Understanding the Impact of Long-Term Memory on Self-Disclosure with Large Language Model-Driven Chatbots for Public Health Intervention

Eunkyung Jo*
University of California, Irvine
United States
eunkyuj@uci.edu

Yuin Jeong
NAVER Cloud
Republic of Korea
youin.jeong@gmail.com

SoHyun Park
NAVER Cloud
Republic of Korea
sohyun.s.park@navercorp.com

Daniel A. Epstein
University of California, Irvine
United States
epstein@ics.uci.edu

Young-Ho Kim
NAVER AI Lab
Republic of Korea
yghokim@younghokim.net

ABSTRACT

Recent large language models (LLMs) offer the potential to support public health monitoring by facilitating health disclosure through open-ended conversations but rarely preserve the knowledge gained about individuals across repeated interactions. Augmenting LLMs with long-term memory (LTM) presents an opportunity to improve engagement and self-disclosure, but we lack an understanding of how LTM impacts people's interaction with LLM-driven chatbots in public health interventions. We examine the case of CareCall—an LLM-driven voice chatbot with LTM—through the analysis of 1,252 call logs and interviews with nine users. We found that LTM enhanced health disclosure and fostered positive perceptions of the chatbot by offering familiarity. However, we also observed challenges in promoting self-disclosure through LTM, particularly around addressing chronic health conditions and privacy concerns. We discuss considerations for LTM integration in LLM-driven chatbots for public health monitoring, including carefully deciding what topics need to be remembered in light of public health goals.

CCS CONCEPTS

• **Human-centered computing → Empirical studies in HCI**; **Natural language interfaces**; • **Computing methodologies →** *Natural language generation.*

KEYWORDS

Chatbot, Large language models, Open-domain dialog systems, Long-term memory, Public health, Check-up calls, Social isolation

*Eunkyung Jo conducted this work as a research intern at NAVER AI Lab.

CHI '24, May 11–16, 2024, Honolulu, HI, USA

ACM ISBN 979-8-4007-0330-0/24/05
https://doi.org/10.1145/3613904.3642420

ACM Reference Format:
Eunkyung Jo, Yuin Jeong, SoHyun Park, Daniel A. Epstein, and Young-Ho Kim. 2024. Understanding the Impact of Long-Term Memory on Self-Disclosure with Large Language Model-Driven Chatbots for Public Health Intervention. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24), May 11–16, 2024, Honolulu, HI, USA.* ACM, New York, NY, USA, 21 pages. https://doi.org/10.1145/3613904.3642420

1 INTRODUCTION

Public health monitoring plays a pivotal role in safeguarding and promoting the health and wellbeing of populations by tracking various factors for containing the spread of infectious diseases [48, 69] and promoting awareness of the public's wellbeing routines [20, 66]. Across different contexts, public health monitoring often requires recurrent data collection from populations, demanding immense time and effort from public health authorities [26–28, 53]. For instance, a press article estimated that contact tracers during COVID-19 called about ten contacts per case and followed up with each contact every other day, expecting that 37 out of 50 states in the United States did not have enough contact tracers to effectively contain the spread of the virus in its early stages [67]. To reduce the burden of public health authorities in monitoring people at scale, technology, such as chatbots and mobile apps, has increasingly been used or proposed to automate aspects of the collection of personal health information for various public health monitoring contexts, including contact tracing [35, 47, 48], maternal health education [26–28, 53, 79], and social isolation intervention [30].

A key challenge in leveraging technology for public health monitoring is how to elicit health disclosure effectively from individuals. People often avoid revealing their vulnerabilities [13, 73], which makes them hesitant to disclose sensitive information, such as their health issues. Further, while public health monitoring often requires maintaining regular engagement, it is more challenging to sustain health disclosure over time [39, 44]. Research has suggested that chatbots can be potentially effective at eliciting self-disclosure about sensitive topics [18, 43, 50, 60] as they are often perceived as non-judgemental [32, 49, 70]. Recent advances in large language models (LLMs) have brought breakthroughs in chatbots' abilities to support free-form conversations on open-ended topics, offering the potential to be particularly effective at eliciting disclosure about broader aspects of personal health. However, current LLM-driven chatbots rarely offer the capability to store and reference information from previous sessions. As a result, when introduced to support public health monitoring, LLM-driven chatbots often struggled with following up on personal health history based on past conversations [30], potentially reducing user engagement and self-disclosure needed for public health monitoring.

To support continued interactions, research in Natural Language Processing (NLP) has increasingly attempted to augment LLMs with *long-term memory* (LTM), an ability to remember information, such as an individual's interests or health status, beyond the current conversation session[1]. LLM-driven chatbots with LTM can store information about certain topics from each session and feed it to the model along with other input in each conversation turn, providing cues for the chatbot to refer to. Some exemplary LLM-driven chatbots such as OpenAI's ChatGPT [54] and Google Bard [19] have recently introduced features like *"memory"* or *"custom instructions"* which statically retain certain user background information or stylistic preferences across sessions and globally apply them to the chatbot behaviors [55]. However, in public health contexts, it is essential to dynamically update the memory to reflect users' constantly changing health statuses and generate responses accordingly. LLM-driven chatbots with LTM, particularly when designed to dynamically store, update, and reference information from previous sessions, have the potential to promote disclosure in public health contexts by offering personalized conversations that bring up specific health concerns one may have. Such a memory capability also offers a valuable opportunity to maintain user engagement in public health settings, which is crucial for understanding of the population's health and wellbeing. Despite the potential of LTM to support public health monitoring through chatbots, there is a limited understanding of how LTM impacts user interactions with and perception of LLM-driven chatbots, particularly in the context of public health interventions. In this study, we therefore seek to understand the utility of LTM for public health monitoring, with particular attention to self-disclosure.

To understand the impact of LTM on people's health disclosure and their impressions of chatbots, we explore the case of CareCall, an LLM-driven voice chatbot that monitors socially isolated individuals' health via check-up phone calls. CareCall was designed to support public health monitoring and used in conjunction with public health agencies in local municipalities in South Korea. Initially developed without LTM, CareCall later integrated this capability, which automatically stores and updates summarized information about five topics— *Health, Meals, Sleep, Visited Places, and Pets*— from each session and feeds it to the LLM in the following sessions to provide cues for the chatbot to refer to appropriately. The unique setting provides a useful case for comparing user experiences with and without LTM, particularly around self-disclosure. Through the case of CareCall, we specifically ask: **How does LTM impact users' self-disclosure with and impressions of LLM-driven chatbots in the context of public health intervention?**

To answer this question, we quantitatively and qualitatively analyzed 1,252 call logs from the real-world deployment for two distinct user groups—those who engaged with CareCall with LTM ($N = 66$) and those without it ($N = 81$). To understand users' perceptions of LTM in public health monitoring chatbots, we supplemented the call log analysis with interviews with nine users who used CareCall with LTM. We found that those who used CareCall with LTM disclosed more details about their health compared to those who used CareCall without it. In addition, repeated experiences with LTM led to greater disclosure with CareCall. Users often showed more positive and less negative reactions toward the chatbot as they experienced more LTM events, perceiving such conversations as personal and emotionally supportive. However, we observed some challenges of LTM in promoting self-disclosure, leaving some users less engaged with repeated questions on chronic health issues and raising privacy concerns around sensitive health topics.

Based on the findings, we discuss design opportunities for LTM integration in LLM-driven chatbots for public health monitoring. We propose that designers should carefully decide the topics for LTM to remember, balancing the need to support open-ended conversations about broader aspects of people's lives and adhere to specific public health monitoring goals. We also highlight LTM's potential to mitigate the impersonality of chatbots and demonstrate care. Lastly, we suggest the need to consider tensions in memory needs for public health utility versus privacy sensitivity.

Key contributions of this work include:

- An empirical understanding of the impact of LTM on people's health disclosure and their impressions of LLM-driven chatbots for public health intervention, gained from analysis of 1,252 call logs from the real-world deployment of CareCall with and without LTM and interviews with nine users who were using the chatbot with LTM. The implementation and repeated experiences of LTM in CareCall led to greater health disclosure. LTM also fostered positive impressions of the chatbot by offering familiarity, leading users to perceive the system as personal and emotionally supportive. However, we observed some challenges of LTM in promoting self-disclosure, particularly around chronic health issues and privacy concerns.
- Implications for future research and implementation of LTM in LLM-driven chatbots for public health monitoring, which include: (1) designing LTM with careful topic selection in light of public health monitoring goals; (2) designing thoughtful LTM-triggered questions to mitigate the impersonality of chatbots and demonstrate care; and (3) balancing the memory needs for public health utility and privacy sensitivity.

2 RELATED WORK

Understanding the impact of LTM on LLM-driven chatbots in public health builds on past work on chatbots for enhancing self-disclosure on health and chatbots driven by LLMs and LTM augmentation.

[1]In this paper, we follow the definition of long-term memory that is commonly used in the NLP community [2, 75, 77, 78, 83], which draws parallels to the one in human cognition (*e.g.*, [75, 83]). Note that in our study context, the term *'long-term'* is not confined to a specific time frame. This contrasts with frequent conversations in HCI around the length of deployment of technology interventions, such as evaluating long-term engagement or behavior change [33, 36].

2.1 Chatbots for Enhancing Self-Disclosure on Health

Self-disclosure, a process where a person reveals personal or sensitive information to others [1], is associated with many benefits, such as stress relief [22] and closer relationships between conversational partners [1]. However, eliciting disclosure, particularly on sensitive topics, poses challenges due to people's reluctance to reveal vulnerabilities [13, 73]. Prior work indicated that people might feel more comfortable disclosing sensitive information to a digital system rather than a human because technology is often perceived as anonymous [47, 48, 50] and non-judgemental [32, 47, 49, 70]. In particular, chatbots can make people feel more comfortable expressing themselves, given their dialogue-driven nature. Prior work has thus frequently proposed chatbots as a promising approach to encourage self-disclosure on sensitive topics, including mental health [18, 43, 50, 59] and sexual health [46]. Studies have explored how different designs of chatbots might impact people's self-disclosure. For example, chatbots that disclose information about themselves have been shown to foster mutual self-disclosure [43, 44, 65]. However, prior work predominantly relied on one-time user studies, limiting our understanding of how chatbots can sustain self-disclosure over time [39, 44].

Prior work suggests that augmenting chatbots with the ability to remember and reference previous conversations has the potential to support sustained self-disclosure over time. Studies demonstrated that when chatbots remember information across multiple sessions, such as users' names or preferences, people perceive them as empathetic [29, 52, 63] and conscientious [8, 17]. However, studies indicated that users might have feelings of privacy violations around chatbots' memory [14, 52]. Recently, Cox *et al.* suggested that the design of how chatbots reference past conversations with users brings an interesting tension in users' perceptions of chatbots [14]. While users perceived a chatbot that used verbatim or paraphrased references as more engaging and intelligent, they raised privacy concerns about such reference formats. On the other hand, when a chatbot used non-explicit references, users doubted that the chatbot could understand them and were hesitant to provide details about their health habits [14]. While implementing memory can potentially improve LLM-driven chatbots' ability to elicit and sustain self-disclosure, the sensitivity of disclosure in health and the need for sustained engagement warrants deeper exploration into the impact of memory. In this study, we thus seek to understand how LTM impacts self-disclosure in the public health space, where sustained engagement is critical to developing an understanding of the health and wellbeing of individuals.

2.2 Chatbots driven by Large Language Models and Long-Term Memory Augmentation

The recent introduction of large language models (*e.g.*, GPT [5], HyperCLOVA [34], PaLM [11], LLaMA [72]; An intensive survey in [81]) gave birth to a new development paradigm for chatbots, moving beyond the traditional rule-based or retrieval-based ones. Applied to chatbots, an LLM is typically prompted to generate responses considering both an instruction (*e.g.*, personality and behavioral guidance for the agent) and the current dialogue. With their large-scale (*i.e.*, over billions of) parameters trained with a tremendous amount of human-produced text corpus, LLMs tend to generate responses that coherently and organically follow up the conversation and flexibly respond to emergent topics [76]. Due to these benefits, LLM-driven chatbots are increasingly developed or proposed by both practitioners (*e.g.*, ChatGPT [54], Bard [19], Character AI [7], Pi [24]) and researchers (*e.g.*, [10, 41, 76, 82]).

Most exemplary LLM-driven chatbots, represented by ChatGPT [54] and Bard [19], did not suppose repetitive interaction scenarios in their early versions, resulting in each session not informing the following ones, because they were mainly designed as assistants performing single-shot tasks such as code generation and reasoning. Recently, ChatGPT and Bard implemented *"memory"* or *"custom instructions"* so that the systems can remember some background information (*e.g.*, *'I'm a software developer and solely use Python.'*) or stylistic preferences (*e.g.*, *'When I ask for code, just give me the code without any explanation on how it works.'*) of users across sessions [55]. However, such memory features are not designed to automatically infer what information is *important* to remember (*e.g.*, inferring that the user's primary programming language is Python when someone continues to request code examples in Python) and instead expect individual users to manually indicate it. Further, these features do not dynamically update the memory based on their most recent conversations with users. Replika [56] is a rare example of a commercial LLM-driven chatbot that is designed to automatically store and refer back to information from previous interactions—such as hobbies, preferences, or names—with the goal of providing support for mental wellbeing. However, a recent study showed that Replika often failed to remember important information about users despite the promise that the chatbot remembers previous conversations through its "memory bank," [51] which interfered with their ability to develop bonding with users. This finding suggests the ability to understand and appropriately reference key information from past conversations is a critical aspect of LLM-driven chatbots that aim to provide emotional support.

