ContextBot

Improving Response Consistency in Crowd-Powered Conversational Systems for Affective Support Tasks

Ma, Yao; Abbas, Tahir; Gadiraju, Ujwal

DOI
10.1145/3603163.3609031

Publication date
2023

Document Version
Final published version

Published in
HT 2023 - The 34th ACM Conference on Hypertext and Social Media

ContextBot: Improving Response Consistency in Crowd-Powered Conversational Systems for Affective Support Tasks

Yao Ma
Delft University of Technology
Delft, The Netherlands
yaom1886@gmail.com

Tahir Abbas
Delft University of Technology
Delft, The Netherlands
t.abbas-1@tudelft.nl

Ujwal Gadiraju
Delft University of Technology
Delft, The Netherlands
u.k.gadiraju@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. 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 3×3 factorial between-subjects study. Experimental conditions varied based on (i) whether or not context was elicited and informed by *motivational interviewing* techniques (MI-adherent guidance, general guidance, and no guidance), and (ii) different conversational entry points for workers to produce responses (early, middle, and late). Our findings show that: (a) workers who entered the conversation earliest were more likely to produce highly consistent responses after interacting with *ContextBot*; (b) showed better user experience after they interacted with *ContextBot* with a long chat history to surf; (c) produced more professional responses as endorsed by psychologists; (d) and that interacting with *ContextBot* through task completion did not negatively impact workers' cognitive load. Our findings shed light on the implications of building intelligent interfaces for scaffolding strategies to preserve consistency in dialogue in CPCS.

CCS CONCEPTS

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

KEYWORDS

Crowd-powered Conversational Systems, Motivational Interviewing, Dialogue Context, Chatbots, Real-time Crowdsourcing

HT '24, September 4–8, 2023, Rome, Italy

ACM ISBN 979-8-4007-0232-7/23/09.
https://doi.org/10.1145/3603163.3609031

ACM Reference Format:
Yao Ma, Tahir Abbas, and Ujwal Gadiraju. 2023. *ContextBot*: Improving Response Consistency in Crowd-Powered Conversational Systems for Affective Support Tasks. In *34th ACM Conference on Hypertext and Social Media (HT '23), September 4–8, 2023, Rome, Italy.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3603163.3609031

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 [27, 35]. 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 [35]. Evorus builds on Chorus by adding an automated module to select high-quality responses from workers [25]. 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 [26]. 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.

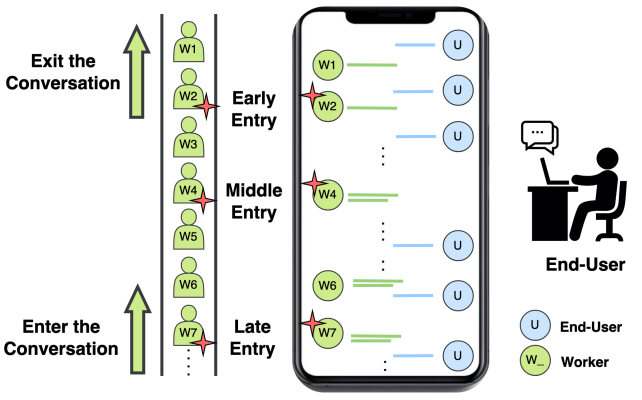

Figure 1: Conversation flow between an end-user and crowd workers in CPCS. New workers constantly enter and exit the conversation. Their entry points (early, middle, and late) directly affect the number of historical contexts that workers need to understand before coherently and effectively responding to the end-user.

Meanwhile, the time spent 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 illustrates the conversation flow between an end-user and crowd workers. 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 [35]. 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 procedural way. Another concern is that for affective crowdsourcing where emotional intelligence of crowds is applied to peer-to-peer mental health support [43], workers are involved in counselling conversations that demand higher comprehension than information retrieval tasks. Existing crowd-powered systems about affective support focus more on the effectiveness of counselling strategies [45] and the availability of real-time recruitment [2], thus ignoring that inconsistent worker responses may fail to relate to users' feelings and problems mentioned in the chat history. This can be particularly harmful in online psychological interventions.

We investigate how to systematically provide context with a text-based conversational interface driven by a chatbot to help new workers quickly grasp key information from an affective support conversation in a procedural way. 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 [39]. Researchers have explored the efficacy of chatbots in training workers to complete therapeutic tasks [3], demonstrating the potential of conversational interfaces in taking on pragmatic roles. Inspired by [53, 54], we developed a chatbot called *ContextBot* to help workers systematically track context about the current dialogue without recording it themselves.

Our goal in this paper is to address response inconsistency in affective support tasks in CPCS with a context-tracking tool mediated by a conversational interface. We employed an affective crowdsourcing task, which required workers to deliver affective support by *Motivational Interviewing (MI)* [42]. To this end, workers used a client-centred conversation to stimulate users to change. 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. Therefore, we decomposed context into multiple dimensions to model the chat history systematically. In CPCS, new crowd workers are required to participate in the discussion at any stage; it might be an earlier stage when the user is just joining CPCS and revealing his problems (early entry point). A worker can participate in the midst when crowd workers pose well-targeted inquiries to analyze users' concerns further (middle entry point). Or, it may occur in the last stages, when the crowd has already examined and comprehended the user's challenges and is now offering pragmatic steps based on the user's skills (late entry point). It is essential to understand how effectively an intelligent interface can provide contextual cues to workers based on the conversation stage in CPCS. We anticipate that a *ContextBot* can be more effective in assisting workers in the middle and late entry points when they must analyze a significant amount of information to make sense of a continuing conversation.

In addition, since we employed stress management tasks, we also investigated the efficacy of providing suggestions based on psychological interventions from Psychology (such as Motivational interviewing (MI) [42]) to guide workers toward more consistent answers. These counseling strategies offer specific recommendations for connecting effectively with users and what types of inquiries to ask at each conversational stage (e.g., focusing or evoking questions). Therefore, an intelligent interface that can guide workers based on the counseling stage may be more effective and assist workers in generating more consistent and high-quality responses. Thus, we used three different variants of *ContextBot*; one based on MI skills, one based on general guidelines (e.g., being compassionate with users), and one that does not involve any help or indication about affective support.

We designed a 3×3 between-subjects study to understand the role of contextual guidance in shaping the consistency of responses in CPCS. We considered three entry points (early, middle, late) and three types of contextual guidance (MI-adherent guidance, general guidance, and no guidance). All data pertaining to our research is accessible from this anonymous link.[1] We addressed the following research question — **RQ:** How does *ContextBot* used for providing context in CPCS affect the response quality and interaction experience of workers in emotion-centric dialogues?

- 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.

