

# Enhancing Trust in Mental Health Agents: A DevOps Approach to Continuous Feedback Reflection

EUNKI JOUNG, KAIST School of Computing, Korea

NGUYEN LINH, KAIST School of Computing, Korea

UICHIN LEE, KAIST School of Computing, Korea

Trust is a crucial factor for effective conversational mental health agents. However, the unpredictability of language models and the lack of standardized development practices are making the agents easily produce conversational errors, which reduces users' trust. Frequent reflection of user feedback is an effective way to iteratively find out and address these errors. Though the concept of continuous feedback is widely adopted in software engineering and DevOps practices, feedback collection and reflection are not well supported in current conversational agent ecosystems. In this study, we present a case study of integrating user feedback into the workspace that the conversational designers and domain experts both use, within a context of a two-week beta test of mental health agent. We introduced a tool for logging and presenting feedback and suggested two findings from this case about the development task management and inspection of domain experts.

CCS Concepts: • **Human-centered computing** → **Human computer interaction (HCI)**.

Additional Key Words and Phrases: Conversational Agent, Continuous Feedback, DevOps, Mental Health, Trust

## ACM Reference Format:

Eunki Joung, Nguyen Linh, and Uichin Lee. 2024. Enhancing Trust in Mental Health Agents: A DevOps Approach to Continuous Feedback Reflection. In . ACM, New York, NY, USA, 5 pages.

## 1 INTRODUCTION

Digital mental health interventions are expected to address mental health needs, because they can require low cost, overcome the place limit, reduce burden of therapeutic institutes and experts, and avoid problems emerging from stigma[1, 9]. Conversational agents have benefits compared to more traditional forms of digital mental health interventions such as digital workbooks because their interactivity can lead to more engagement[3]. For these mental health agents, trust is the factor that should be established for more therapeutic effect. Losing trust towards the agents reduces users' motivation, which leads to less effectiveness. In this paper, we focused on the trust in mental health agents related to the perceived competence of the agents that is collaboratively built by users, developers, and domain experts.

However, conversational agents can lose this trust by producing *errors* in conversation[10]. Myers et al.[8]'s observation of voice agent usage shows common errors that users of conversational agents can encounter. Their categorization consists of four obstacles: users' unfamiliar intents not fitting agent capabilities, NLP errors leading to misinterpretation and misclassification of user utterances, agent's failed feedback not well captured to the user, and system errors such as unexpected behavior of cancel intents.

Reducing these errors is not easy. The innate unpredictability of machine learning technologies makes it difficult to develop robust agents. Both NLU (Natural Language Understanding) and response generation are sources of unpredictable results which are hard to deal with. Also, the lack of a common understanding of interaction models for conversational agents causes faults. Heo and Lee[4] framed this lack as a discrepancy between the 'form' model, which is a classic

interaction model consisting of fixed input areas, and the 'flow' model, which requires recognition of user intents and entities (information to collect) from users' free responses. In the flow model, understanding its own design patterns is required to properly address the flexibility of user input. However, users and most practitioners, including conversational designers and domain experts, lack this understanding, and there are no standardized practices. With this lack, many unpredicted errors can be found in the phase of release and deployment, which increases a need for actively reflecting user feedback.

To build a successful mental health agent, frequent reflection of feedback from users is an effective practice. The case of a mental health agent designed for older adults [7] showed how continuous feedback helps build successful practice using conversational agents. In this case, the users initially had difficulties engaging with the agent, reporting a perception of inconsistent answers from the agent. The researchers continuously checked the user feedback by conducting regular check-in calls and found out that they did not articulate replies in a way that the agent could understand; they added unnecessary words and used complex forms, but the agents only understood simple utterances. Based on this feedback, the research team later deployed educational materials and held a workshop, thereby achieving high user engagement and willingness to use the agent more.

In the software engineering field, the philosophy of DevOps advocates integration of development and operation, which facilitates fast development cycles with continuous delivery and rapid refinements by learning from end users[5, 6]. As a part of DevOps practice, a lean reflection of user feedback can contribute to trust-building by facilitating fast development cycles for addressing the reported conversation errors. The current DevOps ecosystem of conversational agents supports the collection of user feedback with conversational contexts in a database, such as Google BigQuery[2]. However, there is a gap among the tons of feedback from users, the inspection of domain experts, and translation to technical development tasks that would be addressed by developers. In the current conversational agent development process, it is difficult to reflect immediate feedback from users.