Augmenting LLMs to 'remember' past information—often referred to as 'long-term memory' [2, 75, 77, 78, 83]—presents significant challenges for two main reasons. First, LLMs can receive input text only within a limited context window (input size). Including the entire conversation session history in the input prompt is thus not feasible for longer-term interactions. One common approach is to include summarized information of the conversation history instead of a raw knowledge base (*e.g.*, [2, 41, 75]). Second, designing how chatbots should refer to stored information back in conversation involves complex considerations. For example, Cox *et al.* [14] found that the phrasing style of user messages in past conversations impacts the perceived intelligence of and engagement with chatbots as well as privacy concerns around them, suggesting the importance of careful LTM design, particularly in sensitive health domains. Motivated by the gap in understanding the utility of LTM in LLM-driven chatbots, we explore the case of CareCall, a rare example of an LLM-driven chatbot that automatically stores and updates key information from previous conversations to support public health monitoring.

3 STUDY CONTEXT: CLOVA CARECALL AND LONG-TERM MEMORY

In this section, we describe the study context of CareCall with LTM as an example of an LLM-driven chatbot deployed to support a public health intervention. This section builds on prior work on the underlying technology (*c.f.*, [2, 3, 34]) and the design documents written by the designers and developers of CareCall. Building upon previous studies that contributed to the novel implementation of CareCall and LTM, we treat CareCall as a case study for understanding how LTM impacts users' self-disclosure with LLM-based chatbots in the context of public health intervention.

3.1 Motivation and Deployment of CareCall

CareCall is an LLM-driven voice chatbot designed for socially isolated individuals [6]. The chatbot calls the users weekly and engages in an open-ended conversation about their daily lives for about 2 to 3 minutes. The motivation behind the system stems from the recent act on the Prevention and Management of Lonely Death in South Korea [37]. The act highlights the importance of identifying and supporting individuals who are at risk of lonely deaths, which is a frequently discussed concern in South Korea. Thus, CareCall engages in conversations with users to check in with their health and overall wellbeing and provide emotional support.

CareCall was first rolled out in a municipality in South Korea in November 2021 [6] and expanded to others over time. As of October 2022, around the end of our data collection period, CareCall was being deployed to around 6,000 individuals across different municipalities in Korea. CareCall was deployed to middle-aged (40s to 60s) and older adults (60s or older) living alone. Motivated by the significant increase in lonely deaths among low-SES populations [80], such as recipients of the National Basic Livelihood Security (below 50% of median household income), CareCall users were largely from this demographic. These users were recommended to use CareCall by public officers who provide social care services in their municipalities. In most municipalities where CareCall was deployed, including the ones we include in this work, public officers monitored the call recordings and reached out to the person if any negative health signals were detected (*e.g.*, skipping meals, poor sleep, health issues) or if the person did not answer CareCall calls a certain number of times in a row.

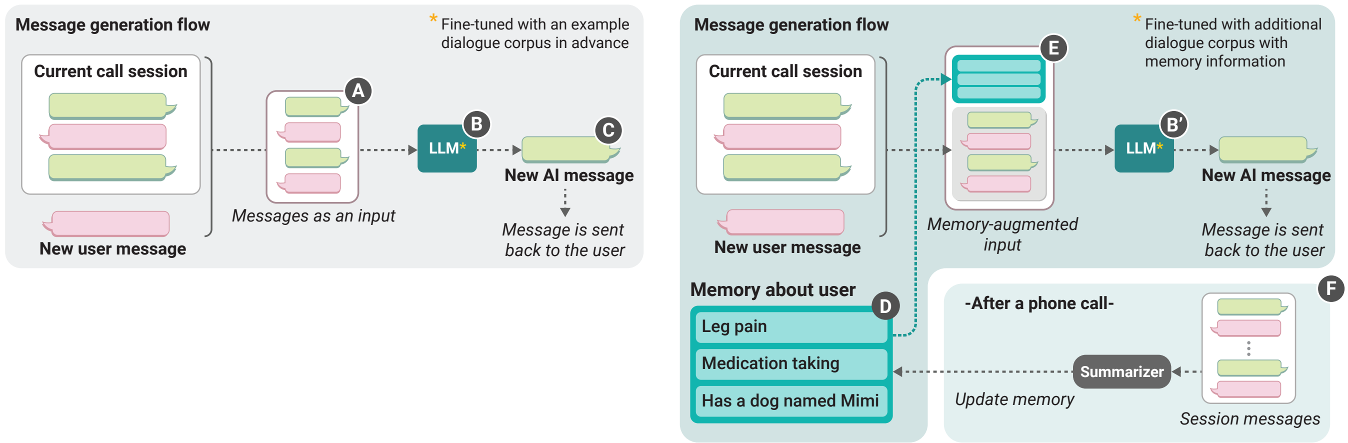
3.2 CareCall and Long-Term Memory

CareCall was first developed without LTM and deployed from November 2021 to September 2022. LTM was integrated into the existing deployments in September 2022. In this work, we compare user experiences with and without LTM.

Architecture of CareCall without LTM. Figure 1a illustrates the architecture of CareCall prior to the implementation of LTM. CareCall was designed as an open-ended dialogue system powered by an LLM called HyperCLOVA [34] (Ⓑ in Figure 1a). Once a new user message is received (Ⓐ in Figure 1a), the system feeds the current dialogue history into the LLM (Ⓑ in Figure 1a) to generate a response (Ⓒ in Figure 1a) that naturally continues the conversation. The LLM was fine-tuned in advance with a large-scale dialogue corpus, which was generated with a novel data augmentation technique. In this method, a machine learning model generates synthetic dialogues from a small set of human-written dialogues and trained workers flag and fix errors in the synthetic dataset (please see [3] for more details about the data augmentation technique). The example dialogues, which consist of dialogues relevant to five health topics (meals, sleep, health, going out, and physical activity), steer the LLM to generate appropriate responses in the public health intervention context while aiming to avoid selecting undesirable or potentially harmful phrases [2].

Architecture of CareCall with LTM. Figure 1b illustrates the architecture of CareCall with LTM. (Please see Appendix A for a more detailed description of the design process, design rationales, technical implementation, and evaluation of LTM in CareCall.) LTM was incorporated into CareCall to improve its ability to provide emotional support by offering familiarity with users. At the end of each session, a summarizer driven by an LLM (Ⓕ in Figure 1b; [2]) generates summary sentences that are relevant to the five LTM topics (Section 3.2). The memory management layer (Ⓓ in Figure 1b) stores and updates the summary sentences upon each call (*e.g.*, Removing the "Regular visit to a clinic due to leg pain" status after a user reports that they have completed the treatment). Unlike CareCall without LTM, the stored information from previous sessions is included in the model input (Ⓔ in Figure 1b), providing cues for the chatbot to refer to (*e.g.*, "*How is your leg feeling?*"). In this version, the underlying LLM (Ⓑ' in Figure 1b) was further fine-tuned with additional example dialogue corpus designed as a multi-session chat with memory to reflect the new memory-augmented input format (Ⓔ in Figure 1b). Given that LLM-driven chatbots are aimed at generating responses that coherently and organically follow up on the conversation [76], whether an LTM event is triggered is dependent on many aspects of the conversation, and there is no guarantee that an LTM event will be triggered in a given dialogue. Both CareCall without and with LTM incorporated the 6.9B parameter model of HyperCLOVA for response generation, maintaining the same conversation style, mimicking a social worker who engages in casual conversations to check in with socially isolated individuals. The model is almost equivalent to the Curie engine of the legacy GPT-3 family in terms of the training method and model size but was pre-trained on a large-scale Korean language corpus, yielding a robust performance in Korean language tasks [34]. Unlike more recent GPT models such as InstructGPT-3, GPT-3.5, and GPT-4, the HyperCLOVA series does not incorporate alignment tuning, RLHF (Reinforcement Learning from Human Feedback, [57]), an approach that improves the model's ability to follow up the natural language instructions in the model input. Therefore, rather than depending solely on natural language instructions, CareCall incorporated a fine-tuning strategy with an example corpus across conditions to effectively steer conversations in alignment with the agent persona. This alternative strategy has been shown to be cost-efficient and performant, allowing for the use of a smaller-sized model and ensuring faster response times for users [9, 64].

LTM Topics. To offer familiarity with users, LTM stores summarized information about five topics: (1) *Health* (*e.g.*, whether users have any health issues, what the health issues are, whether they are receiving clinical care, what type of clinical care they are seeking),



(a) CareCall without LTM

(b) CareCall with LTM

Figure 1: Architecture of the two different versions of CareCall chatbots, an open-ended dialogue system powered by an LLM called HyperCLOVA [34]. (a) In the initial version of CareCall without LTM, the system generates a response (Ⓒ) by feeding the current dialogue history (Ⓐ) into the LLM (Ⓑ) that was fine-tuned in advance with an example dialogue corpus that covers five health topics—meals, sleep, health, going out, and physical activity. The user information obtained from previous calls did not affect future calls since this version did not have long-term memory. (b) CareCall with LTM retains user information from the call logs. At the end of each session, a summarizer driven by an LLM (Ⓕ) generates summary sentences that are relevant to the five LTM topics (see below), which are stored and updated by the memory management layer (Ⓓ). The summary sentences are then included in the model input (Ⓔ) so that the underlying LLM (Ⓑ') can take that knowledge into account when generating responses in the following sessions. In this version, the LLM (Ⓑ') was further fine-tuned with an additional example dialogue corpus designed as a multi-session chat in memory-augmented format.

(2) *Meals* (*e.g.*, whether and why users are having difficulty eating, how they are managing difficulty in eating), (3) *Sleep* (*e.g.*, whether users are having difficulty sleeping, what difficulty they are experiencing related to sleep, how they are managing difficulty in sleeping), (4) *Pets* (*e.g.*, whether people have pets, what kind of pets they have, what their names are, what they do with their pets), and (5) *Visited Places* (*e.g.*, what places users visit frequently, what they do in those places). LTM remembers any noteworthy information that comes up during conversations relevant to the five LTM topics, including both positive and negative health experiences. For example, when users mention that they have been seeing a doctor for leg pain, CareCall would ask LTM-triggered questions in later sessions, such as *"How does your leg feel?"* or *"You mentioned having knee joint issues last time. Are you still seeing the doctor?"* Similarly, when users mention that they are regularly engaging in physical activity, CareCall would ask LTM-triggered questions in later sessions, such as *"Are you still regularly going to the park to take a walk?"* Driven by CareCall's primary goals—to check up on individuals' health and wellbeing, and to provide emotional support—the example dialogue corpus for LTM was curated with the following priority order: (1) *Health*, (2) *Meals* and *Sleep*, and (3) *Visited Places* and *Pets*.

4 METHODOLOGY

To understand how LTM impacts users' self-disclosure and how users react to LTM-triggered exchanges in the context of public health intervention, we analyzed 1,252 call logs from the real-world deployment for different user groups who used CareCall with and without LTM. Taking a mixed-method approach, we analyzed the call logs both quantitatively and qualitatively. We also conducted interviews with nine users who have been using CareCall with LTM to supplement the call log analysis.

The call log analysis leveraged pre-existing data that had been collected through the deployment of CareCall with participants' informed consent, which included collecting health information through conversations with the system and agreeing to analyze the data for research purposes. The call log analysis was classified as exempt by the guidelines from the Ministry of Health and Welfare of South Korea. The interview study was approved by the public institutional review board affiliated with the Ministry of Health and Welfare of South Korea.