2 BACKGROUND AND RELATED WORK

2.1 Crowd-Powered Conversational Systems

Our research is closely related to solutions for providing context in CPCS, which allow workers to help users with online tasks through an interactive conversation. Chorus enabled a group of synchronized workers to help users retrieve information in a conversational interface [35]. New workers were guided to a "Working Memory"

[1] https://osf.io/wr4ea/?view_only=603a84a247ca423ebaf90db8a137e608

section for important aspects of the current conversation. A voting mechanism was also developed to guarantee consistent and most agreed-upon responses were selected to users. Evorus further included a machine learning module to automatically select responses from workers [25], thereby reducing real-time latency and improving output qualities. Guardian was a spoken dialogue system that asked crowd workers to identify parameters from web APIs and convert the query results to responses [27]. A similar voting system was implemented to help achieve a consensus on responses among active workers. InstructableCrowd was a dialogue system where synchronized workers collectively programmed IF-THEN rules based on users' needs [24]. Their interface and experiments also followed the setup in Chorus. Although the above systems are designed for information retrieval tasks, it is evident that combining the voting mechanism and the memory board maintained manually by workers is still the primary approach to providing context in the current CPCS.

Applications of affective crowdsourcing [43] extend the use of crowd-powered systems to online psychological interventions. They solicited the collective emotional intelligence of crowds and implemented the concept of peer-to-peer mental health support [48], which has been demonstrated in systems such as in-person and online support forums [5]. The web-based system, Panoply, trained crowd workers to perform cognitive reconstructing and generate new appraisals for users who sought emotional help [45, 46]. KokoBot further extended the idea and acted as a computer agent to guide users by predicting appropriate responses from an existing crowdsourced peer support corpus [44]. However, such research required users to log in to the system or get help from peers through agents [34]. 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 [2]. These chatbot-based affective support systems have become important alternatives to traditional one-to-one health consultations due to their cost-effective advantages such as the 24-hour accessibility, privacy and anonymity. 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 [26, 35], addressing the challenge of uncovering important context from the lengthy chat history while not burdening workers in affective support.

2.2 Context in Conversations

The concept of "context" was originally used in linguistics to refer to accompanying text [19], 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 [8], namely the linguistic, semantic, cognitive, social and physical context. Different context dimensions contribute to the pragmatic knowledge [60] about the current situations and mutual understanding for both the speaker and the addressee [11].

Context modelling in current dialogue systems models the context as persona-specific statements [36, 64] or the entire chat history. Despite its simplicity from a computational perspective, it cannot comprehensively link the current utterance with multi-dimensional contexts, such as semantic facts and speaker intentions. Moreover, both participants in dialogue usually perform specific actions to pursue a communicative goal [40] while existing research in automated dialogue systems ignores the contextual information that drives the goal. We decide to guide workers to explore the conversation by our contextual guidance, where specific cognition of the conversation from them is expected.

2.3 Motivational Interviewing

Motivational Interviewing (MI) is a client-centred, collaborative conversation widely used in behaviour change [7, 47] and psychotherapy [58]. Miller and Rollnick described four basic processes for using MI: engaging, focusing, evoking and planning [42]. Each stage requires a different amount and type of context to be focused on and elicited. In engaging, the therapist develops a safe, interactive environment where the patient can seek support. In focusing, the therapist attempts to work with the patient to establish a central goal and focus of the conversation. In evoking, the therapist uses various questions to explore the deep reasons for the change. In planning, the therapist helps the patient decide on an associated action plan to change. Micro-skills including open questions, affirmations, reflective listening and summaries (OARS) are commonly used to facilitate a deep understanding of the patient and help move the conversation forward.

Previous research has shown that MI could be integrated into chatbots to deliver psychological interventions where promising users' acceptance and perceived enjoyment have been achieved [37], but more contextualized responses were still required [50]. Therefore, we adopted MI-adherent guidance in our crowd-powered system to emphasize contextual understanding when workers responded to users with empathy.

2.4 Conversational Interfaces in Crowdsourcing

Text-based conversational interfaces in microtask crowdsourcing have been effectively used in various tasks such as image annotation and information finding [39], yielding comparable results and satisfaction from workers compared to traditional web interfaces. Crowdsourced studies in decision-making tasks with conversational interfaces have also been shown to yield relatively more user trust [21]. The impact of conversational styles on the output quality, user engagement, and cognitive load was investigated in conversational interfaces [53]. The results revealed that with a more enthusiastic conversational style, workers generated better output and experienced less cognitive load in more difficult microtasks [52]. Worker avatars and metaphors for conversational agents have also been proposed to increase worker engagement with varying effects across tasks [31, 51]. Besides traditional tasks, Trainbot adopted conversational interfaces to train workers on MI to provide emotional support for stressed users [3]. Workers reported less stress and performed better in answering quizzes when taking the test. Building upon such prior work [3, 39], we introduced the conversational interface as a tool for providing context and investigated how

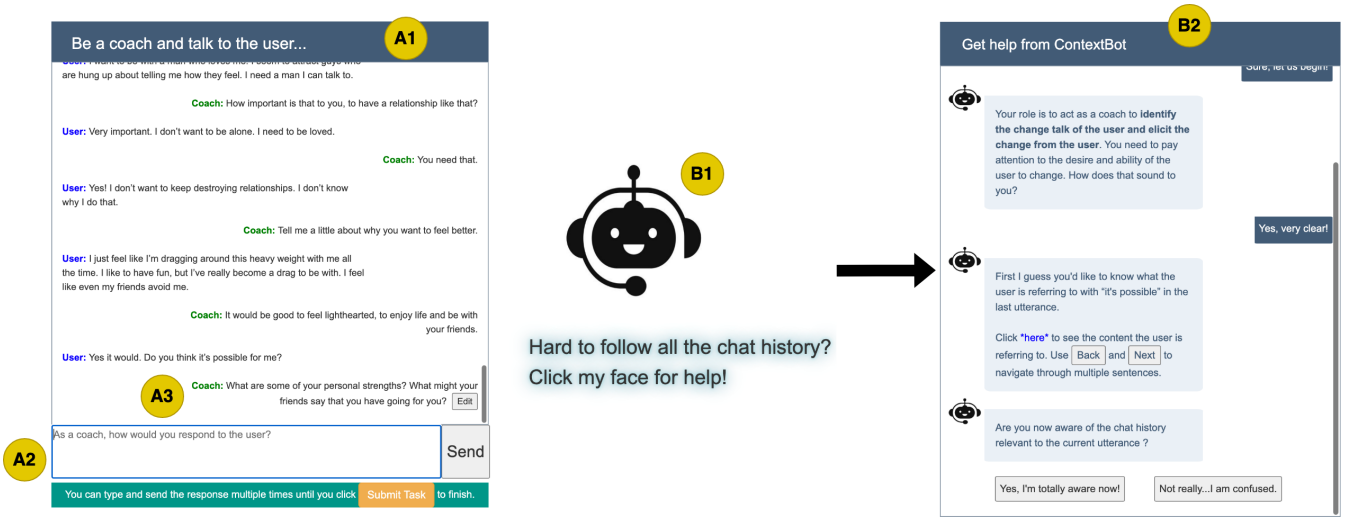

Figure 2: Conversational interfaces of the system with the task panel on the left and assistance from *ContextBot* on the right.

the interface affected the output quality and interaction experience in the microtask of delivering support in MI.