In this study, we present a case study of continuously collecting and reflecting user feedback during the development of a mental health conversational agent. The team consists of HCI researchers with computer science backgrounds where the first author was mostly developing the agents and two mental health experts who designed the intervention scenario. Prior to deploying the conversational agent for evidence-based mental health intervention, we conducted a beta test with 10 users recruited from the campus. We developed a tool to support getting continuous feedback from users with the conversation context, which is integrated with a workspace for task management that is collaboratively used by the developers and mental health experts. By observing this case, we suggest implications for collecting and reflecting feedback in conversational agent development involving domain experts.

## 2 CASE OVERVIEW

Before the study, we first identified the stakeholders' needs through informal interviews among team members to design a tool for continuous feedback. The conversation designers wanted to collect and manage the comments from users during the beta test period. They also reported a need for utilizing the agent version and other related information of the conversation to iterate the agent and inspect the improvement of each version. Mental health experts wanted to improve the intervention content and examine the impact and quality of conversation for users through comments.

Based on the identified needs, we designed a system to be used in the beta test of the agent for two weeks. Each user was requested to engage with the agent three times per day and leave comments while using the agent.

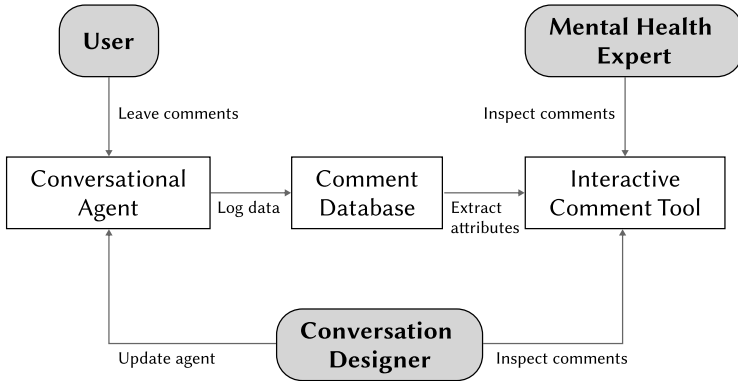


Fig. 1. The overview of an interactive comment system and related users.

As shown in Fig. 1, our system consists of the conversational agent and the interactive comment tool. The conversational agent, built on Google’s Dialogflow CX platform, logs user inputs and related conversational context into our own database in MongoDB. The interactive comment tool consists of the extraction program and the comment view page. In detail, the extraction is done by the Python program which takes logs from the database, extracts predefined attributes from the raw data, and makes a CSV file as output. This CSV file is effortlessly copied and pasted to the comment page daily by a research team member. The comment page was implemented in tabular form in Notion, a web workspace which can be used for managing development tasks as shown in Fig.2.

In the comment page, predefined attributes are extracted and presented as follows. The attributes ‘userID’, ‘timelog’, ‘sessionID’, ‘version’, and ‘comment’ are recorded as the feedback of each user and related information. The attributes for conversational context consist of the details of the dialog state machine including the current ‘flow’ (a unit of a single state machine in Google Dialogflow CX, which is mapped to each intervention scenario in our case) and the ‘page’ (a single state). ‘Hyperlink’ shows links to the conversational context including the full dialogue. The ‘comment category’, which helped us code issues and set priorities, was iteratively designed. Initially, the designers manually inspected the comments during the first week but it was time-consuming. Consequently, we created predefined categories from past issues, allowing users to categorize their comments accordingly. The researchers could leave comments about the specific user feedback in the ‘researcher comment’ attribute.

### 3 FINDING 1: USER FEEDBACK AS A SOURCE OF INFORMATION FOR DEFINING AND PRIORITIZING DEVELOPMENT TASKS

The system helped us set and rank development tasks. Using the presented system, we conducted meetings utilizing dialog histories and user feedback and collaboratively planned scenario change and technical developments. The qualitative feedback from users conveyed rich expression of user’s perceptions, which provided a better understanding to the developers and domain experts. Also, the development team could review and mark (e.g., not started / in progress / done) the comments in the task management workspace.