4.1 Data Collection

4.1.1 Call Log Collection. Figure 2 illustrates our sampling and screening process of the CareCall users and the number of call logs. In this study, we specifically focused on two metropolitan cities with similar populations in South Korea (City A: 2.96 million, City B: 3.33 million as of May 2023), located in different regions (City A: northwestern, City B: southeastern). Both cities have a mix of urban and suburban areas. In 2021, the average household income in City A was around 5.5K USD, marginally below the national average of 5.6K USD, while City B's average income was around 4.96K USD, notably lower, indicating a modest economic disparity between the two cities [38]. CareCall was deployed to middle-aged (40s to 60s) and older adults (60s or older) living alone in both

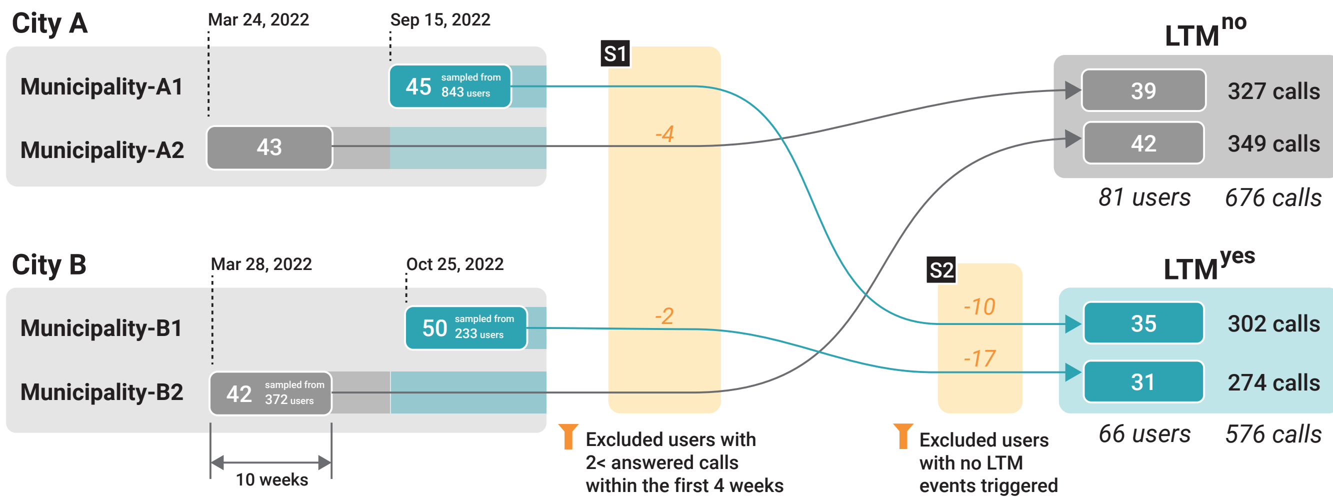


Figure 2: Overview of sampling and screening users from municipalities and the final datasets for the LTM^{yes} and LTM^{no} groups.

cities. We sought to select cities where CareCall had been deployed both with and without LTM. Versus comparing between cities, this approach helped minimize cultural influences on participants' use and perceptions of the technology.

To compare user experiences with and without LTM, we organized two groups: LTM^{yes}—users who started conversing with CareCall with LTM—and LTM^{no}—users who only used CareCall without LTM—by assigning municipalities from each city to both groups. From the two cities, we selected four municipalities (two each) that meet two inclusion criteria. First, we only included municipalities that had deployed CareCall for more than ten weeks to ensure opportunities for LTM to be triggered. We therefore excluded municipalities that recently started their deployment of CareCall at the time of the data collection (June 2023). Second, for LTM^{yes} group, we only included municipalities that had not previously deployed CareCall without LTM before deploying the version with LTM (Municipality-A1 and Municipality-B1 in Figure 2). Since prior use of CareCall without LTM could influence users' perceptions towards CareCall with LTM, we screened municipalities to exclude the ones that introduced LTM in the middle of their deployment.

We gathered call logs in the first 10-week window from LTM^{yes} group (45 individuals from Municipality-A1 and 50 individuals from Municipality-B1 who used CareCall with LTM) and LTM^{no} group (43 individuals from Municipality-A2 and 42 individuals from Municipality-B2 who used CareCall without LTM). The four municipalities deployed CareCall with varying user group sizes (843 in Municipality-A1, 43 in Municipality-A2, 233 in Municipality-B1, and 372 in Municipality-B2). To avoid over-reliance on one municipality and ease the analysis burden, we included all users from Municipality-A2 and randomly sampled a similar number from other municipalities (See Figure 2). We opted to gather the call logs within the first ten weeks of deployment for consistent comparison, though deployment in these municipalities was longer.

To ensure a certain level of engagement for examining the impact of LTM, we screened individuals in each group, only including those who answered the weekly calls twice or more in the first four weeks (excluding four from the LTM^{no} group and two from the LTM^{yes} group; S1 in Figure 2). Because our goal was to compare the influence of LTM on conversations, we further screened individuals in the LTM^{yes} group, only including those who experienced CareCall triggering LTM at least once in the first 10-week window (excluding 27 from the LTM^{yes} group; S2 in Figure 2).

As a result, our final dataset included: (1) LTM^{yes} group: 576 call logs in the 10-week window from 66 individuals (35 from City A, 31 from City B; LTM^{yes}-1 – LTM^{yes}-66) and (2) LTM^{no} group: 676 call logs in the 10-week window from 81 individuals (39 from City A, 42 from City B; LTM^{no}-1 – LTM^{no}-81). In total, we analyzed 1,252 call logs from 147 individuals. The LTM^{yes} group consisted of 17 males and 49 females, while the LTM^{no} group included 32 males and 49 females. We did not have access to other demographic information of the users included in the call log analysis.

4.1.2 Interviews with Participants Using CareCall with LTM. After completing the call log analysis, we still had a few open questions

Table 1: Demographics of interview participants and duration of CareCall use at the point of the interviews

Alias	Age	Gender	Duration of CareCall Use
P1	65	Male	10 months
P2	66	Male	10 months
P3	61	Male	10 months
P4	61	Female	10 months
P5	65	Male	10 months
P6	76	Male	4 months
P7	66	Female	10 months
P8	77	Male	10 months
P9	81	Female	10 months

about how users perceived and experienced the LTM feature, motivating us to conduct a small number of clarifying interviews. We recruited nine individuals from Municipality-B1 who were receiving check-up calls from CareCall with LTM (P1 – P9) by distributing flyers to their community service centers. The interview participants included six males and three females, aged from 61 to 81. All but P6 had been receiving check-up calls through CareCall with LTM once a week for over ten months at the point of data collection (August 2023). We asked a focused set of questions around user experiences of LTM, including (1) memorable conversations with CareCall in either a positive or negative way, (2) experiences of LTM events including what information the agent remembered from their past conversations, how the agent followed up, and how they felt about such follow-ups, (3) experiences where the agent failed to remember necessary information, and (4) perspectives around what types of information they wish the agent does or does not remember. We conducted all interviews via phone, recording the conversations, and each interview lasted for 20 to 30 minutes. We compensated each participant 50,000 KRW (approximately 37.3 USD as of Aug 2023) as a gift card.

4.2 Data Analysis

We conducted both quantitative and qualitative analyses on the call logs, as well as qualitative analysis of the interview transcripts. Each user response during the calls was auto-transcribed prior to feeding into the LLM and the recordings of interviews were auto-transcribed through an AI transcription tool. We manually corrected the automatic speech recognition errors in these transcripts. Utilizing a method frequently applied in conversation analysis [31, 45], we also included some non-verbal cues to the transcripts—including pauses, sighs, chuckling, animated tone, and agitated tone. The inclusion of such non-verbal cues was aimed at gaining a more nuanced understanding of users' emotional responses during their interactions with CareCall. The first author, who is a native Korean and is fluent in English, translated the transcripts of the call logs and interviews during the analysis, paraphrasing some idioms and phrasings to sound more natural in English.

4.2.1 Quantitative Analysis: Call Logs. We quantitatively analyzed the call logs, specifically looking to develop a codebook to recognize disclosure around the five health topics (meals, sleep, health, going out, and physical activity) and users' reactions to the chatbot. With this goal, the first author first open-coded call logs from 40 users (around 25% of the data), 20 from each group. The entire research team regularly met to generate and iterate on the codebook. The final codebook (See Appendix B) had nine categories (*Meals, Sleep, Health, Clinical, Activity, Wellbeing, Hobbies, Positive Reactions*, and *Negative Reactions*) and 19 codes (*Meals-simple, Meals-detail, Sleep-simple, Sleep-detail, Health-simple, Health-detail, Health-attitudes, Clinical-simple, Clinical-detail, Activity-simple, Social-detail, Physical-detail, Wellbeing-simple, Wellbeing-detail, Hobbies, Appreciation, Anthromorphization, Negative Feedback*, and *Disregard*).

Some categories closely aligned with the five health topics that CareCall was designed to ask. We further split health information into categories for *Health* (disclosure about health issues that warrant clinical care), *Clinical* (disclosure about clinical care that people are seeking), and *Wellbeing* (disclosure about general wellbeing that does not necessarily warrant clinical care). We divided these three categories, as well as *Meals* and *Sleep*, into codes for *-simple* (*e.g., Health-simple*: "I'm not feeling well.") and *-detail* (*e.g., Health-detail*: "I still have some pain in my legs.") based on the level of elaboration that participants provided on these categories. We added a separate code for *Health-attitude* because users occasionally mentioned their feelings and opinions about their health status and management, which differed from other informational details about health. We combined simple information about social and physical activities into *Activity-simple* as users' remarks about those two topics frequently overlapped with each other. Although both depth and breadth are important dimensions of self-disclosure [1], our quantitative analysis specifically focused on depth of disclosure as CareCall imposes predefined boundaries on conversations by design. Unlike natural conversations, CareCall leads conversations by asking questions relevant to the LTM topics, which limits opportunities for users to disclose information about other broader topics. The breadth of disclosure was further constrained by the conversation turn limit of CareCall (up to 15 agent-user turn pairs before ending the call), suggesting that depth was a better indicator of disclosure in our study context. In addition, in our coding process, we considered any disclosure as relevant whether it contained "new" information compared to what has been brought up in previous sessions. Even if an individual disclosed the same information, it still offers updates about how the person is doing at different time points, which can be valuable for public health monitoring. For example, if a person consistently reports having regular meals and sleeping over multiple sessions, public health workers can be reassured about their wellbeing. Conversely, if someone regularly mentions poor eating habits and lack of sleep across multiple sessions, the repetition signals to public health workers that this person may need additional interventions or care. We thus consider our approach to be valid for examining disclosure with LLM-driven chatbots in our study context.

Informed by prior work on conversation analysis [16, 31, 45], we developed the following coding guidelines: (1) The basic unit of coding is the agent-user turn pair since the call is mostly driven by the agents' questions; (2) Multiple codes can be applied to the same pairs; (3) When a *-detail* code is applied to a pair, the corresponding *-simple* code is applied to the pair as well. Three researchers used the initial codebook to code the call logs from eight users (four from both of the LTM^{yes} and LTM^{no} groups), reaching an initial agreement of 80% or higher for 17 out of 19 codes and revising the definitions of those codes to resolve ambiguities, and then coded the remaining call logs.

With the coded data, we statistically compared (1) users' self-disclosure and (2) reactions to CareCall between LTM^{yes} and LTM^{no} groups using a linear mixed-effects model for each code. We treated code counts as a dependent variable, groups as a categorical fixed effect, and participants as a random effect. We report this analysis in Section 5.1.1 and Section 5.2. In the initial analysis, we also fitted models with gender, cities, and call index (*i.e.*, nth call) as fixed effects. However, we found that factors other than LTM groups did not have a significant impact on most of the codes except a couple at the 5% significance level: gender (*Health-attitudes, Appreciation*), cities (*Sleep-simple, Physical-detail, Anthromorphization*), and the call index (*Sleep-simple, Health-simple, Wellbeing-simple*). Given the insignificance of these factors, we excluded them from the analysis. We further examined how LTM^{yes} group's self-disclosure changed as they experienced more LTM events using a linear mixed-effects model for each code; we treated code counts as a dependent variable, the cumulative number of LTM events as a continuous fixed effect, and participants as a random effect. We report this analysis in Section 5.1.2 and Section 5.2.2.

4.2.2 Qualitative analysis: Call Logs and Interview Transcripts. We analyzed the call logs and interview transcripts using thematic analysis [4]. The qualitative analysis aimed to add more nuances to the quantitative findings in describing users' reactions and perceptions toward LTM. We thus organized our qualitative findings around the quantitative findings on self-disclosure and users' reactions to the chatbot. The first author coded the call logs and interview transcripts around our quantitative findings, going through several rounds of iteration. The full research team then discussed and identified patterns and themes through multiple rounds of peer debriefing meetings. We incorporate the qualitative analysis into different sections of the findings.

5 FINDINGS

We found that participants who used CareCall with LTM disclosed more health details compared to those who used CareCall without it, and their repeated experiences with LTM led to greater disclosure. We also observed that LTM promoted positive reactions and mitigated negative reactions toward the chatbot by offering familiarity. Our interview participants described such conversations as personal and emotionally supportive. However, some LTM events revealed potential challenges in promoting self-disclosure, particularly around chronic health issues and privacy concerns.

5.1 LTM Increased Elaboration on Health Information Over Time.

We found that the LTM^{yes} group disclosed more health details than the LTM^{no} group. We further observed that repeated experiences of LTM (in the LTM^{yes} group) led to greater disclosure.

5.1.1 LTM Encouraged Disclosure on Health Information. Overall, the LTM^{yes} group disclosed more information about themselves compared to the LTM^{no} group both in terms of *-simple* ($p = 0.01$; 95% CI 0.05–0.41 higher code counts per call) and *-detail* codes ($p < 0.001$; 95% CI 0.32–0.74 higher code counts per call). Note that *-simple* code counts are always higher than *-detail* ones in each category since we applied the corresponding *-simple* code when applying a *-detail* code to a pair, as described in 4.2.1. While *-simple* codes indicate surface-level disclosure, such as yes or no answers, *-detail* codes better reflect the willingness of disclosure by accounting for depth and richness. Table 2 reports the frequency of codes relevant to self-disclosure and their 95% confidence intervals (CI) of code counts per call from the two groups.