3 METHOD

To answer our research questions, we developed different experimental conditions and manipulated *ContextBot* to provide dialogue context within CPCS systematically. As described earlier, we selected a counselling dialogue using MI as the treatment to verify the role of the system in delivering affective support. Our study received approval from our institutional ethics review board.

3.1 Task Design

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. Figure 2 shows the system interfaces. 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. Workers then compose a message and send it to the user based on the guidance from *ContextBot*. Due to the scripted nature of the task, we allowed one response from workers; however, since humans tend to convey their message in fragments through multiple turns, we allowed them to edit their answers multiple times A3 and only submit the task when satisfied. 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.

3.2 Design of *ContextBot*

We designed *ContextBot* based on *Chatbot-driven* dialogue instead of *User-driven* dialogue [16]. A chatbot-driven dialogue function with a highly specified interaction design, i.e., the interaction is primarily directed or driven by the chatbot. This strategy is beneficial for tasks with a transactional nature, in which we want the chatbot to persuade users to accomplish a goal. On the other hand, user-driven dialogues allow for a greater variety of user input options and are more sensitive to fluctuations in user input. This latter style is more suitable for small social talk than the former. Therefore, for this study, the locus of control was transferred to *ContextBot* by providing workers with a limited number of options for standard content and by employing scripted dialogue. The scripted nature of the dialogue was desirable for this controlled study, where we wish to account for differences in user input and *ContextBot* instruction. In the following sections, we describe how we structured the dialogue's content.

3.2.1 Conceptualizing Context. Different context dimensions are related to each other and together constitute the factors for comprehending the current sentence. We consider four dimensions of context mentioned in earlier work [8]: 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. Figure 3 provides an overview of all dimensions of context.

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. It is placed in the first step of interaction to give global goals (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 can be elucidated by directly giving the previous sentences, which contain events, places, and pronouns that have been referred to in the current sentence. The authors of our paper jointly determined the relevant linguistic context that should be extracted from the selected current utterance. The criteria include, (i) finding an unspecified noun or pronoun in the current utterance; (ii) and

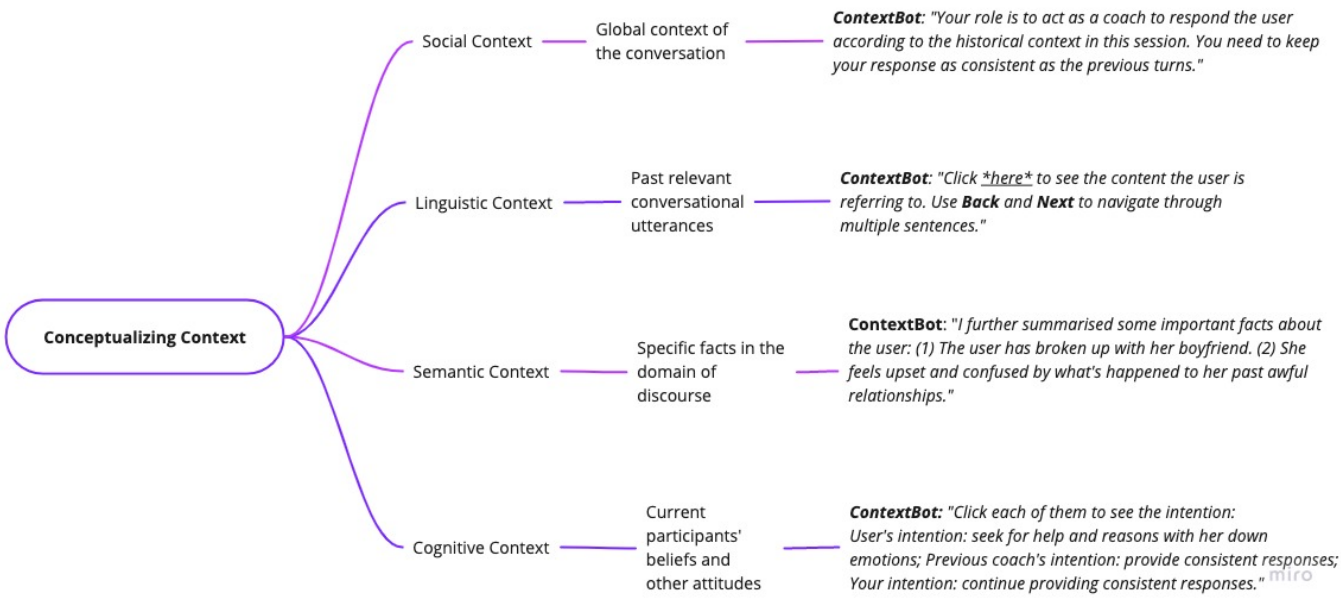

Figure 3: This figure illustrates a summary of different dimensions of context based on the work of Bunt [8]. The first node describes the context, the next node briefly explains it, and the final node provides examples of how we operationalized it in *ContextBot*.

identifying sentences related to the unspecified word in the preceding text. Finally, all of the authors reached an agreement on the identified sentences as linguistic context.

Semantic Context is known to be "specific facts in the domain of discourse; the current state of the underlying task" [8]. 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. We manually summarized the corresponding number of facts based on the position of the current utterance in the chat. Text summarization can be performed automatically by advanced natural language processing tools [18]. However, this was not the current focus of our research.

Cognitive Context reflects the intention and attentional state of participants toward the current utterance. Grosz and Sidner proposed the attentional state as the "information about the objects, properties, relations, and discourse intentions that are most salient at any given point" [20]. The attentional state of a local utterance has very close relations to the cognitive context, where "current participants' beliefs, intentions, and other attitudes" are reflected [8].

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 explicitly summarized 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 modified the templates previously used in [3, 50] and added instructions relevant to the corresponding MI stage.

3.2.2 Mapping Conversation Flow. We aim to provide contextual information in a procedural 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. We chose the specific sequence (social → linguistic → semantic → cognitive) based on the following reasons. We draw on the idea of analyzing the context from both global and local perspectives in the work of [8]. For instance, it was vital to provide the *social context* up front to familiarize workers with the broader objectives of interacting with ContexBot. Then, based on past work in the CPCS [35], we supplied *linguistic context*

that reveals the most important lines from previous conversations. This allows workers to be directed to the most crucial portions of a conversation without having to go through long conversational history. Providing social and linguistic context may be sufficient for shorter conversations with a limited number of facts. Still, for longer and more complex discussions, such as emotion-centric dialogues, workers must comprehend what has been talked around the most significant conversational exchanges. When workers read, "I simply feel like I'm dragging around this heavyweight with me all the time," they cannot determine what is wrong with the person. Consequently, it is essential to summarize a list of stressors that induce stress (e.g., "The user has broken up with her boyfriend"). Therefore, we added *semantic context* next to the linguistic context. Finally, the preceding three contexts train a worker for consistent response. Still, they do not account for differences in the workers' beliefs, which might result in bias and inconsistent responses. Therefore, we added *cognitive context* to address this. Please note that the precise order and content was decided after multiple iterations of testing and brainstorming by the authors of this work to arrive at the final better sequence.

