it just made me crazy. I feel like there's something wrong with me, that I always screw up my relationships.

BOT: You'd like to feel happy and good about yourself again. What else?

USER: I want to be with a man who loves me. I seem to attract guys who are hung up about telling me how they feel. I need a man I can talk to.

BOT: How important is that to you, to have a relationship like that?

USER: Very important. I don't want to be alone. I need to be loved.

BOT: You need that.

USER: Yes! I don't want to keep destroying relationships. I don't know why I do that.

## A. SYSTEM IMPLEMENTATION ELEMENTS

---

BOT: Tell me a little about why you want to feel better.

USER: I just feel like I'm dragging around this heavy weight with me all the time. I like to have fun, but I've really become a drag to be with. I feel like even my friends avoid me.

BOT: It would be good to feel lighthearted, to enjoy life and be with your friends.

USER: **Yes it would. Do you think it's possible for me?**

BOT: Well I was just going to ask you about that. What are some of your personal strengths? What might your friends say that you have going for you?

USER: I don't know. They'd probably say that I'm stubborn. I've been pretty persistent in the past.

BOT: Give me an example. When have you done something or made a change in your life that really took some effort, maybe something you weren't sure at first that you could do?

USER: Moving here. I had always lived in Ireland near my father and my sisters, and I moved out here all by myself. I just wanted to be on my own for a change, to get away. But now I feel too much on my own.

BOT: It takes a lot to move to a new place on your own. How did you do it?

USER: I had to find a job here, and I started going to classes at the university. I had to get used to a different culture, figure out the stores and banking system, and make some friends.

BOT: And you did it. That's a lot of change to manage.

USER: I never thought much about it. I guess I can do it when it's important enough.

BOT: So use your imagination here. Suppose that we work together and you are successful in making these changes. How might your life be different, say, 5 years from now?

USER: I'd be married, maybe have a family. I'd have a better job that I enjoy more. I just wouldn't be stressing out like this all the time. A calmer life.

BOT: Well, thanks for all you've told me. You've been feeling a lot of turmoil in your life and I appreciate how honest and open you have been with me. Let me just pull together what you've told me, and then we'll see what the next step is. You would really like to start feeling better, you've been feeling really bad, like you have been dragging a great weight around with you, and you're rather tired of feeling like your friends avoid you. In the longer run you also want to have an open, warm, and loving relationship with a man. You're a pretty resourceful person. You moved here from Ireland all on your own and set up a new life for yourself. You're also a stubborn and persistent person. Once you make up your mind to do something it's likely to happen. You would like to understand what's been going wrong in your relationships because somewhere down the line you would like to be married, and you also mentioned finding a better job as a longer-term goal. You've been through a lot of emotional pain, and you're eager for a calmer and happier life. First of all, I would like to know what ideas you have for how you might start feeling better. No one knows you better than you do, and I'm sure you have tried some things in the past to lift your mood a bit. Tell me about those.

USER: Sometimes I have gone to a funny, romantic movie and it makes me smile, but it also reminds me of what I don't have in a relationship.

BOT: A pleasant movie is one thing that can lift your spirits sometimes. What else?

USER: Getting out of my apartment. If I just sit there watching television with the curtains drawn, that's not good for me.

BOT: You know that about yourself—getting out helps. And what do you do when you leave your apartment?

USER: **I might just take a walk or see my friends. But like I said, it seems like they don't want to be around me so much anymore because I bring them down with me. Do you have some suggestions on what I should do?**

BOT: Yes, let's consider some options. You already know some things that have helped lift your mood in the past, like getting out of your apartment to take a walk, see friends, or go to a pleasant movie. I'm very interested in your own hunches about what you need and what will help you, so let's talk about some possibilities. What do you already know about other treatments used to make you feel better?

USER: Not much. I've seen ads for pills. What others do you have?

BOT: It seems that you have wondered if medication would help you. Your strongest hunch also seemed to be about how you run yourself down in your mind—things you tell yourself that get you feeling worse about yourself. A treatment that helps with this is called cognitive therapy.

USER: I definitely do that.

BOT: Great. One plan, then, could be to start with cognitive therapy and see how that goes for you. We can always keep other options open depending on your experience.

USER: That sounds fine to me.

BOT: Are you willing to do that—come once or twice a week, work together for about 2 months, and see how it goes?

USER: Yes definitely.





## Appendix B

---

# Consent Form and Questionnaires

In this appendix, we give an overview of the questionnaires used to evaluate ContextBot, including three versions (MI, Non-MI, History) of consent forms in the pre-task questionnaire, the post-task questionnaire and the evaluation questionnaire provided for crowd workers and two professional psychologists to evaluate the response quality.