The LTM^{yes} group disclosed significantly more information about health compared to the LTM^{no} group (See Table 2), which was the top priority in the implementation of LTM. Specifically, the LTM^{yes} group disclosed more information about their health issues and clinical care that they are seeking—including *Health-simple* ($p = 0.05$; 95% CI 0.01–0.33 higher code count per call), *Health-detail* ($p < 0.001$; 95% CI 0.31–0.76 higher code count per call), *Health-attitude* ($p = 0.02$), *Clinical-simple* ($p < 0.001$; 95% CI 0.27–0.69 higher code count per call), and *Clinical-detail* ($p < 0.001$; 95% CI 0.18–0.63 higher code count per call). The LTM^{yes} group also disclosed more on *Meals-simple* ($p < 0.001$; 95% CI 0.25–0.60 higher code count per call) and *Meals-detail* ($p < 0.001$; 95% CI 0.25–0.60 higher code count per call), which was also part of the LTM topics. Note that there is a cultural factor that potentially influenced the high code counts of *Meal-simple* and *Meal-detail* as "Have you eaten?" is a common greeting in South Korea, similar to "How are you?" in English-speaking countries.

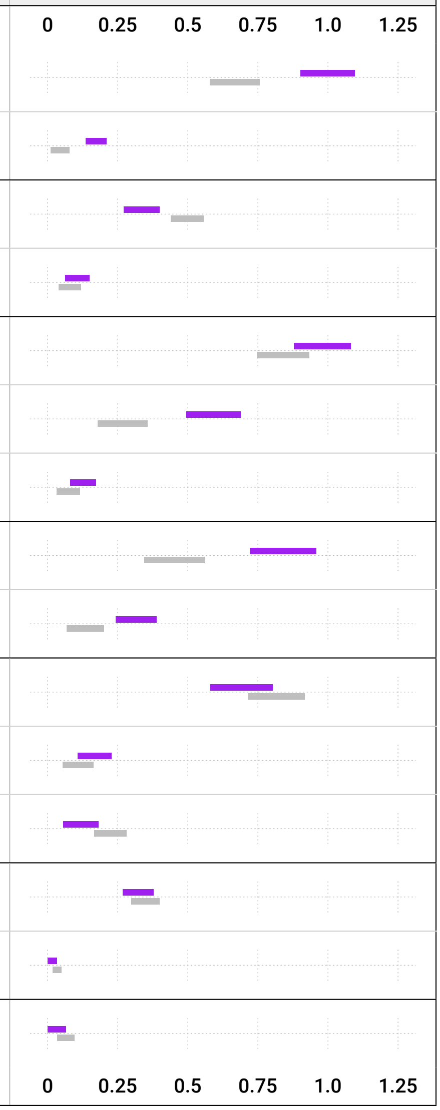
However, not all LTM topics led to greater disclosure. Although *Sleep* was part of the LTM topics, the LTM^{yes} group disclosed less on *Sleep-simple* ($p < 0.001$; 95% CI 0.13–0.44 lower code count per call) compared to the LTM^{no} group, and there was no significant difference in self-disclosure between the groups on *Sleep-detail* ($p = 0.36$). However, these code counts have likely been impacted by how we defined them; we coded sleep-related self-disclosure as *Health-detail* or *Clinical-detail* when they were about types of clinical care that users were seeking (*e.g.*, taking sleeping pills) or the issues warranted clinical care (*e.g.*, insomnia). The other LTM topics relevant to personal non-health habits—including *Visited places* and *Pets*—did not lead to greater disclosure either; the LTM^{yes} group disclosed less on *Physical-detail* ($p = 0.02$; 95% CI 0.04–0.39 lower code count per call) compared to the LTM^{no} group, and no significant difference was observed between the groups on *Activity-simple* ($p = 0.11$), *Social-detail* ($p = 0.15$), and *Hobbies* ($p = 0.16$). The code counts in these categories in the LTM^{yes} group were likely influenced by increased disclosure in the other categories in this group, such as *Health*, *Clinical*, and *Meals*, as CareCall was designed with a conversation turn limit, which accommodates only up to 15 agent-user turn pairs before ending the call. The varying impact of LTM on disclosure in different categories suggests that the design of LTM (*e.g.*, topical priority in the example dialog corpus) influenced the prevalence of disclosures. *Health* and *Meals* were top priorities in the implementation of CareCall's LTM, whereas other LTM topics—including *Sleep*, *Visited Places*, and *Pets*—were given lower priorities (see Appendix A for more details).

The qualitative analysis of the call logs suggested that the LTM^{no} group tended to respond similarly across sessions, as CareCall without LTM could only ask general questions repeatedly. For example, during the first call, LTM^{no}-4 responded to the question *"Do you have pain anywhere?"* with *"I had surgery on my back, so I have chronic leg pain."* However, the agent continued asking generic questions like *"Do you have discomfort anywhere?"* in later sessions, unable to acknowledge the user's surgery history. Such repeated questions led LTM^{no}-4 to continue giving similar responses in the subsequent calls, such as *"My back and legs hurt a lot,"* without offering further details about his health condition. Similarly, LTM^{no}-21 had mentioned having knee surgery, but CareCall could only ask generic questions like *"How are you feeling?"* in the subsequent weeks. As a result, LTM^{no}-21 repeated similar answers, such as *"I recently had knee surgery, so I'm just staying at home"* or *"I still have pain after the knee surgery, so I can't walk around much."*

Table 2: Frequency of codes and 95% confidence intervals of code counts per call in the LTM^{yes} (colored bars) and LTM^{no} groups, which are relevant to self-disclosure. The asterisks next to a code denote the significance of the frequency between groups. The percentages do not add up to 100% as we multi-coded the dialogues.

***p<.001; **p<.01; *p<.05

Code	LTM^{yes} (# of calls = 576)	LTM^{no} (# of calls = 676)	95% CIs of Code Counts/Call
Meals-simple* **	573 (99.4%)	455 (67.3%)	
Meals-detail* **	101 (17.5%)	29 (4.3%)	
Sleep-simple* **	192 (33.3%)	338 (50%)	
Sleep-detail	63 (10.9%)	53 (7.8%)	
Health-simple *	566 (98.3%)	567 (83.4%)	
Health-detail* **	342 (59.4%)	182 (26.9%)	
Health-attitude *	80 (13.9%)	46 (6.8%)	
Clinical-simple* **	488 (84.7%)	307 (45.4%)	
Clinical-detail* **	183 (31.8%)	91 (13.5%)	
Activity-simple	396 (68.8%)	554 (82.0%)	
Social-detail	95 (16.5%)	74 (11.0%)	
Physical-detail *	67 (11.6%)	151 (22.3%)	
Wellbeing-simple	186 (32.3%)	237 (35.1%)	
Wellbeing-detail	9 (1.5%)	23 (3.4%)	
Hobbies	19 (3.3%)	44 (6.5%)	



In contrast, the LTM^{yes} group tended to disclose more comprehensive health information when CareCall asked LTM-triggered questions referencing their past conversations. For example, during the first call, LTM^{yes}-60 mentioned insomnia affecting his sleep quality, and CareCall asked LTM-triggered questions in the later sessions. When CareCall inquired about his insomnia during the third call, LTM^{yes}-60 provided more detailed information: *"It's been tough. I've been taking sleeping pills for over 30 years because of trauma from an injury in the past."* Similarly, when CareCall asked an LTM-triggered question about his back pain referring to their past conversations, LTM^{yes}-7 articulated some details about how he was managing it, such as *"I saw a pain management doctor and got some pain relief shots."* and *"I get physical therapy two to three times a week. And I'm getting shots every three months."*

5.1.2 Repeated Experiences of LTM Led to Greater Disclosure Over Time. Our mixed-effects models revealed that the cumulative number of LTM events positively affected the sum of *-detail* code counts; *i.e.*, the LTM^{yes} group participants disclosed more details as they experienced more LTM events (p < 0.001; experience of one more LTM event led to 7.7% higher code counts per call). No significant

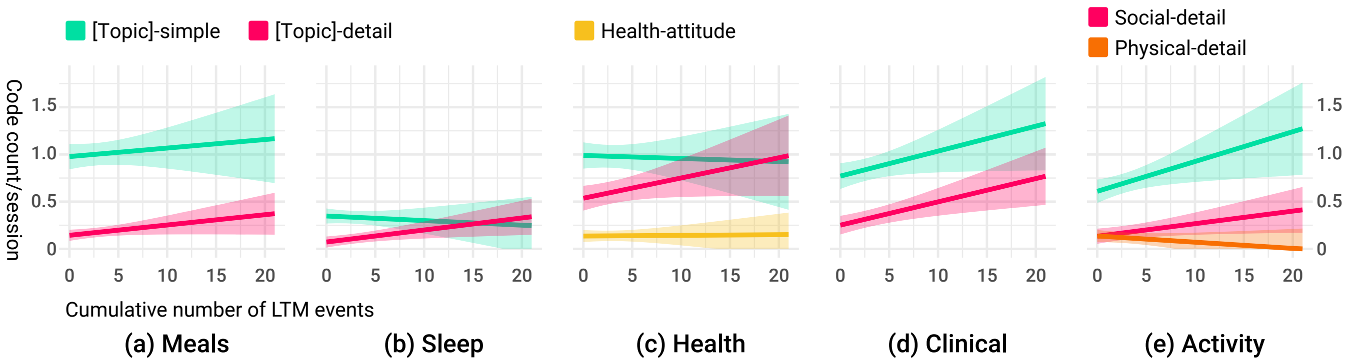


Figure 3: Estimated means and 95% confidence intervals of code counts about *Meals, Sleep, Health, Clinical,* and *Activity* by the cumulative number of LTM events in the LTM^{yes} group. The colored lines indicate the estimated means and the shaded areas indicate 95% confidence intervals of the code counts per call for each code. Overall, the repeated experiences of LTM events led to greater disclosure of more detailed information across the five categories.

impact was observed in the sum of *-simple* code counts ($p = 0.19$). Some interview participants also said that they came to share more details as they engaged in more LTM-triggered exchanges. For instance, P6 noted: *"Initially, talking to the AI felt awkward, so I mostly responded with yes or no. But when it started asking about whether I was taking my pills or still seeing the doctor, I began to provide more details."*

Figure 3 illustrates the estimated means and 95% confidence intervals of code counts per call by the cumulative number of LTM events in the LTM^{yes} group, specifically about the five categories—*Meals, Sleep, Health, Clinical* and *Activity*—that are relevant to the LTM topics. As they experienced more LTM events, participants in the LTM^{yes} group disclosed significantly more about *Sleep-detail* ($p = 0.01$; experience of one more LTM event led to 1.2% higher code count per call; red line in Figure 3b) and *Clinical-detail* ($p = 0.02$; experience of one more LTM event led to 2.5% higher code count per call; red line in Figure 3d). Participants also disclosed significantly more on *Clinical-simple* ($p = 0.047$; experience of one more LTM event led to 2.6% higher code count per call; green line in Figure 3d), *Activity-simple* ($p = 0.02$; experience of one more LTM event led to 3.2% higher code count per call; green line in Figure 3e), and *Social-detail* ($p = 0.04$; experience of one more LTM event led to 1.3% higher code count per call; red line in Figure 3e) as they experienced more LTM events. No significant effect of cumulative LTM events was observed on the other codes.