ContextBot adopts a rule-based method in selecting and generating answers [28]. Three components are taken into account to program the process: a responder to control the interface, a classifier to group the input, and a graphmaster for information storage [4]. The interface receives the worker's input, which is then parsed and matched with the predefined pattern designed for determining whether the input is positive or negative to the previous question. Finally, the corresponding responses from a predefined static knowledge base are selected and sent to the interface. More advanced techniques such as AIML (Artificial Intelligence Mark-up Language) can be added to build a conversation flow allowing more flexible inputs from workers [28].

❶ 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. ❷ After receiving the acceptance, *ContextBot* first introduces the global social context, including what role the worker should generally act as and the corresponding requirement 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 [40], we designed 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 [59]. ❸ Next, *ContextBot* prompts the guidance of exploring linguistic context. Another turn of rephrasing is provided as well. ❹ After the confirmation, the worker is prompted to read the summarized semantic context. ❺ 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. ❻ The final prompt for MI techniques comes after the acknowledgement of the current information from the worker.

3.2.3 Rationale for Choosing Stress Management task. We focus on stress management tasks since stress-related disorders are becoming more prevalent [12]. Therefore, we must support this societal need by building technological innovations. Given the limitation of AI, improving the consistency of responses in mental-health CPCS is desirable. Secondly, we added the stress task because of its open-ended nature. These types of tasks do not have an objective answer and require answering user complaints by comprehending complicated and multi-layered stressors and entail higher cognitive complexity and time investment. As a result, it is critical to investigate the usefulness of intelligent user interfaces in improving the quality of affective support activities by assisting crowds toward quick input and minimizing their cognitive strain. Furthermore, it is easier to extrapolate the results of this complex task to any other conversational tasks that demand an organized debate where the dialogue systematically moves from one topic to another. For *e.g.*, a moderator chatbot for deliberative debate is one example [33].

3.3 Experimental Conditions

A 3 × 3 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*. For the study data, we selected a complete conversation [2] which covered the four stages of MI by describing how a therapist helped a depressed user from understanding her problems to developing a plan for change. We chose this professional case because it contained the full four stages of MI. The rich explanations on the use of micro-skills in the original case also helped us form MI prompts and extract relevant context.

3.3.1 Entry Points. 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 MI processes 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 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.

3.3.2 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 (indicated as `history` group in later discussions). 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. 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

[2] https://www.guilford.com/add/miller2/julia.pdf

Table 1: Choice of entry points corresponding to the user's utterance in each experimental condition.

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?

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*.

3.4 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 [17], 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).

3.5 Procedure

(1) Introduction. Participants were first informed about the goal of the study and how their data would be used. Once they gave their consent, they were randomly assigned to one of the nine experimental conditions. Specific instructions on how to interact with the system were provided to introduce the participants to the main components of the system.

(2) Main Task. When workers clicked the "Start Task" button, they started the main task as described in Section 3.1. The end of the task was marked by clicking the "Submit Task" button after messages were sent.

(3) Post-task Questionnaire. Workers were redirected to a post-task questionnaire after finishing the task. We asked workers to fill in a short User Experience Questionnaire (UEQ-S) to understand their perceived pragmatic and hedonic quality of the system design [56]. Next, workers were asked to report the cognitive load by completing the NASA Task Load Index (NASA-TLX) [23], which contained six items (mental demand, physical demand, temporal demand, self-performance, effort, and frustration). We further included several questions for measuring the user perceptions of general chatbots and design choices of *ContextBot* on a 7-point Likert scale. Finally, workers were allowed to give a satisfaction score on a 10-point Likert scale and leave comments about *ContextBot* and the system design.

3.6 Evaluation Metrics and Hypotheses

Standardized survey tools for examining response qualities in affective support tasks have not yet been developed. Existing works adopt metrics from information retrieval (IR), using ranking methods to compare generated responses with the expected response [49, 61, 62]. Manual evaluations may also be included, which requires subjective scoring inputs from multiple annotators. Jadeja and Varia [29] proposed four evaluation perspectives: user experience, IR, linguistic and AI perspectives. We designed survey questions to evaluate the consistency and professionalism of responses from the linguistic perspective, supplemented by standardized UEQ-S and NASA-TLX questionnaires. To establish a standardized survey tool for future studies, an exhaustive list of evaluation criteria that captures key aspects of response qualities is needed. Then a survey instrument to incorporate these criteria into structured questions is expected for diverse psychotherapy use cases. The tool will be finalized after testing the reliability and validity [13].

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 [41, 57]. 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 [32]. 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 [62]. 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 following criteria:

- the extent to which the response is related to the user's concern;

- whether or not the persons/events referred to by the response have been talked about in the dialogue;
- whether or not the response is natural without being obtrusive when embedded in the dialogue.

Since *ContextBot* was expected to convey context more effectively, we formed the following hypothesis:

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

Professionalism in Responses. The evaluation of a task-oriented chatbot usually considers the goals required by the task [38]. For example, in negotiation tasks, the percentage of games in which negotiation decisions are agreed upon is used to evaluate the chatbot's performance [63]. 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. Since different entry points required different MI techniques to focus on, customized evaluation criteria were designed for each condition. While the responses from the early entry were evaluated on whether the worker engaged the user with empathy, the middle entry was evaluated on whether the worker evoked the user to make changes. Each response was rated based on a 7-point Likert scale (*1: Highly unprofessional, 7: Highly professional*). MI-adherent guidance was expected to help workers respond more professionally than general guidance, we thus hypothesized:

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

User Experience. We measured eight constructs by using the UEQ-S where each item was scored on a 7-point Likert scale. 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. Therefore, we formed the hypothesis:

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

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. 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. Thus, we hypothesized:

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

4 RESULTS

4.1 Response Quality

We did not force workers to engage with *ContextBot*, resulting in two classes of workers: those who did interact and those who did not. 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 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 4. 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 post-hoc 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.

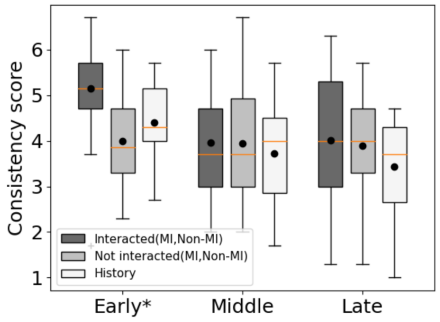

Figure 4: Response consistency scores across interacted types and entry points by combining the MI and non-MI conditions. * = statistically different (`interacted` vs. `not interacted` vs. `history`).

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. 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.

Table 2: Execution time (*mean* ± *std*, in seconds) of different conditions. Responses with consistency scores higher than 4 are regarded as "High" level of consistency. * indicates no samples in this group.

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

	MI *Mean* ± *Std*	Non-MI *Mean* ± *Std*
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

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.