### B.1 Consent Form: MI Condition

Crowd-powered systems combine computation with human intelligence, drawn from large groups of people connecting and coordinating online. Recently, researchers have empowered the text-based messaging and chatbots systems with crowd support in real-time. These are known as crowd-powered conversational systems (CPCS). In the CPCS, multiple humans are involved in generating chatbot responses in real-time. Then either computational intelligence or human evaluators are employed to choose the best quality response in real-time. We have implemented CPCS in this study under the context of supportive conversations.

1. After signing this digital consent form, you will be redirected to a web page where you will see detailed instructions on how to use the system. You need to play the role of a coach to give responses to the depressed user in a text-based supportive conversation.

2. After reading the instructions, you will be redirected to another web page where you are required to respond to the user. You can ask for help from the chatbot if you find it hard to respond to the user. Furthermore, the chatbot will give you specific tips that you might need to improve your response quality.

3. Then, you will be asked to fill out a survey form.

4. Finally, you will be redirected back to the completion URL on Prolific.

#### Confidentiality

Data will be handled anonymously and only the researchers of this study will have direct

access to it. We will not store any personal information that will allow participants to be identified from their data. We will only gather basic demographic (age, gender, qualifications). This data is provided by prolific by default for each study.

By checking this you have decided to volunteer as a research participant for this study, and that you have read and understood the information provided above.

Note: It is up to you to decide whether to take part or not; choosing not to take part will not disadvantage you in any way. You are allowed to leave the study even during the execution of the study.

Do you consent to this study?

Yes, I do.; No, I don't.

### **B.2 Consent Form: Non-MI Condition**

Crowd-powered systems combine computation with human intelligence, drawn from large groups of people connecting and coordinating online. Recently, researchers have empowered the text-based messaging and chatbots systems with crowd support in real-time. These are known as crowd-powered conversational systems (CPCS). In the CPCS, multiple humans are involved in generating chatbot responses in real-time. Then either computational intelligence or human evaluators are employed to choose the best quality response in real-time. We have implemented CPCS in this study under the context of supportive conversations.

1. After signing this digital consent form, you will be redirected to a web page where you will see detailed instructions on how to use the system. You need to play the role as a coach to give responses to the depressed user in a text-based supportive conversation.

2. After reading the instructions, you will be redirected to another web page where you are required to respond to the user. You can ask for help from the chatbot if you find it hard to respond to the user.

3. Then, you will be asked to fill out a survey form.

4. Finally, you will be redirected back to the completion URL on Prolific.

#### Confidentiality

Data will be handled anonymously and only the researchers of this study will have direct access to it. We will not store any personal information that will allow participants to be identified from their data. We will only gather basic demographic (age, gender, qualifications). This data is provided by prolific by default for each study.

By checking this you have decided to volunteer as a research participant for this study, and that you have read and understood the information provided above.

Note: It is up to you to decide whether to take part or not; choosing not to take part will not disadvantage you in any way. You are allowed to leave the study even during the execution of the study.

Do you consent to this study?

Yes, I do.; No, I don't.

### **B.3 Consent Form: History Condition**

Crowd-powered systems combine computation with human intelligence, drawn from large groups of people connecting and coordinating online. Recently, researchers have empowered the text-based messaging and chatbots systems with crowd support in real-time. These are known as crowd-powered conversational systems (CPCS). In the CPCS, multiple humans are involved in generating chatbot responses in real-time. Then either computational intelligence or human evaluators are employed to choose the best quality response in real-time. We have implemented CPCS in this study under the context of supportive conversations.

1. After signing this digital consent form, you will be redirected to a web page where you will see detailed instructions on how to use the system. You need to play the role of a coach to give responses to the depressed user in a text-based supportive conversation.

2. After reading the instructions, you will be redirected to another web page where you are required to respond to the user.

3. Then, you will be asked to fill out a survey form.

4. Finally, you will be redirected back to the completion URL on Prolific.

#### **Confidentiality**

Data will be handled anonymously and only the researchers of this study will have direct access to it. We will not store any personal information that will allow participants to be identified from their data. We will only gather basic demographic (age, gender, qualifications). This data is provided by prolific by default for each study.

By checking this you have decided to volunteer as a research participant for this study, and that you have read and understood the information provided above.

Note: It is up to you to decide whether to take part or not; choosing not to take part will not disadvantage you in any way. You are allowed to leave the study even during the execution of the study.

Do you consent to this study?

Yes, I do.; No, I don't.