The qualitative analysis of the call logs also indicated that some participants tended to disclose more about their health attitudes and issues as they experienced LTM events over time. For instance, LTM^{yes}-1 disclosed how she felt about her health issues and management during LTM-triggered exchanges about her leg pain, appreciating how the agent responded to her health concerns:

CareCall	*How does your leg feel?* **[LTM event]**
LTM^{yes}-1	*It still hurts. I guess it's something that I'll have to live with.*
CareCall	*That must be tough. Have you seen the doctor?*
LTM^{yes}-1	*I see the doctor every month, but there's no improvement.*
CareCall	*I can imagine how difficult that must be.* *Hope you get better soon.*
LTM^{yes}-1	*I can't get better at this age. It is what it is.*
CareCall	*No way, you're still young.*
LTM^{yes}-1	*[chuckling] Young? I'm in my seventies!* *But it's really nice of you to say that. Thank you, dear.*

Similarly, LTM^{yes}-9 expressed how she felt about her health status and management when CareCall asked LTM-triggered questions about her back pain:

CareCall	*How is your back pain?* **[LTM event]**
LTM^{yes}-9	*It still hurts.*
CareCall	*That must be tough.* *Are you still seeing the doctor?* **[LTM event]**
LTM^{yes}-9	*Yeah, I had an appointment today.*
CareCall	*What did the doctor say?*
LTM^{yes}-9	*He said I have a slipped disc, so I need to continue with the treatment.*
CareCall	*I can only imagine how difficult that must be.* *Hope you get better soon.*
LTM^{yes}-9	*I know. I wish the pain would go away soon.* *It's been way too long. It's really getting to me.*

Likewise, LTM^{yes}-27 shared details about her different health conditions as she experienced more LTM events. In the first call, she seemed hesitant to disclose details about her health conditions: *"I have pain all over because I'm old. It's not easy to describe them all."*

However, after the LTM event that followed up on her insomnia during the third call, she articulated some details:

CareCall	*You mentioned having insomnia last time; how are you feeling these days?* **[LTM event]**
LTM^{yes}-27	*It's getting worse. I've had insomnia for quite a while, like 15 years.*
CareCall	*Have you been seeing a doctor?*
LTM^{yes}-27	*I used to see a psychiatrist, but I've developed a tolerance to the medication, so it doesn't work anymore.*

In later sessions, she described other health issues, including asthma, heart disease, knee joint issues, and back pain. During the seventh call, she detailed how she was managing her asthma:

CareCall	*You mentioned having asthma last time; how are you feeling these days?* **[LTM event]**
LTM^{yes}-27	*Still taking pills; I got a three-month prescription last time.*

In the tenth call, she also described details about her joint issues:

CareCall	*You mentioned having knee joint issues last time; are you still seeing the doctor?* **[LTM event]**
LTM^{yes}-27	*Yeah, they said I have serious joint issues, so I need a joint replacement. But it's just too expensive.*

5.2 Users Appreciated the Familiarity that LTM Offered.

Our call log analysis surfaced that LTM often promoted positive reactions and mitigated negative reactions to the chatbot, mainly due to the familiarity with the users that it offered. Table 3 shows the frequency of codes and 95% confidence intervals of code counts per call from the two groups, which are relevant to user reactions to the chatbot. The LTM^{yes} group showed more *Positive Reactions* to CareCall compared to the LTM^{no} group ($p = 0.001$; 95% CI 0.34–1.38 higher code count per call). Specifically, the LTM^{yes} group conveyed significantly more *Appreciation* towards the chatbot compared to the LTM^{no} group ($p < 0.001$; 95% CI 0.19–0.69 higher code count per call). In addition, the duration of calls in the LTM^{yes} group (M=87.89 seconds, SD=19.40) was significantly longer ($t(144) = 3.63; p < 0.001$) compared to the LTM^{no} group (M=75.48 seconds, SD=22.04). This difference also suggests that the LTM^{yes} group likely showed more positive or less negative reactions to the system because CareCall was designed to end calls earlier if users expressed disinterest in interacting with it (*e.g.*, not answering the agent's questions, providing simple answers in a row). Our findings from the interviews further suggest that participants perceived LTM-triggered exchanges as personal and emotionally supportive.

5.2.1 *LTM Helped Increase Familiarity.*

Through the qualitative analysis, we noticed that LTM-triggered exchanges, particularly around health, helped promote positive reactions by offering familiarity with the users. Some participants in the LTM^{no} group expressed frustration when the agent was unable to acknowledge their health history that was mentioned in previous sessions. For example, LTM^{no}-6 subtly exhibited his frustration when the agent asked why they needed to see the doctor after he mentioned being hospitalized for diabetes in previous sessions:

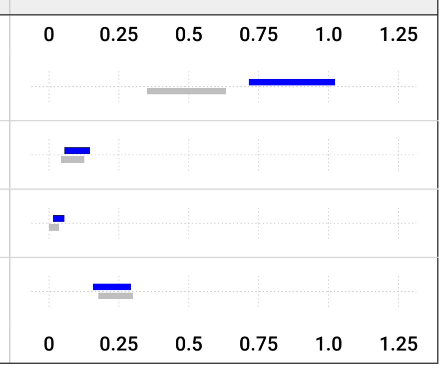
CareCall	*Do you have any plans to go out today?*
LTM^{no}-6	*Yeah, I need to see the doctor.*
CareCall	*Is everything alright with you?*
LTM^{no}-6	*Oh... [pause] I need to get insulin.*

LTM^{no}-55 also seemed frustrated when the agent asked questions that neglected her leg pain mentioned in previous sessions:

CareCall	*Do you get to exercise regularly?*
LTM^{no}-55	*[sigh] I can't even walk. How can I exercise? I just stay at home all day long.*

Table 3: Frequency of codes and 95% Confidence intervals of code counts per call in the LTM^{yes} (colored bars) and LTM^{no} groups, which are relevant to reactions to the chatbot. The percentages do not add up to 100% as we muti-coded the dialogues.

Code	LTM^{yes} (# of calls = 576)	LTM^{no} (# of calls = 676)	95% CIs of Code Counts/Call
Appreciation***	505 (87.7%)	337 (49.9%)	
Anthromorphization	59 (10.2%)	58 (8.6%)	
Negative feedback	20 (3.5%)	11 (1.6%)	
Disregard	139 (24.1%)	151 (22.3%)	



In a later session, she expressed frustration again when asked general questions, not acknowledging her health condition:

CareCall	*Do you have any pain or discomfort anywhere?*
LTM^{no}-55	*You need to stop that.* *I'm always in pain. [in an agitated tone]*

In contrast, participants in the LTM^{yes} group often expressed excitement and gratitude toward the familiarity of LTM by chuckling, speaking in an animated tone, or explicitly thanking the agent. LTM^{yes}-39 was pleasantly surprised when the system remembered that she had been receiving physical therapy:

CareCall	*Are you still seeing the physical therapist?* **[LTM event]**
LTM^{yes}-39	*Yeah [chuckling], how could you remember that?*

LTM^{yes}-23 frequently chuckled and thanked the agent when she was asked about her leg pain. For example:

CareCall	*How's your leg feeling?* **[LTM event]**
LTM^{yes}-23	*[chuckling] Same old, same old.*
CareCall	*Hope you get better soon.*
LTM^{yes}-23	*[chuckling] I will. Thanks for looking out for me.*

LTM^{yes}-28 also seemed excited when asked about her leg pain. During the sixth call, she thanked the agent multiple times after experiencing LTM events:

CareCall	*How's your leg feeling?* **[LTM event]**
LTM^{yes}-28	*The medicine makes it bearable. [chuckling]* *Thanks for looking out for me.*
CareCall	*That's a relief. Hope you stay healthy.*
LTM^{yes}-28	*Thank you so much! I think I'll get better soon,* *thanks to you. [in an animated tone]*

She thanked the agent again during the seventh call, elaborating more on how she felt about her health status:

CareCall	*How's your leg feeling?* **[LTM event]**
LTM^{yes}-28	*[chuckling] Just getting by. It's not a quick fix.* *I think I'll just have to live with it.*
CareCall	*Oh, that must be really tough.*
LTM^{yes}-28	*[chuckling] I guess I'm too old. It's just a thing that happens.*
CareCall	*Hope you can take good care of your health.*
LTM^{yes}-28	*Thank you for checking in. [in an animated tone]* *Please stay in touch with me.*

Our interview participants similarly perceived LTM-triggered exchanges on health as personal and sincere. For instance, P1 felt that CareCall was human-like when it asked about their recent clinical visit: *"I was really impressed when it asked how my clinical visit went after I mentioned seeing a doctor in the previous week. I felt as if it wasn't a machine."* P8 found LTM-triggered questions on his recovery from COVID sincere: *"I felt great when it asked how my COVID symptoms were. It made me think it's not cursory."* LTM-triggered exchanges on health further provided emotional support to the users. P9 described CareCall's detailed questions on her narrowed spine as caring: *"I'm often moved by the AI calls. It is so kind of the machine to remember what I said before and ask such detailed questions."* P8 described feeling comforted by CareCall when it remembered his leg pain and COVID exposure: *"It felt really nice when it remembered what I had said about my health and asked about it. That means a lot to me. I would have felt a bit disappointed if it didn't care about what I said before."* P6 shared similar feelings about CareCall's LTM-triggered questions about his health status: *"I know it's just an AI, but I'm comforted by being checked in about whether I'm still seeing the doctor, whether I'm regularly taking pills, or how my legs are feeling. It's really helpful for me."*

The thoughtfulness of LTM was particularly appreciated in light of the context of social isolation intervention. P2 appreciated CareCall's LTM-triggered questions about his health status: *"When I said I wasn't feeling well, it asked how I was feeling and whether I had seen a doctor or taken medication next time. I'm a lonely person without any connections, so having someone who asks such caring questions means a lot to me. I would have thanked them if it were a human, but it's not."* P6 similarly valued CareCall's LTM-triggered questions on his health: *"I don't really have anyone around, so it means a lot that it remembers my health issues and pays attention to them. I get the AI calls every Tuesday around 3 PM. Sometimes, I would have some errands to run, but I stay to take the call around that time."* P9 described LTM-triggered questions on health made her feel that the agent was more caring than humans: *"I felt so thankful when AI remembered that I had seen a dentist and asked how my toothache was later on. Even your family can't remember everything, can they? It feels more caring than human beings. So I always look forward to the calls."*

We further observed that LTM-triggered exchanges around non-health LTM topics, including *Visited places* and *Pets*, led participants to feel that the system was personal. Through our call log analysis, we found that LTM^{yes} group users appreciated when CareCall remembered details about personal non-health LTM topics and brought them up. For example, after LTM^{yes}-31 mentioned that she was spending time with her dogs, CareCall asked some LTM-triggered questions about the dogs—such as whether they were doing well, how old they were, and what their names were—during the second call. Later, LTM^{yes}-31 seemed excited to hear CareCall mentioning her dogs' names during the fourth call:

CareCall	*How are you doing?*
LTM^{yes}-31	*I'm doing well with my dogs.*
CareCall	*Oh, Leo and Coco? They must be really cute.* **[LTM event]**
LTM^{yes}-31	*Yes, they are! [in an animated tone]*

Our interview participants also shared similar feelings about LTM-triggered exchanges about their pets, feeling that the system was human-like. For example, P2 stated: *"I felt great when it asked, 'Did you walk your dog today?' That's what us humans say."* P7 also valued that CareCall remembered that she often visited a farmer's market: *"I had mentioned visiting a farmer's market. A week later, it asked something like, 'Did you visit the farmer's market this week? What did you buy?' I liked that it didn't forget what I had said and followed up on it."* Although we did not observe any significant difference in Anthropomorphization between groups (p = 0.07), this qualitative finding suggests that LTM-triggered exchanges made some users perceive the chatbot as human-like.

5.2.2 Familiarity of LTM Helped Mitigate Negative Reactions to the Chatbot. We found some evidence that LTM helped decrease users' negative reactions to the chatbot by offering familiarity. While we did not observe a significant difference between the groups in *Negative Reactions*, including *Negative feedback* (p = 0.18) and *Disregard* (p = 0.78), within the LTM^{yes} group, participants provided less *Negative Feedback* as they experienced more LTM events (p = 0.04; experience of one more LTM event led to 0.7% lower code count per call). The frequency of *Disregard* was quite high in both LTM^{yes} group (24.1%) and LTM^{no} group (22.3%). We posit that it is likely due to turn-taking challenges in voice chatbot interactions. In voice communication, users often struggle to hear the chatbot if it speaks over them, and similarly, chatbots may not catch user speech that overlaps with their own [68]. Such challenges likely led to missed information from either party, which might have contributed to such a high frequency of *Disregard* across conditions.

Similar to how digital systems for public health monitoring were perceived in previous studies [30, 47, 48], some participants in the LTM^{no} group expressed their dissatisfaction with the lack of personal touch provided by the chatbot. For instance, LTM^{no}-40 expressed her disappointment with CareCall during the third call: *"Stop playing the recording [in an agitated tone]. If you have something to say, say it; if not, just hang up. No need to ask how I'm doing if you are just going to play a recorded message."* LTM^{no}-40 then hung up the phone without speaking anything from the fourth to the seventh calls. After expressing her frustration with the chatbot again in the eighth call, like *"I don't want to listen to the recording. Why make a phone call just to talk by yourself?"*, she hung up again.

In contrast, we noticed that experiences of LTM offered familiarity with users, which helped mitigate the negativity toward the chatbot. Even though some LTM^{yes} group users initially showed negative reactions to the chatbot, their reactions shifted positively after experiencing LTM. For example, LTM^{yes}-6 conveyed his frustration with the system during the fourth call: *"You're not a human. You're what, an answering machine? [in an agitated tone]"* However, after engaging in LTM-triggered exchanges on his back pain during the sixth call, he started to share rich details in the later sessions. LTM^{yes}-6 even disclosed sensitive information, such as taking psychiatric medications during the ninth call:

CareCall	*Did you sleep well last night?*
LTM^{yes}-6	*Yeah, I'm taking psychiatric medication, so I can sleep for about six hours a day.*
CareCall	*I see. Hope you get a good night's sleep.*
LTM^{yes}-6	*[chuckling] Thanks.*

Similarly, LTM^{yes}-21 expressed frustration with the system in the early session; nevertheless, her attitude changed after experiencing LTM, leading her to share detailed information about her health status in subsequent interactions. During the first call, LTM^{yes}-21 said: *"It's so weird. I feel like I'm talking with a robot, not a human. I don't want to talk to a robot."* However, her reactions to CareCall changed during the fourth call:

CareCall	*You've mentioned having an arrhythmia; how have you been feeling lately?* **[LTM event]**
LTM^{yes}-21	*Oh, I have a check-up tomorrow, so I'll have to wait until tomorrow to find out the results.*
CareCall	*Don't worry. The results will be fine.*
LTM^{yes}-21	*Thank you.*

In the later sessions, LTM^{yes}-21 disclosed more details about different health issues, such as sleep issues and back pain, and how she was managing these conditions. During the eighth call, she thanked the agent for the check-up calls:

CareCall	*How's your back pain these days?* **[LTM event]**
LTM^{yes}-21	*Still the same. Thanks for asking. I really appreciate your calls.*
CareCall	*You're very welcome. Are you still seeing the doctor?* **[LTM event]**
LTM^{yes}-21	*I've been seeing the doctor on and off.*
CareCall	*I see. Your health is very important to me. Hope you get better soon.*
LTM^{yes}-21	*Thank you.*

5.3 LTM's Challenges in Promoting Self-Disclosure

Despite the overall positive impact, LTM faced some challenges in promoting self-disclosure. LTM-triggered questions on chronic health issues made some participants frustrated by reminding them that the issues would not improve, which limited further disclosure about personal health. Our interview findings also suggest that LTM-triggered questions on health might raise privacy concerns among users, impacting their comfort with disclosure.