4.1.1 Trade-Offs Between Response Consistency 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 2. 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 [35], 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

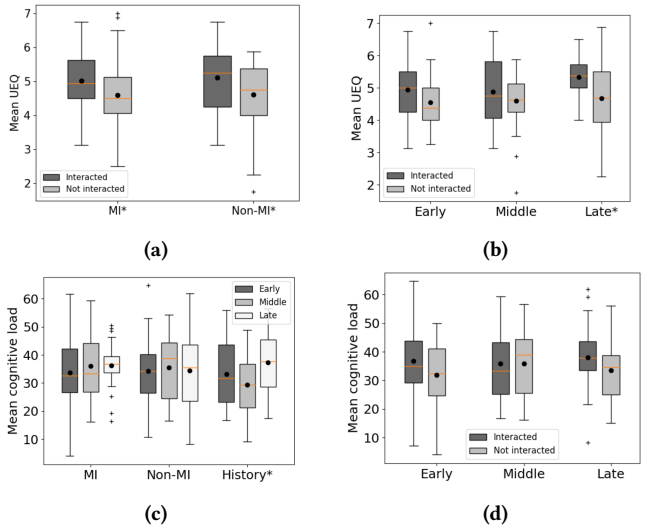

Figure 5: Boxplots of UEQ-S scores (Fig (a), (b), * = statistically significant between `interacted` group and `not interacted` group) and NASA-TLX scores (Fig (c), (d), * = statistically significant among early, middle, and late groups in the History condition in Fig (c)) in terms of the interacted types across different conditions. Black points indicate the average value of this group.

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 history conditions (*early,p=.003; middle,p=.041; late,p=.004*) under α*=.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 [1].

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.

4.1.2 Professionalism in Responses. Table 3 shows the average scores for sampled responses assessed by two professional psychologists. Since hiring psychologists was expensive, we limited our sample size in each condition to *N=5*. Workers following MI-adherent guidance consistently produced more professional responses across three entry points as shown in Table 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.

4.2 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. 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 (a)). Based on the data distributions, we performed a two-tailed independent T-test 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 α*=.05* level. When workers entered into long conversations with longer chat history (Figure 5 (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 (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 α*=.05* level. A post-hoc 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, workers who interacted with *ContextBot* showed higher cognitive load on average compared to those who did not, but the difference was not significant at α*=.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.

- To conclude, our results on response consistency, execution time, UEQ, and cognitive load show the potential benefits of interacting with *ContextBot*. Figure 6 presents the relations between the four variables and the context type.
- Constrained by sample sizes, we observed that workers who read all contexts on average could systematically produce more consistent responses, had better user experience, comparable cognitive load, and less execution time.

4.3 Qualitative Analysis

We followed an inductive thematic analysis approach to analyse the open-ended comments from 20 workers who have interacted with *ContextBot* [6]. 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 pseudonymized representations.

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.

Effectiveness of Guidance. 15% of workers felt that *ContextBot* provided them with limited information, resulting in a lack of effective guidance to help them.

> **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.

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 [3].

> **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 [26].

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

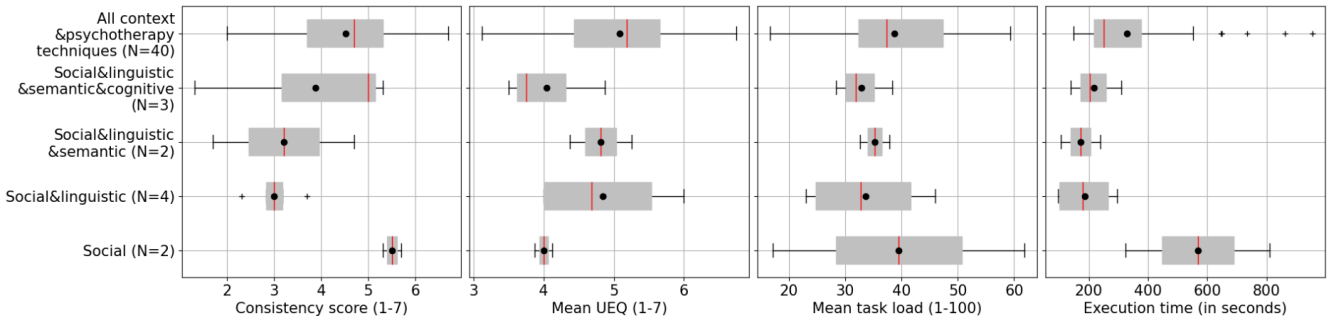

Figure 6: Relations between context types (social, linguistic, semantic and cognitive) and relevant variables (samples were selected from those who had interacted with *ContextBot*).

Table 4: Summary of our key findings.

Hypothesis	Supported?	Main Takeaways
H1: Interacting with *ContextBot* yields more consistent responses as compared to other conditions, across varying entry points and contextual guidance.	Partially	• Contextual instruction had no significant effect on consistency at any entry point. • Workers who entered the early dialogue following the interaction with *ContextBot* provided more consistent responses. • In the early and middle entry point conditions, workers who were consistent took less time on average.
H2: When using *ContextBot* with MI-adherent guidance, crowd workers respond with more professional responses than those who interact with general guidance.	Supported	Workers following MI-adherent guidance consistently produced more professional responses across the three entry points.
H3: When workers enter the dialogue late, interacting with *ContextBot* yields a better user experience compared to other conditions.	Supported	Workers who interacted with *ContextBot* reported a better user experience in long chats.
H4: Interacting with *ContextBot* does not increase the cognitive load of workers as compared to the condition of only providing chat history.	Supported	*ContextBot* had no discernible effect on workers' perceived cognitive load.

5 DISCUSSION

Our results suggest that *ContextBot* can serve as an effective tool for providing context in CPCS in affective support tasks. *ContextBot* does not have the ability to automatically generate responses, but checks workers' inputs and retrieves multiple dimensions of contexts manually predefined by annotators. The setting of MI-adherent contextual guidance in this paper initially realized the future prospect of [50], 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. Generally speaking, human chats often occur in a natural conversational flow, which may appear chaotic and illogical. Contexts are hidden in different sentences throughout the chat. Even with the help of advanced statistical models, it remains challenging to identify contexts from chat histories and link contexts to the current utterance [64]. Moreover, there is currently no sufficient evidence showing that organizing chat histories in a way that follows a specific contextual structure can lead to easier navigation. The value of *ContextBot* is thus to provide a possibility of structurally segmenting hidden contexts in chat histories and guiding the user in a humanized way. 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 [30].

Our findings of the response quality corroborated results from the work of [39] where the difference in the output quality was not significant as a result of using conversational interfaces. However, 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 2 have revealed that more time may be required when workers enter the dialogue late. 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 a conversational interface to provide the context in CPCS is that the trade-off between context amount and context quality is essential. When the task requires specific guidance (e.g., MI-adherent techniques), 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. In addition, enhancing the confidence and collective identity of workers in performing affective support tasks cannot be ignored.