## B.4 Post-Task Questionnaire Items

1. Did you ask ContextBot for help before responding to the user? (multi-choice question)

2. Please answer the following questions about your previous experience with chatbots. (7-point Likert scale)

- I am familiar with chatbot technologies.
- I use text-based chatbots frequently.
- I like getting help from a chatbot when I have trouble online.

3. Please answer the following questions about your impressions of ContextBot. (7-point Likert scale)

- The ContextBot has explained every context well for me to understand.
- The ContextBot is coherent and maintains a clear conversational flow.
- The ContextBot has provided easy-to-use quick reply buttons.
- When using the ContextBot, I feel in control.
- I find that the highlighted texts help me understand the chat history.
- I find that the summary of the user helps me understand the context better.
- I find that the collapsible section of showing intentions helps me understand the context better.

4. If you have any additional thoughts about your experience with ContextBot, please share them with us. (open-ended question)

5. Decide as spontaneously as possible which of the following conflicting terms better describes the system design. There is no "right" or "wrong" answer. Only your personal opinion counts! (7-point bipolar scale)

- obstructive-supportive
- complicated-easy
- inefficient-efficient
- confusing-clear
- boring-exciting
- not interesting-interesting
- conventional-inventive
- usual-leading edge

6. Choose your response in the slider below according to your impression of the task load. (range from 1 to 100)

- How mentally (e.g. thinking, calculating, deciding, remembering, searching, etc) demanding was the task?
- How physically (e.g. pushing, pulling, controlling, turning, etc) demanding was the task?
- How hurried or rushed was the pace of the task?
- How successful were you in accomplishing what you were asked to do?
- How hard did you have to work to accomplish your level of performance?
- How insecure, discouraged, irritated, stressed, and annoyed were you?

7. Indicate your overall satisfaction with our system on a 10-point scale ranging from (1) 'very dissatisfied' to (10) 'very satisfied'.

8. Please share any additional thoughts, remarks, or feedback that you may have regarding your experience interacting with our system. (open-ended question)

9. Did you feel that the ContextBot you interacted with was pre-programmed by people? (multi-choice question)

## **B.5 Evaluation Questionnaire for Response Consistency**

You are required to rate the consistency of 25 responses.

In terms of consistency, you can pay attention to the following items:

- (1) if the response is related to the user's concern;
- (2) if the events referred by the response have been talked about in the dialog;
- (3) if the persons referred by the response have been talked about in the dialog;
- (4) if the response is natural without being obtrusive when you put it into the dialog, trust your intuition as a human :)

We display 10 responses per page. The dialog and criteria will be placed at the top of the page for you to look up.

**Q:** For each response, how consistent do you think it is with the history context of the dialog? ("1" : highly inconsistent; "7": highly consistent)

**Q:** What is your overall opinion of the consistency of the responses you've just reviewed? Are there any good/bad examples that you can point to?

## **B.6 Evaluation Questionnaire for Professionalism in Responses**

In the next page, we will first show you a snippet of dialog which happened between a depressed user and a professional psychologist who was using Motivational Interviewing (MI) as the therapy method.

## B. CONSENT FORM AND QUESTIONNAIRES

---

Next, we will show you 25 candidate responses collected from different crowd workers who played the role as a therapy coach. They provided the response towards the user's last utterance.

We would like to know how professional these responses are. You are required to rate each response based on your own professional opinion. We've also included some criteria below for your reference if you find them useful.

**Condition #1 (for the early entry point)** (1) It shows empathy to the user;

(2) It provides reflective listening to the user's concerns;

(3) It affirms the user's effort of expressing herself in a positive way;

(4) It helps engage the user in this dialog session.

**Condition #2 (for the half entry point)** (1) It provides reflective listening to the user's concerns;

(2) It asks open questions to elicit the user's ability of change;

(3) It affirms the user's effort of expressing herself in a positive way;

(4) It helps evoke the user's intention of changing in this dialog session.

**Condition #3 (for the late entry point)** (1) It summarizes the user's concerns discussed in the previous dialog;

(2) It provides reflective listening to the user's concerns;

(3) It asks open questions to elicit the user's ability of change;

(4) It affirms the user's effort of expressing herself in a positive way;

(5) It helps the user form a plan to change her depression situation in this dialog session.

(For a better reading experience, we display 10 responses per page. The dialog, criteria and standard response will be placed at the top of each page for you to look up.)

**Q:** For each response, what do you think about the quality of the response? (1: highly unprofessional; 7: highly professional).

**Q:** What is your overall opinion of the quality of the responses you've just reviewed? Are there any good/bad examples that you can point to. In their responses, did any of them adopt solid MI-adherent practices? What do you believe could be improved? (open-ended question)