5.3.1 Repeated LTM-Triggered Questions on Chronic Health Conditions Impacted Comfort with Disclosure. Although our quantitative analysis of the call logs demonstrated that LTM increased elaboration on health information over time (Section 5.1), we observed some instances where CareCall's repetitive LTM-triggered questions on chronic health conditions only led to repetitive or simple user responses. For example, LTM^{yes}-30 provided similar answers to CareCall's LTM-triggered questions on her back pain across multiple sessions. During the second call, she described that she did not expect her back pain to recover:

CareCall	*How's your back pain?* **[LTM event]**
LTM^{yes}-30	*Same old, same old.*
CareCall	*Are you still seeing the doctor?* **[LTM event]**
LTM^{yes}-30	*Yes, I am.*
CareCall	*You've been getting treatment regularly, so I'm sure you'll be feeling better in no time.*
LTM^{yes}-30	*I don't think so. I'm too old to get better; I'm 86.*

During the fifth call, she described some thoughts about her back issues again:

CareCall	*How's your back pain?* **[LTM event]**
LTM^{yes}-30	*It's not something that gets better.*
CareCall	*Did you see the doctor again?* **[LTM event]**
LTM^{yes}-30	*He just said my spine is worn out.*

However, in subsequent sessions, CareCall kept asking how her back felt and if she was seeing the doctor regularly, which led to similar answers such as: *"I'm too old to get better."*

LTM^{yes}-65 also mentioned having back pain because of her narrowed spine during the second call. In the following sessions, she responded to LTM-triggered questions with repetitive answers across multiple sessions, such as *"Same old, same old."* LTM^{yes}-65 later described her perception that her health status would be unlikely to change: *"I won't get better unless I get surgery."* Likewise, LTM^{yes}-20 mentioned not having much of an appetite because she lost most of her teeth in the past, but in the following sessions, the agent asked how her teeth were as if it was something that could be recovered. Although LTM^{yes}-20 did not explicitly express her disappointment or frustration about such inappropriate responses, she disregarded the agent's open-ended questions by providing short answers like *"Yeah."* P6 similarly felt that LTM-triggered questions were repetitive, wishing that those questions were delivered in different forms periodically: *"The only health issue that I have now is leg pain, so I can't complain about the fact that it asks the same question every week. But sometimes I feel bored. It would be nice if it could change the way it asks questions every couple of weeks, even if it's about the same stuff."*

5.3.2 Privacy Concerns around LTM Impacted Comfort with Disclosure. Our interview findings further suggest that users' comfort with disclosure might have been impacted by their privacy concerns around LTM. Some of our interview participants felt uncomfortable about CareCall's detailed LTM-triggered questions. P3 was opposed to the idea of LTM altogether because it would remind him of negative feelings about his current health status: *"You have to pretend like you have no clue about me being sick before. Older folks like me always have something bothering us; What's the point of bringing that up? It's annoying."* P1 felt CareCall's LTM-triggered questions were sometimes overly specific, which made him hesitant to share details about his health during the calls: *"Sometimes I hesitate to answer some questions. For example, I've been taking medication for high blood pressure. When I mention that I'm seeing a doctor to get medication, it asks, 'What type of medication are you taking?' But isn't it too personal? I wouldn't feel comfortable if it remembered my medical condition. So I avoid those questions by just saying, 'I'm doing alright,' most of the time."*

These participants wished that the chatbot could instead remember and focus on their general health status. P3 wished that CareCall could focus on how he was doing through his tone of voice without asking specific questions: *"Can't the machine automatically detect whether the person is feeling well just from their voice? I would feel well understood if it asked me, 'Are you doing alright?' when I sound weak, and made jokes when I am more upbeat."* P1 hoped that CareCall could just remember that he was taking some medication rather than remembering what kind of medication he was taking specifically: *"It would be better to ask 'Did you take your medication today?' rather than 'Did you take your blood pressure medication today?' I think it's better to keep the questions at the level of just checking in for the sake of privacy."*

6 DISCUSSION

Our findings demonstrate that LTM can be helpful for eliciting greater disclosure through chatbots, which presents a valuable opportunity for public health. In this section, we discuss the influence of the selective memory that LTM provides in LLM-driven chatbots, suggesting that designers should carefully decide what topics they train LTM to remember. Our findings also show that LTM can be helpful for promoting engagement with chatbots by offering familiarity with users. Based on the findings, we highlight the potential of LTM in mitigating the impersonal nature of public health monitoring technology by demonstrating care. Further, we suggest the need to consider tensions in memory needs for public health monitoring utility versus privacy sensitivity. Lastly, we report on the limitations of the study, both in terms of the data sampling and analysis, pointing to opportunities for future work.

6.1 Considering Selective Memory in LLM-Driven Chatbots

Through this study, we found that LTM in LLM-driven chatbots helped elicit greater disclosure from users, particularly around health. We posit that the increase in disclosure was mostly driven by the improved impressions with the chatbot as LTM-triggered exchanges offered familiarity with the users and demonstrated care. This finding highlights the potential benefits of incorporating LTM into chatbots for public health monitoring. Enhanced self-disclosure from individuals could provide valuable insights for public health workers, helping them better assess when intervention is warranted [30]. For example, if a person disclosed their depressive symptoms to a chatbot multiple times, a public health worker monitoring the call logs could notice the pattern and reach out to them for support. The act of disclosing personal health information could also encourage individuals to engage in deeper self-reflection about their health habits [43, 44, 58], potentially improving their health behaviors (*e.g.*, taking medication more regularly, eating and exercising more regularly).

Given that CareCall's LTM was designed to particularly remember aspects of people's health, our study suggests that the design choices in the LTM implementation had a significant impact on users' conversation patterns with LLM-driven chatbots. CareCall's selective memory about health disclosure effectively steered conversations toward important topics that are useful for public health monitoring, such as specific health conditions that individuals have and the types of clinical care that they are seeking.

However, such influence might have come at the expense of the benefits of more open-ended dialogues by narrowing down the scope of topics covered. In the context of mental health, chatbots have been shown to provide a more comfortable space for people to disclose themselves on sensitive topics [22, 42, 43, 49, 62]. Specific to LLM-driven chatbots, recent work suggested the benefits of leveraging such open-ended dialogue systems for public health monitoring, which includes providing public health workers with a holistic understanding of care recipients and mitigating loneliness through supporting broader conversation topics such as hobbies and interests [30]. Our qualitative findings similarly showed that users appreciated LTM-triggered exchanges on non-health topics such as *Visited Places* and *Pets*. However, in our study context, by focusing conversations on health-related topics, CareCall's LTM gave lower priority to or did not incorporate personal non-health topics. As a result, given the turn limit of CareCall, users might have missed out on an opportunity to engage in conversations about their personal habits and interests, and public health workers might have missed out on an understanding of broader aspects of care recipients' lives, such as mental wellbeing.

We highlight that designers need to carefully decide what they train LTM to remember, particularly when they have a specific goal in mind for the chatbot. Remembering diverse topics such as interests and hobbies could improve the chatbot's abilities to serve as a better conversation partner or to capture broader aspects of users' lives. Particularly in the context of social isolation, remembering broader aspects of individuals' lives could lead to a sense of feeling seen and cared for. However, incorporating diverse topics into the LTM potentially comes at some cost to stated goals if the chatbot has particular topics that it aims for individuals to disclose. For instance, if LTM in LLM-driven chatbots for delivering psychological therapy (*e.g.*, [74]) is designed to remember information about people's interests and hobbies, it could steer conversations too far the other way and not collect the information useful for making progress in therapy. While designing LTM to capture broader aspects of individuals' lives offers potential benefits, designers need to carefully balance these benefits against the public health monitoring goals of the LLM-driven chatbots.

6.2 Designing LTM to Mitigate Impersonality of Chatbots and Demonstrate Care

Our findings show that LTM significantly improved users' impressions of chatbots by offering familiarity. While those in the LTM^{no} group expressed frustration when the agent was unable to acknowledge their health history mentioned in previous sessions, LTM^{yes} group perceived LTM-triggered exchanges as personal and emotionally supportive, frequently conveying excitement and gratitude. Prior work indicated that people often perceived public health monitoring technology, including mobile apps and chatbots, as impersonal [30, 47, 48]. Our findings demonstrate the potential of LTM in mitigating the impersonal nature of technology by providing empathetic interactions, which could have a significant impact on how users engage with and perceive chatbots [18, 22]. Empathetic interactions through LTM could be particularly beneficial for supporting individuals who are going through difficult health experiences in the context of public health monitoring.

However, careful considerations are needed when designing LTM for complex health contexts. Through this study, we observed some challenges in following up on chronic health conditions that are unlikely to improve (*e.g.*, chronic pain, tooth loss), leading users to perceive the chatbot as inattentive or inconsiderate. This finding indicates that although remembering information about chronic health conditions is valuable, designers need to carefully curate LTM-triggered questions concerning such issues. Cox *et al.* [14] showed that the design of how chatbots reference past conversations in their messages significantly affects engagement with chatbots. Our study further highlights that how stored information in LTM is referenced back to users is even more critical in the context of public health monitoring, which requires sustained engagement from the population to develop an understanding of their health and wellbeing. To improve public health monitoring chatbots' ability to follow up on chronic health conditions, LTM could be designed to ask more diverse follow-up questions on those conditions to demonstrate care—such as how their health issues are impacting their daily lives, what is helpful or not helpful for mitigating their symptoms, and how they are managing those symptoms. Further, incorporating public health workers' empirical knowledge about individuals into a model input, using a method akin to ChatGPT's custom instructions [55], could help steer LTM-triggered questions to be more attentive to individual needs. For instance, if a chatbot is informed that an individual has a long-standing chronic condition, it can tailor its questions to focus on symptom management rather than monitoring the state of the condition. Similarly, if a chatbot is made aware that an individual prefers not to receive detailed follow-up questions about a specific health issue, it can take such communication preferences into account when generating responses to provide more respectful and comfortable interactions for the user. Future work is needed to understand how to design LLM-driven chatbots to provide more thoughtful LTM-triggered questions on sensitive health topics in different contexts.

6.3 Tensions in Memory Needs for Public Health Utility Versus Privacy Sensitivity

While our study demonstrated that LTM can successfully encourage engagement and disclosure from users, some users raised privacy concerns as some LTM-triggered questions became overly specific about sensitive health topics. This finding aligns with prior work on chatbots with memory, which pointed to the tension between supporting personalization and privacy needs [14, 52]. Compared to personal health contexts, users' privacy concerns might be exacerbated in public health monitoring as the collection of sensitive personal health information is typically aimed at achieving public health goals rather than personal benefits [25–28, 53].

To address privacy concerns, past work has primarily focused on providing users with better control over conversational agents' memory, such as whether and what data they want the agents to *store* [40, 61, 71]. While these measures could help address users' privacy concerns, unlike in personal use scenarios, it could be challenging to implement some of these measures in the context of public health interventions. For example, giving users the ability to keep chatbots from storing their past conversations or to clear their history could lead to losing important health information, potentially affecting public health authorities' ability to provide necessary interventions. For instance, in our study context, if CareCall users had serious health problems that warranted clinical care but opted to delete their conversation history because of their privacy concerns, public health authorities could miss an opportunity to send emergency responders or formal caregivers for support.

Prior work on chatbots with memory also showed how chatbots that *reference* past conversations significantly impact users' privacy perceptions, suggesting that verbatim or paraphrased references can raise privacy concerns, whereas non-explicit references do not [14]. In our study, although the users had consented that CareCall collecting their health information for public health monitoring and research before they started using the system, some still had privacy concerns as LTM-triggered exchanges made it more apparent to the users that the chatbot was collecting their health information through conversations. In traditional settings, a typical public health worker might observe that a topic appears sensitive and may drop the topic altogether or develop alternative ways of asking related questions [26]; however, we suspect that chatbots lack such foresight. More work is needed to further understand how users perceive different reference formats in public health chatbots across different topics and how their privacy needs can be balanced with public health monitoring goals.