5.1 Implications for Designing Intelligent Crowd Agents for Mental Health

Given the shortage of resources (e.g., lack of therapists, stigma) [9], several research communities are becoming interested in developing mental health chatbots to meet the rising demand for mental health support [14, 22, 55]. The limitations of AI, on the other hand, contribute to a high attrition rate and a poor adoption rate for these mental health chatbots [10]. Although the CPCS has the potential to bridge this socio-technical gap, it is essential to note that sustaining quality might be difficult when using labor markets such as MTurk, Prolific, or Toloka. *ContextBot* contributes to the broader goals of constructing intelligent user interfaces for crowd computing to support mental health tasks by proposing a chatbot built on a structured approach using motivational interviewing and employing different dimensions of context. The *ContextBot* can assist workers in providing appropriate and high-quality emotional support in near real-time while reducing cognitive load.

This strategy can also surpass standard training methods that train workers before the primary conversational task, such as Trainbot [3] or cognitive reappraisal technique [46], by training workers while they are busy crafting a response. Thus, our proposed solution can potentially reduce the logistical costs involved with maintaining a separate pool of workers for training and the delay caused by training, thereby assisting in developing CPCS for real-time emotional support.

Moreover, given that crowd-powered systems rely on diverse workers with different backgrounds and skills, providing them with varying levels of context and MI-adherent guidance has the potential to eliminate cognitive bias in stress management tasks. This technique can also help preserve the consistent personality of CPCS by guiding workers to build on each others' answers, which is essential for a positive user experience. Furthermore, existing work in this area maintains consistent dialogues using only linguistic context (such as keeping critical facts). We enhanced this work by applying several context dimensions developed from language theory and employing a chatbot as a mediator to support and coach workers toward quality input.

This study's findings apply to numerous application domains that allow workers to engage in structured discussions that follow a set sequence or plan. Kim et al. [33], for instance, built a chatbot named DebateBot, which leverages the Think-pair-share technique to facilitate deliberate dialogue and persuade reluctant individuals to participate. Similarly, *ContextBot* can be utilized when we want a group of individuals to reach a conclusion using a "pre-authored structure", making the conversation more focused, and facilitating a small group's decision-making [15].

6 CONCLUSIONS AND FUTURE WORK

Online peer-to-peer psychological support has the advantages of instantaneity and easy availability that cannot be easily replaced by traditional counselling. CPCS used in affective support tasks 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 adherent to a counselling dialogue, 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. Future research needs to focus on the effect of conversational interfaces in other task types, where automated methods could be adopted to extract contextual guidance from the dialogue and embed different context dimensions into the conversation flow.

ACKNOWLEDGMENTS

This work was partially supported by the Delft Design@Scale AI Lab, the 4TU.CEE UNCAGE project, and the Convergence Flagship "ProtectMe" project. We finally thank all participants from Prolific.

REFERENCES

[1] Tahir Abbas, Ujwal Gadiraju, Vassilis-Javed Khan, and Panos Markopoulos. 2021. 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*, Vol. 9. 2–14.

[2] Tahir Abbas, Vassilis-Javed Khan, Ujwal Gadiraju, Emilia I. Barakova, and Panos Markopoulos. 2020. Crowd of Oz: A Crowd-Powered Social Robotics System for Stress Management. *Sensors* 20, 2 (2020), 569. https://doi.org/10.3390/s20020569

[3] Tahir Abbas, Vassilis-Javed Khan, Ujwal Gadiraju, and Panos Markopoulos. 2020. Trainbot: A Conversational Interface to Train Crowd Workers for Delivering On-Demand Therapy. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 8. 3–12.

[4] Sameera A. Abdul-Kader and John Woods. 2015. Survey on chatbot design techniques in speech conversation systems. *International Journal of Advanced Computer Science and Applications* 6, 7 (2015). https://doi.org/10.14569/ijacsa.2015.060712

[5] Kathina Ali, Louise Farrer, Amelia Gulliver, and Kathleen M Griffiths. 2015. Online peer-to-peer support for young people with mental health problems: a systematic review. *JMIR mental health* 2, 2 (2015), e4418.

[6] Virginia Braun and Victoria Clarke. 2012. *Thematic analysis.* American Psychological Association.

[7] Christine Bundy. 2004. Changing behaviour: using motivational interviewing techniques. *Journal of the royal society of medicine* 97, Suppl 44 (2004), 43.

[8] Harry Bunt. 1994. Context and dialogue control. *Think Quarterly* 3, 1 (1994), 19–31.

[9] James A. Cartreine, David K. Ahern, and Steven E. Locke. 2010. A roadmap to computer-based psychotherapy in the United States. *Harvard review of psychiatry* 18 (3 2010), 80–95. Issue 2. https://doi.org/10.3109/10673221003707702

[10] Helen Christensen, Kathleen M Griffiths, and Louise Farrer. 2009. Adherence in Internet Interventions for Anxiety and Depression: Systematic Review. *J Med Internet Res* 11, 2 (24 Apr 2009), e13. https://doi.org/10.2196/jmir.1194

[11] Herbert H Clark and Catherine R Marshall. 1981. Definite knowledge and mutual knowledge. (1981).

[12] Tom Van Daele, Dirk Hermans, Chantal Van Audenhove, and Omer Van den Bergh. 2012. Stress Reduction Through Psychoeducation: A Meta- Analytic

Review. *Health Education & Behavior* 39, 4 (2012), 474–485. https://doi.org/10.1177/1090198111419202 arXiv:https://doi.org/10.1177/1090198111419202 PMID: 21986242.

[13] Dipankar De, Kamal Kishore, Vidushi Jaswal, and Vinay Kulkarni. 2021. Practical guidelines to develop and evaluate a questionnaire. *Indian Dermatology Online Journal* 12, 2 (2021), 266. https://doi.org/10.4103/idoj.idoj_674_20

[14] Munmun De Choudhury. 2022. Employing Social Media to Improve Mental Health: Pitfalls, Lessons Learned, and the Next Frontier. In *27th International Conference on Intelligent User Interfaces* (Helsinki, Finland) *(IUI '22)*. Association for Computing Machinery, New York, NY, USA, 1. https://doi.org/10.1145/3490099.3519389

[15] Shelly Farnham, Harry R. Chesley, Debbie E. McGhee, Reena Kawal, and Jennifer Landau. 2000. Structured Online Interactions: Improving the Decision-Making of Small Discussion Groups. In *Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work* (Philadelphia, Pennsylvania, USA) *(CSCW '00)*. Association for Computing Machinery, New York, NY, USA, 299–308. https://doi.org/10.1145/358916.359001

[16] Asbjørn Følstad, Marita Skjuve, and Petter Bae Brandtzaeg. 2019. Different chatbots for different purposes: Towards a typology of chatbots to understand interaction design. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 11551 LNCS (2019), 145–156. https://doi.org/10.1007/978-3-030-17705-8_13/TABLES/1

[17] Ujwal Gadiraju, Alessandro Checco, Neha Gupta, and Gianluca Demartini. 2017. Modus Operandi of Crowd Workers: The Invisible Role of Microtask Work Environments. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 3 (2017), 49:1–49:29. https://doi.org/10.1145/3130914

[18] Mahak Gambhir and Vishal Gupta. 2017. Recent automatic text summarization techniques: a survey. *Artificial Intelligence Review* 47, 1 (2017), 1–66.