In the future, we need to investigate ways to automatically interpret users’ qualitative feedback on conversational agents. When the deployment becomes larger, massive qualitative feedback will

Aa session	version	user	time	comment	page	flow	history	status	category	Researcher Comment
dfMessenger-b	Test0103	U10	2024-01-23 9:3	We started chatting in the morning, but the first greeting was lunch.	howareyou	Default Start Fl	https://dialogfl	Done	First greetin...	I adjusted it depending on the

Fig. 2. The comment viewer of the developed interactive comment tool.

not be feasible to manually handle. It is reasonable to expect that the recent progress in natural language processing will enable extraction of comprehensive information from massive qualitative feedback.

#### 4 FINDING 2: DOMAIN EXPERTS ARE ENCOURAGED TO INSPECT THE AGENT FROM THE USERS' POINT OF VIEW

The mental health experts (domain experts) want to check whether the agent works as intended and collaboratively develop the agent so that they can trust it. In our case, the transformation from a raw database to a user-friendly task management tool encouraged the engagement of domain experts to inspect the user comments. For example, using the system, an expert in our team found that users feel bored when they perceive the repetitive elements of the agent. Nevertheless, she thought that doing similar interventions routinely is needed as a part of psychotherapy. Thus, she proposed the need for more guidance and education about counseling to address user needs and still preserve the meaning of counseling.

In our case, mental health experts could get feedback from users, but the communication was one-way. In the future, facilitating two-way communication between users and experts, such as affording direct replies to users' comments, will improve user satisfaction and provide a richer understanding of what users think.

#### ACKNOWLEDGMENTS

This research was supported by LGE-KAIST Digital Health Research Center (DHRC).

#### REFERENCES

- [1] Amit Baumel, Theresa Fleming, and Stephen M Schueller. 2020. Digital micro interventions for behavioral and mental health gains: core components and conceptualization of digital micro intervention care. *Journal of medical Internet research* 22, 10 (2020), e20631.
- [2] Google Cloud. 2024. *Answer feedback*. Retrieved Feb 29, 2024 from <https://cloud.google.com/dialogflow/cx/docs/concept/answer-feedback>
- [3] Kathleen Kara Fitzpatrick, Alison Darcy, and Molly Vierhile. 2017. 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 (2017), e7785.
- [4] Jeongyun Heo and Uichin Lee. 2023. Form to Flow: Exploring Challenges and Roles of Conversational UX Designers in Real-world, Multi-channel Service Environments. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2 (2023), 1–24.
- [5] Jez Humble and David Farley. 2010. *Continuous delivery: reliable software releases through build, test, and deployment automation*. Pearson Education.
- [6] Lucy Ellen Lwakatare, Terhi Kilamo, Teemu Karvonen, Tanja Sauvola, Ville Heikkilä, Juha Itkonen, Pasi Kuvaja, Tommi Mikkonen, Markku Oivo, and Casper Lassenius. 2019. *DevOps in practice: A multiple case study of five companies*.

*Information and Software Technology* 114 (2019), 217–230.

- [7] Rachel McCloud, Carly Perez, Mesfin Awoke Bekalu, and K Viswanath. 2022. Using smart speaker technology for health and well-being in an older adult population: Pre-post feasibility study. *JMIR aging* 5, 2 (2022), e33498.
- [8] Chelsea Myers, Anushay Furqan, Jessica Nebolsky, Karina Caro, and Jichen Zhu. 2018. Patterns for How Users Overcome Obstacles in Voice User Interfaces. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–7. <https://doi.org/10.1145/3173574.3173580>
- [9] Emre Sezgin, Yungui Huang, Ujjwal Ramtekkar, and Simon Lin. 2020. Readiness for voice assistants to support healthcare delivery during a health crisis and pandemic. *NPJ Digital Medicine* 3, 1 (2020), 122.
- [10] Diana-Cezara Toader, Grațiela Boca, Rita Toader, Mara Măcelaru, Cezar Toader, Diana Ighian, and Adrian T Rădulescu. 2019. The effect of social presence and chatbot errors on trust. *Sustainability* 12, 1 (2019), 256.