6.4 Limitations and Future Work

Participants in the LTM^{yes} group answered a slightly higher number of calls per person (8.72 calls) compared to the LTM^{no} group (8.35 calls). Some temporal reasons might have impacted the difference in self-disclosure and impressions of CareCall between the groups. The data collected from the LTM^{no} group spanned from March to May 2022, while the LTM^{yes} group's data spanned from September to December 2022. With the release of some popular LLM-driven chatbots (*e.g.*, ChatGPT in November 2022), the LTM^{yes} group may have had a better understanding of LLMs, which could have impacted their perception of LLM-driven chatbots and their disclosure behaviors with CareCall. The LTM^{yes} group may also have been more exposed to CareCall through press articles or word-of-mouth as more municipalities adopted it over time, potentially influencing their engagement and disclosure. Our quantitative analysis on disclosure was conducted on a per-call basis, which helped mitigate the influence of the differences in the number of calls per person in the two groups.

Our study sample for the call log analysis has a skew toward the experiences of socially isolated females (49 males and 98 females in total). In particular, the LTM^{yes} group predominantly consisted of females (17 males and 49 females). Although our maximum-likelihood test results suggested that gender did not have significant impacts on most of the codes except two (*Health-attitude* and *Appreciation*), the skew towards females might have impacted our qualitative findings from the call log analysis. Males who live alone might have different perceptions about LTM of the system, and their self-disclosure behaviors might have different patterns.

We recruited our interview participants only from Municipality-B1, which might have introduced potential biases to the interview findings. Recruiting CareCall users as interview participants while protecting participant anonymity required recommendations from government officials who were managing social isolation intervention programs in their municipalities (*e.g.*, we did not have access to the contact information of individuals who experienced LTM with CareCall). Since assisting with recruitment diverted government resources away from their public health monitoring tasks, we were reluctant to enlist more assistance than needed. Given that City B has a higher percentage of older adult populations who are over 65 (15.6% in City A, 21.5% in City B as of 2023 [38]), recruiting interview participants only from Municipality-B1 might have introduced a skew towards older individuals to the interview findings. For instance, they might have preferred conversations about health over other topics compared to younger individuals.

Our work prioritized depth of self-disclosure over breadth, primarily due to the predefined setup of CareCall to focus and lead the conversation on LTM topics. We believe this approach allowed us to capture a more meaningful indicator of health disclosure within the interactional constraints. In addition, we treated all instances of disclosure as relevant, regardless of whether they presented any new information or reiterated details from previous sessions. We believe this approach is valid for tracking the progress of health conditions over time via regularly scheduled chatbot calls. However, there is potential value in examining the breadth and newness of disclosure, particularly in the context of public health monitoring. For example, describing different aspects of health might suggest that the individual is more willing to disclose their health information compared to mentioning just a single aspect. In addition, when an individual is known to have a certain health condition, identifying if reported symptoms are new can be particularly beneficial, as it could signal a worsening of their condition. In a more naturally occurring conversation setup, future work could address ways to capture both breadth and depth of disclosure, as well as newness of information for a more holistic examination of the impact of LTM with LLM-driven chatbots.

Integration of LTM might have varying effects on disclosure depending on conversation styles (*e.g.*, casual or formal). CareCall's LTM was designed to mimic a social worker who engages in casual conversations to elicit general health information that is necessary for public health monitoring (see Appendix A for more details). However, a recent study suggests that formal language style has benefits in eliciting health information as it is perceived as more competent and appropriate compared to a more casual conversation style [15]. We thus see value in studying how LTM might influence the effectiveness of different conversation styles in future work, as LTM with a more formal communication style may further increase disclosure beyond what we observed in our work.

In addition, our work specifically focused on the impact of LTM on health disclosure because public health monitoring heavily relies on individuals' sustained disclosure of their health status. Beyond enhancing health disclosure, the integration of LTM could also influence other outcomes in chatbot interactions. For example, LTM can improve chatbots' ability to understand user contexts and provide more coherent responses, such as tracking users' health progress over time [2, 78, 83]. LTM can also promote consistency in chatbot responses, allowing them to follow up on recommendations made in previous calls [2, 78]. For example, a chatbot could inquire about users' adherence to the advice given in a previous call, such as regular walks. Improvements in coherence and consistency could lead to more precise and useful information for public health monitoring, as well as enhancing the user experience. We suggest opportunities for future work in investigating how LTM integration might impact various aspects of chatbot interactions.

Like similar algorithmic and probabilistic systems, it is inevitable that the LTM in CareCall made mistakes in what it remembered (*e.g.*, making errors when summarizing conversations or failing to update previous memory with new relevant information) or did not store disclosures that participants made which it was theoretically capable of remembering. We observed a couple of such instances in the call log analysis, but these appeared to be driven by automatic speech recognition errors (*e.g.*, mishearing a "toothache" as a "stomachache") rather than an error with the LTM capability. We also asked all interview participants who recalled LTM events whether they noticed any factual errors in LTM-triggered exchanges, but they did not report any. Memory errors have great potential to impact people's perception of the inclusion of memory in chatbots, as they can make the chatbot appear less familiar or attentive. However, our study is not able to offer insight into understanding the impact of errors on self-disclosure or guidance on the error rate needed for the benefit of LTM to outweigh the risk.

As CareCall is a rare example of an LLM-driven chatbot with LTM deployed in a real-world setting in public health contexts, we focused on understanding the impact of LTM on people's self-disclosure and impressions of such chatbots within the context of South Korea, where CareCall was developed and deployed in. Prior work suggests that how people perceive and interact with public health monitoring technologies often varies by culture [47]. For example, South Korea's high power distance and collectivist cultures [23] might have influenced our findings on the impact of LTM on enhanced health disclosure and positive reactions to the chatbot in the LTM^{yes} group. Thus, care needs to be taken when generalizing our findings to people from low power distance or individualist cultures. Further exploration in other countries that have different cultures would be valuable to have a deeper understanding of how LTM in LLM-driven chatbots should be designed to support public health monitoring in culturally sensitive ways.

7 CONCLUSION

Through comparing call logs from people who engaged with CareCall with and without LTM, as well as interviews with people who used CareCall with LTM, we found that LTM led users to engage in greater health disclosure. LTM fostered positive impressions of the chatbot by offering familiarity, leading users to perceive the system as personal and emotionally supportive. However, LTM introduced some challenges around promoting self-disclosure as it reminded some users of negative feelings about their chronic health issues and raised privacy concerns. Our work points to the influence of selective memory in LLM-driven chatbots, highlighting that the design of LTM warrants careful consideration about what topics need to be remembered in light of the design goals like public health monitoring. We also surface that incorporating LTM in LLM-driven chatbots can help mitigate the impersonality of public health monitoring chatbots, suggesting opportunities for designing more thoughtful follow-ups to demonstrate care. Lastly, we highlight the need to consider the tension between accommodating users' privacy concerns about LTM and public health monitoring goals when designing with LTM.

ACKNOWLEDGMENTS

We thank our participants for their sincere participation. We are grateful to the members of the Generative Chatbot team at NAVER Cloud, especially Sanghwan Bae, Donghyun Kwak, and Soyoung Kang, for helping us understand the technological details and design rationales behind the CareCall LTM feature. We thank Sang-houn Ok for his assistance with recruiting study participants, Hanju Jo for helping out with extracting call logs, and Myeonghan Ryu for his support in writing scripts to count codes. In addition, we thank the members of inComms for their help with transcribing and coding the call logs. This work was supported through a research internship at NAVER AI Lab and by the National Science Foundation under Award IIS-2237389.

REFERENCES

[1] Altman, Irwin. 1973. *Social penetration: The development of interpersonal relationships.* Holt, Rinehart & Winston.

[2] Sanghwan Bae, Donghyun Kwak, Soyoung Kang, Min Young Lee, Sungdong Kim, Yuin Jeong, Hyeri Kim, Sang-Woo Lee, Woomyoung Park, and Nako Sung. 2022. Keep Me Updated! Memory Management in Long-term Conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 3769–3787. https://doi.org/10.18653/v1/2022.findings-emnlp.276

[3] Sanghwan Bae, Donghyun Kwak, Sungdong Kim, Donghoon Ham, Soyoung Kang, Sang-Woo Lee, and Woomyoung Park. 2022. Building a Role Specified Open-Domain Dialogue System Leveraging Large-Scale Language Models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Seattle, United States, 2128–2150. https://doi.org/10.18653/v1/2022.naacl-main.155

[4] Virginia Braun and Victoria Clarke. 2006. Using Thematic Analysis in Psychology. *Qualitative Research in Psychology* 3, 2 (2006), 77–101. https://doi.org/10.1191/1478088706qp063oa

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems (NeurIPS '20)*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 1877–1901. https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

[6] Hye-jin Byun. 2022. NAVER launches AI call service aimed at seniors - The Korea Herald. Retrieved Sep 14, 2023 from https://www.koreaherald.com/view.php?ud=20220530000643

[7] Character AI. 2023. Character AI. Retrieved Sep 14, 2023 from https://character.ai/

[8] Ana Paula Chaves and Marco Aurelio Gerosa. 2021. How Should My Chatbot Interact? A Survey on Social Characteristics in Human–Chatbot Interaction Design. *International Journal of Human–Computer Interaction* 37, 8 (2021), 729–758. https://doi.org/10.1080/10447318.2020.1841438

[9] Baian Chen, Chang Shu, Ehsan Shareghi, Nigel Collier, Karthik Narasimhan, and Shunyu Yao. 2023. FireAct: Toward Language Agent Fine-tuning. arXiv:2310.05915 [cs.CL]

[10] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An Open-Source Chatbot Impressing GPT-4 with

90%* ChatGPT Quality. https://lmsys.org/blog/2023-03-30-vicuna/

[11] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. PaLM: Scaling Language Modeling with Pathways. https://doi.org/10.48550/ARXIV.2204.02311

[12] Shammur Chowdhury. 2017. *Computational Modeling of Turn-Taking Dynamics in Spoken Conversations.* Ph. D. Dissertation. University of Trento. https://doi.org/10.13140/RG.2.2.35753.70240

[13] Patrick Corrigan. 2004. How stigma interferes with mental health care. *American Psychologist* 59, 7 (Oct. 2004), 614–625. https://doi.org/10.1037/0003-066X.59.7.614

[14] Samuel Rhys Cox, Yi-Chieh Lee, and Wei Tsang Ooi. 2023. Comparing How a Chatbot References User Utterances from Previous Chatting Sessions: An Investigation of Users' Privacy Concerns and Perceptions. http://arxiv.org/abs/2308.04879 arXiv:2308.04879 [cs].

[15] Samuel Rhys Cox and Wei Tsang Ooi. 2022. Does Chatbot Language Formality Affect Users' Self-Disclosure?. In *Proceedings of the 4th Conference on Conversational User Interfaces.* ACM, Glasgow United Kingdom, 1–13. https://doi.org/10.1145/3543829.3543831

[16] Magdalena A. Denham and Anthony John Onwuegbuzie. 2013. Beyond Words: Using Nonverbal Communication Data in Research to Enhance Thick Description and Interpretation. *International Journal of Qualitative Methods* 12, 1 (Feb. 2013), 670–696. https://doi.org/10.1177/160940691301200137

[17] Daniëlle Duijst. 2017. Can we Improve the User Experience of Chatbots with Personalisation? (2017). https://doi.org/10.13140/RG.2.2.36112.92165 Publisher: Unpublished.

[18] 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 (June 2017), e19. https://doi.org/10.2196/mental.7785

[19] Google, Inc. 2023. Bard - Chat Based AI Tool from Google, Powered by PaLM 2. Retrieved Sep 14, 2023 from https://bard.google.com/

[20] Mitchell Gordon, Tim Althoff, and Jure Leskovec. 2019. Goal-setting And Achievement In Activity Tracking Apps: A Case Study Of MyFitnessPal. In *The World Wide Web Conference.* ACM, San Francisco CA USA, 571–582. https://doi.org/10.1145/3308558.3313432

[21] Robert B Hays. 1985. A Longitudinal Study of Friendship Development. (April 1985).