[19] Mohsen Ghadessy. 1999. *Text and context in functional linguistics*. Vol. 169. John Benjamins Publishing.

[20] Barbara J Grosz and Candace L Sidner. 1986. Attention, intentions, and the structure of discourse. *Computational linguistics* 12, 3 (1986), 175–204.

[21] Akshit Gupta, Debadeep Basu, Ramya Ghantasala, Sihang Qiu, and Ujwal Gadiraju. 2022. To trust or not to trust: How a conversational interface affects trust in a decision support system. In *Proceedings of the ACM Web Conference 2022*. 3531–3540.

[22] Katrin Hänsel, Michael Sobolev, Tobias Kowatsch, and Rafael A Calvo. 2022. HEALTHI: Workshop on Intelligent Healthy Interfaces. In *27th International Conference on Intelligent User Interfaces* (Helsinki, Finland) *(IUI '22 Companion)*. Association for Computing Machinery, New York, NY, USA, 7–9. https://doi.org/10.1145/3490100.3511169

[23] Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology*. Vol. 52. Elsevier, 139–183.

[24] Ting-Hao K. Huang, Amos Azaria, Oscar J. Romero, and Jeffrey P. Bigham. 2019. InstructableCrowd: Creating IF-THEN Rules for Smartphones via Conversations with the Crowd. *Hum. Comput.* 6 (2019), 113–146. https://doi.org/10.15346/hc.v6i1.7

[25] Ting-Hao Kenneth Huang, Joseph Chee Chang, and Jeffrey P. Bigham. 2018. 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, CHI 2018, Montreal, QC, Canada, April 21-26, 2018*, Regan L. Mandryk, Mark Hancock, Mark Perry, and Anna L. Cox (Eds.). ACM, 295. https://doi.org/10.1145/3173574.3173869

[26] Ting-Hao Kenneth Huang, Walter S. Lasecki, Amos Azaria, and Jeffrey P. Bigham. 2016. "Is There Anything Else I Can Help You With?" Challenges in Deploying an On-Demand Crowd-Powered Conversational Agent. In *Proceedings of the Fourth AAAI Conference on Human Computation and Crowdsourcing, HCOMP 2016, 30 October - 3 November, 2016, Austin, Texas, USA*, Arpita Ghosh and Matthew Lease (Eds.). AAAI Press, 79–88. http://aaai.org/ocs/index.php/HCOMP/HCOMP16/paper/view/14050

[27] Ting-Hao (Kenneth) Huang, Walter S. Lasecki, and Jeffrey P. Bigham. 2015. Guardian: A Crowd-Powered Spoken Dialog System for Web APIs. In *Proceedings of the Third AAAI Conference on Human Computation and Crowdsourcing, HCOMP 2015, November 8-11, 2015, San Diego, California, USA*, Elizabeth Gerber and Panos Ipeirotis (Eds.). AAAI Press, 62–71. http://www.aaai.org/ocs/index.php/HCOMP/HCOMP15/paper/view/11599

[28] Shafquat Hussain, Omid Ameri Sianaki, and Nedal Ababneh. 2019. A Survey on Conversational Agents/Chatbots Classification and Design Techniques. In *Web, Artificial Intelligence and Network Applications - Proceedings of the Workshops of the 33rd International Conference on Advanced Information Networking and Applications, AINA Workshops 2019, Matsue, Japan, March 27-29, 2019 (Advances in Intelligent Systems and Computing, Vol. 927)*, Leonard Barolli, Makoto Takizawa, Fatos Xhafa, and Tomoya Enokido (Eds.). Springer, 946–956. https://doi.org/10.1007/978-3-030-15035-8_93

[29] Mahipal Jadeja and Neelanshi Varia. 2017. Perspectives for Evaluating Conversational AI. *CoRR* abs/1709.04734 (2017). arXiv:1709.04734 http://arxiv.org/abs/1709.04734

[30] Ian Andrew James, Rachel Morse, and Alan Howarth. 2010. The science and art of asking questions in cognitive therapy. *Behavioural and Cognitive Psychotherapy* 38, 1 (2010), 83–93.

[31] Ji-Youn Jung, Sihang Qiu, Alessandro Bozzon, and Ujwal Gadiraju. 2022. Great chain of agents: The role of metaphorical representation of agents in conversational crowdsourcing. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–22.

[32] Hyunwoo Kim, Byeongchang Kim, and Gunhee Kim. 2020. Will I Sound Like Me? Improving Persona Consistency in Dialogues through Pragmatic Self-Consciousness. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, 904–916. https://doi.org/10.18653/v1/2020.emnlp-main.65

[33] Soomin Kim, Jinsu Eun, Joseph Seering, and Joonhwan Lee. 2021. Moderator Chatbot for Deliberative Discussion: Effects of Discussion Structure and Discussant Facilitation. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 87 (apr 2021), 26 pages. https://doi.org/10.1145/3449161

[34] Rachel Kornfield, David C Mohr, Rachel Ranney, Emily G Lattie, Jonah Meyerhoff, Joseph J Williams, and Madhu Reddy. 2022. 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 (2022), 1–30.

[35] Walter S. Lasecki, Rachel Wesley, Jeffrey Nichols, Anand Kulkarni, James F. Allen, and Jeffrey P. Bigham. 2013. Chorus: a crowd-powered conversational assistant. In *The 26th Annual ACM Symposium on User Interface Software and Technology, UIST'13, St. Andrews, United Kingdom, October 8-11, 2013*, Shahram Izadi, Aaron J. Quigley, Ivan Poupyrev, and Takeo Igarashi (Eds.). ACM, 151–162. https://doi.org/10.1145/2501988.2502057

[36] Jiwei Li, Michel Galley, Chris Brockett, Georgios P. Spithourakis, Jianfeng Gao, and William B. Dolan. 2016. 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. https://doi.org/10.18653/v1/p16-1094

[37] Christine L. Lisetti, Reza Amini, Ugan Yasavur, and Naphtali Rishe. 2013. I Can Help You Change! An Empathic Virtual Agent Delivers Behavior Change Health Interventions. *ACM Trans. Manag. Inf. Syst.* 4, 4 (2013), 19:1–19:28. https://doi.org/10.1145/2544103

[38] Wari Maroengsit, Thanarath Piyakulpinyo, Korawat Phonyiam, Suporn Pongnumkul, Pimwadee Chaovalit, and Thanaruk Theeramunkong. 2019. A Survey on Evaluation Methods for Chatbots. In *Proceedings of the 2019 7th International Conference on Information and Education Technology* (Aizu-Wakamatsu, Japan) *(ICIET 2019)*. Association for Computing Machinery, New York, NY, USA, 111–119. https://doi.org/10.1145/3323771.3323824

[39] Panagiotis Mavridis, Owen Huang, Sihang Qiu, Ujwal Gadiraju, and Alessandro Bozzon. 2019. Chatterbox: Conversational Interfaces for Microtask Crowdsourcing. In *Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization, UMAP 2019, Larnaca, Cyprus, June 9-12, 2019*, George Angelos Papadopoulos, George Samaras, Stephan Weibelzahl, Dietmar Jannach, and Olga C. Santos (Eds.). ACM, 243–251. https://doi.org/10.1145/3320435.3320439

[40] Michael Frederick McTear, Zoraida Callejas, and David Griol. 2016. *The conversational interface*. Vol. 6. Springer.

[41] Mohsen Mesgar, Edwin Simpson, and Iryna Gurevych. 2021. Improving Factual Consistency Between a Response and Persona Facts. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, Paola Merlo, Jörg Tiedemann, and Reut Tsarfaty (Eds.). Association for Computational Linguistics, 549–562. https://doi.org/10.18653/v1/2021.eacl-main.44

[42] William R Miller and Stephen Rollnick. 2012. *Motivational interviewing: Helping people change*. Guilford press.

[43] Robert Morris. 2011. Crowdsourcing workshop: the emergence of affective crowdsourcing. In *Proceedings of the 2011 annual conference extended abstracts on Human factors in computing systems*. Citeseer.

[44] Robert R Morris, Kareem Kouddous, Rohan Kshirsagar, and Stephen M Schueller. 2018. Towards an artificially empathic conversational agent for mental health applications: system design and user perceptions. *Journal of medical Internet research* 20, 6 (2018), e10148.

[45] Robert R. Morris and Rosalind Picard. 2012. Crowdsourcing Collective Emotional Intelligence. *arXiv:1204.3481 [cs]* (April 2012). arXiv:1204.3481 [cs] http://arxiv.org/abs/1204.3481

[46] Robert R Morris, Stephen M Schueller, and Rosalind W Picard. 2015. Efficacy of a Web-Based, Crowdsourced Peer-To-Peer Cognitive Reappraisal Platform for Depression: Randomized Controlled Trial. *Journal of Medical Internet Research* 17, 3 (March 2015), e72. https://doi.org/10.2196/jmir.4167

[47] Katie Morton, Mark Beauchamp, Anna Prothero, Lauren Joyce, Laura Saunders, Sarah Spencer-Bowdage, Bernadette Dancy, and Charles Pedlar. 2015. The effectiveness of motivational interviewing for health behaviour change in primary care settings: a systematic review. *Health psychology review* 9, 2 (2015), 205–223.

[48] John A Naslund, Kelly A Aschbrenner, Lisa A Marsch, and Stephen J Bartels. 2016. The future of mental health care: peer-to-peer support and social media. *Epidemiology and psychiatric sciences* 25, 2 (2016), 113–122.

[49] Maxwell I. Nye, Michael Henry Tessler, Joshua B. Tenenbaum, and Brenden M. Lake. 2021. Improving Coherence and Consistency in Neural Sequence Models with Dual-System, Neuro-Symbolic Reasoning. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (Eds.). 25192–25204. https://proceedings.neurips.cc/paper/2021/hash/d3e2e8f631bd9336ed25b8162aef8782-Abstract.html

[50] SoHyun Park, Jeewon Choi, Sungwoo Lee, Changhoon Oh, Changdai Kim, Soohyun La, Joonhwan Lee, and Bongwon Suh. 2019. Designing a chatbot for a brief motivational interview on stress management: Qualitative case study. *Journal of medical Internet research* 21, 4 (2019), e12231.

[51] Sihang Qiu, Alessandro Bozzon, Max V Birk, and Ujwal Gadiraju. 2021. Using worker avatars to improve microtask crowdsourcing. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–28.

[52] Sihang Qiu, Ujwal Gadiraju, and Alessandro Bozzon. 2020. Estimating conversational styles in conversational microtask crowdsourcing. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW1 (2020), 1–23.

[53] Sihang Qiu, Ujwal Gadiraju, and Alessandro Bozzon. 2020. Improving worker engagement through conversational microtask crowdsourcing. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–12.

[54] Sihang Qiu, Ujwal Gadiraju, and Alessandro Bozzon. 2020. Ticktalkturk: Conversational crowdsourcing made easy. In *Conference Companion Publication of the 2020 on Computer Supported Cooperative Work and Social Computing*. 53–57.

[55] Tanja Schneeberger, Naomi Sauerwein, Manuel S. Anglet, and Patrick Gebhard. 2021. Stress Management Training Using Biofeedback Guided by Social Agents. In *26th International Conference on Intelligent User Interfaces* (College Station, TX, USA) *(IUI '21)*. Association for Computing Machinery, New York, NY, USA, 564–574. https://doi.org/10.1145/3397481.3450683

[56] Martin Schrepp, Andreas Hinderks, and Jörg Thomaschewski. 2017. 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).

[57] Haoyu Song, Wei-Nan Zhang, Jingwen Hu, and Ting Liu. 2020. 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*. AAAI Press, 8878–8885. https://ojs.aaai.org/index.php/AAAI/article/view/6417

[58] Benjamin Tolchin, Gaston Baslet, Steve Martino, Joji Suzuki, Hal Blumenfeld, Lawrence J Hirsch, Hamada Altalib, and Barbara A Dworetzky. 2020. Motivational interviewing techniques to improve psychotherapy adherence and outcomes for patients with psychogenic nonepileptic seizures. *The Journal of neuropsychiatry and clinical neurosciences* 32, 2 (2020), 125–131.

[59] David R Traum. 1999. Computational models of grounding in collaborative systems. In *Psychological Models of Communication in Collaborative Systems-Papers from the AAAI Fall Symposium*. 124–131.

[60] Teun A Van Dijk. 2007. Comments on context and conversation. *Discourse and contemporary social change* 54 (2007), 281–316.

[61] Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and Answering Questions to Evaluate the Factual Consistency of Summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault (Eds.). Association for Computational Linguistics, 5008–5020. https://doi.org/10.18653/v1/2020.acl-main.450

[62] Sean Welleck, Jason Weston, Arthur Szlam, and Kyunghyun Cho. 2019. Dialogue Natural Language Inference. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28-August 2, 2019, Volume 1: Long Papers*, Anna Korhonen, David R. Traum, and Lluís Màrquez (Eds.). Association for Computational Linguistics, 3731–3741. https://doi.org/10.18653/v1/p19-1363

[63] Ran Zhao, Oscar J. Romero, and Alex Rudnicky. 2018. SOGO: A Social Intelligent Negotiation Dialogue System. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents, IVA 2018, Sydney, NSW, Australia, November 05-08, 2018*, Anton Bogdanovych, Deborah Richards, Simeon Simoff, Catherine Pelachaud, Dirk Heylen, and Tomas Trescak (Eds.). ACM, 239–246. https://doi.org/10.1145/3267851.3267880

[64] Peixiang Zhong, Chen Zhang, Hao Wang, Yong Liu, and Chunyan Miao. 2020. Towards Persona-Based Empathetic Conversational Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, 6556–6566. https://doi.org/10.18653/v1/2020.emnlp-main.531