[22] Annabell Ho, Jeff Hancock, and Adam S Miner. 2018. Psychological, Relational, and Emotional Effects of Self-Disclosure After Conversations With a Chatbot. *Journal of Communication* 68, 4 (Aug. 2018), 712–733. https://doi.org/10.1093/joc/jqy026

[23] Hofstede Insights. 2023. Country Comparison - Hofstede Insights. https://www.hofstede-insights.com/country-comparison-tool?countries=south+korea

[24] Inflection AI. 2023. Pi, your personal AI. Retrieved Sep 14, 2023 from https://pi.ai/talk

[25] Azra Ismail, Naveena Karusala, and Neha Kumar. 2018. Bridging Disconnected Knowledges for Community Health. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 75 (nov 2018), 27 pages. https://doi.org/10.1145/3274344

[26] Azra Ismail and Neha Kumar. 2018. Engaging solidarity in data collection practices for community health. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018). https://doi.org/10.1145/3274345

[27] Azra Ismail and Neha Kumar. 2021. AI in Global Health: The View from the Front Lines. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 598, 21 pages. https://doi.org/10.1145/3411764.3445130

[28] Azra Ismail, Divy Thakkar, Neha Madhiwalla, and Neha Kumar. 2023. Public Health Calls for/with AI: An Ethnographic Perspective. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW2, Article 354 (oct 2023), 26 pages. https://doi.org/10.1145/3610203

[29] Mohit Jain, Ramachandra Kota, Pratyush Kumar, and Shwetak N. Patel. 2018. Convey: Exploring the Use of a Context View for Chatbots. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems.* ACM, Montreal QC Canada, 1–6. https://doi.org/10.1145/3173574.3174042

[30] Eunkyung Jo, Daniel A Epstein, Hyunhoon Jung, and Young-Ho Kim. 2023. Understanding the Benefits and Challenges of Deploying Conversational AI Leveraging Large Language Models for Public Health Intervention. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems.* ACM, New York, NY, USA, 1–16. https://doi.org/10.1145/3544548.3581503

[31] Dan Jurafsky, Rajesh Ranganath, and Dan McFarland. 2009. Extracting social meaning: identifying interactional style in spoken conversation. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics on - NAACL '09.* Association for Computational Linguistics, Boulder, Colorado, 638. https://doi.org/10.3115/1620754.1620847

[32] Sin-Hwa Kang and Jonathan Gratch. 2010. Virtual humans elicit socially anxious interactants' verbal self-disclosure. *Computer Animation and Virtual Worlds* (2010), n/a–n/a. https://doi.org/10.1002/cav.345

[33] Evangelos Karapanos, Jens Gerken, Jesper Kjeldskov, and Mikael B. Skov (Eds.). 2021. *Advances in Longitudinal HCI Research.* Springer International Publishing, Cham. https://doi.org/10.1007/978-3-030-67322-2

[34] Boseop Kim, HyoungSeok Kim, Sang-Woo Lee, Gichang Lee, Donghyun Kwak, Jeon Dong Hyeon, Sunghyun Park, Sungju Kim, Seonhoon Kim, Dongpil Seo, Heungsub Lee, Minyoung Jeong, Sungjae Lee, Minsub Kim, Suk Hyun Ko, Seokhun Kim, Taeyong Park, Jinuk Kim, Soyoung Kang, Na-Hyeon Ryu, Kang Min Yoo, Minsuk Chang, Soobin Suh, Sookyo In, Jinseong Park, Kyungduk Kim, Hiun Kim, Jisu Jeong, Yong Goo Yeo, Donghoon Ham, Dongju Park, Min Young Lee, Jaewook Kang, Inho Kang, Jung-Woo Ha, Woomyoung Park, and Nako Sung. 2021. What Changes Can Large-scale Language Models Bring? Intensive Study on HyperCLOVA: Billions-scale Korean Generative Pretrained Transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing.* Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 3405–3424. https://doi.org/10.18653/v1/2021.emnlp-main.274

[35] Junhan Kim, Jana Muhic, Lionel Peter Robert, and Sun Young Park. 2022. Designing Chatbots with Black Americans with Chronic Conditions: Overcoming Challenges against COVID-19. In *CHI Conference on Human Factors in Computing Systems.* ACM, New Orleans LA USA, 1–17. https://doi.org/10.1145/3491102.3502116

[36] Predrag Klasnja, Sunny Consolvo, and Wanda Pratt. 2011. How to Evaluate Technologies for Health Behavior Change in HCI Research. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vancouver, BC, Canada) *(CHI '11)*. Association for Computing Machinery, New York, NY, USA, 3063–3072. https://doi.org/10.1145/1978942.1979396

[37] Korea Law Translation Center. 2020. Act on the Prevention and Management of Lonely Deaths. https://elaw.klri.re.kr/eng_mobile/viewer.do?hseq=55028&type=part&key=38

[38] Korean Statistical Information Service. 2023. Resident Population in Five-Year Age Groups. Retrieved Nov 14, 2023 from https://kosis.kr/statHtml/statHtml.do?orgId=101%26tblId=DT_1B04005N%26vw_cd=MT_ETITLE%26list_id=A_7%26scrId=%26language=en%26seqNo=%26lang_mode=en%26obj_var_id=%26itm_id=%26conn_path=MT_ETITLE%26path=%2Feng%2FstatisticsList%2FstatisticsListIndex.do

[39] Hamutal Kreiner and Yossi Levi-Belz. 2019. Self-Disclosure Here and Now: Combining Retrospective Perceived Assessment With Dynamic Behavioral Measures. *Frontiers in Psychology* 10 (March 2019), 558. https://doi.org/10.3389/fpsyg.2019.00558

[40] Josephine Lau, Benjamin Zimmerman, and Florian Schaub. 2018. Alexa, Are You Listening?: Privacy Perceptions, Concerns and Privacy-seeking Behaviors with Smart Speakers. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (Nov. 2018), 1–31. https://doi.org/10.1145/3274371

[41] Gibbeum Lee, Volker Hartmann, Jongho Park, Dimitris Papailiopoulos, and Kangwook Lee. 2023. Prompted LLMs as Chatbot Modules for Long Open-domain Conversation. In *Findings of the Association for Computational Linguistics: ACL 2023.* Association for Computational Linguistics, Toronto, Canada, 4536–4554. https://doi.org/10.18653/v1/2023.findings-acl.277

[42] Minha Lee, Sander Ackermans, Nena van As, Hanwen Chang, Enzo Lucas, and Wijnand IJsselsteijn. 2019. Caring for Vincent: A Chatbot for Self-Compassion. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems.* ACM, Glasgow Scotland Uk, 1–13. https://doi.org/10.1145/3290605.3300932

[43] Yi-Chieh Lee, Naomi Yamashita, and Yun Huang. 2020. Designing a Chatbot as a Mediator for Promoting Deep Self-Disclosure to a Real Mental Health Professional. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW1 (May 2020), 1–27. https://doi.org/10.1145/3392836

[44] Yi-Chieh Lee, Naomi Yamashita, Yun Huang, and Wai Fu. 2020. "I Hear You, I Feel You": Encouraging Deep Self-disclosure through a Chatbot. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems.* ACM, Honolulu HI USA, 1–12. https://doi.org/10.1145/3313831.3376175

[45] Qian Li. 2022. Sensitizing Social Interaction with a Mode-Enhanced Transcribing Process. *Organizational Research Methods* (Oct. 2022), 109442812211340. https://doi.org/10.1177/10944281221134096

[46] Bingjie Liu and S. Shyam Sundar. 2018. Should Machines Express sympathy and empathy? experiments with a health advice chatbot. *Cyberpsychology, Behavior, and Social Networking* 21, 10 (2018), 625–636. https://doi.org/10.1089/cyber.2018.0110

[47] Xi Lu, Eunkyung Jo, Seora Park, Hwajung Hong, Yunan Chen, and Daniel A. Epstein. 2022. Understanding Cultural Influence on Perspectives Around Contact Tracing Strategies. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 468 (nov 2022), 26 pages. https://doi.org/10.1145/3555569

[48] Xi Lu, Tera L. Reynolds, Eunkyung Jo, Hwajung Hong, Xinru Page, Yunan Chen, and Daniel A. Epstein. 2021. Comparing Perspectives Around Human and Technology Support for Contact Tracing. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–15. https://doi.org/10.1145/3411764.3445669

[49] Gale M. Lucas, Jonathan Gratch, Aisha King, and Louis-Philippe Morency. 2014. It's only a computer: Virtual humans increase willingness to disclose. *Computers in Human Behavior* 37 (Aug. 2014), 94–100. https://doi.org/10.1016/j.chb.2014.04.043

[50] Gale M. Lucas, Albert Rizzo, Jonathan Gratch, Stefan Scherer, Giota Stratou, Jill Boberg, and Louis-Philippe Morency. 2017. Reporting Mental Health Symptoms: Breaking Down Barriers to Care with Virtual Human Interviewers. *Frontiers in Robotics and AI* 4 (Oct. 2017), 51. https://doi.org/10.3389/frobt.2017.00051

[51] Zilin Ma, Yiyang Mei, and Zhaoyuan Su. 2023. Understanding the Benefits and Challenges of Using Large Language Model-based Conversational Agents for Mental Well-being Support. arXiv:2307.15810 [cs.CL]

[52] Indrani Medhi Thies, Nandita Menon, Sneha Magapu, Manisha Subramony, and Jacki O'Neill. 2017. How Do You Want Your Chatbot? An Exploratory Wizard-of-Oz Study with Young, Urban Indians. In *Human-Computer Interaction - INTERACT 2017*, Regina Bernhaupt, Girish Dalvi, Anirudha Joshi, Devanuj K. Balkrishan, Jacki O'Neill, and Marco Winckler (Eds.). Vol. 10513. Springer International Publishing, Cham, 441–459. https://doi.org/10.1007/978-3-319-67744-6_28 Series Title: Lecture Notes in Computer Science.

[53] Chinasa T Okolo, Srujana Kamath, Nicola Dell, and Aditya Vashistha. 2021. "It cannot do all of my work": community health worker perceptions of AI-enabled mobile health applications in rural India. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–20.

[54] OpenAI, Inc. 2022. ChatGPT. Retrieved Sep 14, 2023 from https://chat.openai.com

[55] OpenAI, Inc. 2023. Custom instructions for ChatGPT. Retrieved Sep 14, 2023 from https://openai.com/blog/custom-instructions-for-chatgpt

[56] OpenAI, Inc. 2023. Replika. Retrieved Nov 14, 2023 from https://replika.com/

[57] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. https://doi.org/10.48550/ARXIV.2203.02155

[58] Hyanghee Park and Joonhwan Lee. 2021. Designing a Conversational Agent for Sexual Assault Survivors: Defining Burden of Self-Disclosure and Envisioning Survivor-Centered Solutions. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–17. https://doi.org/10.1145/3411764.3445133

[59] 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 (April 2019), e12231. https://doi.org/10.2196/12231

[60] SoHyun Park, Anja Thieme, Jeongyun Han, Sungwoo Lee, Wonjong Rhee, and Bongwon Suh. 2021. "I wrote as if I were telling a story to someone I knew.": Designing Chatbot Interactions for Expressive Writing in Mental Health. In *Designing Interactive Systems Conference 2021*. ACM, Virtual Event USA, 926–941. https://doi.org/10.1145/3461778.3462143

[61] Rachel Phinnemore, Mohi Reza, Blaine Lewis, Karthik Mahadevan, Bryan Wang, Michelle Annett, and Daniel Wigdor. 2023. Creepy Assistant: Development and Validation of a Scale to Measure the Perceived Creepiness of Voice Assistants. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–18. https://doi.org/10.1145/3544548.3581346

[62] Matthew D. Pickard, Catherine A. Roster, and Yixing Chen. 2016. Revealing sensitive information in personal interviews: Is self-disclosure easier with humans or avatars and under what conditions? *Computers in Human Behavior* 65 (Dec. 2016), 23–30. https://doi.org/10.1016/j.chb.2016.08.004

[63] Manuel Portela and Carlos Granell-Canut. 2017. A new friend in our Smartphone? Observing Interactions with Chatbots in the search of emotional engagement. (2017).

[64] Gonçalo Raposo, Luisa Coheur, and Bruno Martins. 2023. Prompting, Retrieval, Training: An exploration of different approaches for task-oriented dialogue generation. In *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, Svetlana Stoyanchev, Shafiq Joty, David Schlangen, Ondrej Dusek, Casey Kennington, and Malihe Alikhani (Eds.). Association for Computational Linguistics, Prague, Czechia, 400–412. https://doi.org/10.18653/v1/2023.sigdial-1.37

[65] Abhilasha Ravichander and Alan W. Black. 2018. An Empirical Study of Self-Disclosure in Spoken Dialogue Systems. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*. Association for Computational Linguistics, Melbourne, Australia, 253–263. https://doi.org/10.18653/v1/W18-5030

[66] Ali Shameli, Tim Althoff, Amin Saberi, and Jure Leskovec. 2017. How Gamification Affects Physical Activity: Large-scale Analysis of Walking Challenges in a Mobile Application. In *Proceedings of the 26th International Conference on World Wide Web Companion - WWW '17 Companion*. ACM Press, Perth, Australia, 455–463. https://doi.org/10.1145/3041021.3054172

[67] Selena Simmons-Duffin. 2020. As States Reopen, Do They Have The Workforce They Need To Stop Coronavirus Outbreaks? *NPR* (Jun 2020). https://www.npr.org/sections/health-shots/2020/06/18/879787448/as-states-reopen-do-they-have-the-workforce-they-need-to-stop-coronavirus-outbre

[68] Gabriel Skantze. 2021. Turn-taking in Conversational Systems and Human-Robot Interaction: A Review. *Computer Speech & Language* 67 (May 2021), 101178. https://doi.org/10.1016/j.csl.2020.101178

[69] Mark S. Smolinski, Adam W. Crawley, Kristin Baltrusaitis, Rumi Chunara, Jennifer M. Olsen, Oktawia Wójcik, Mauricio Santillana, Andre Nguyen, and John S. Brownstein. 2015. Flu Near You: Crowdsourced Symptom Reporting Spanning 2 Influenza Seasons. *American Journal of Public Health* 105, 10 (Oct. 2015), 2124–2130. https://doi.org/10.2105/AJPH.2015.302696

[70] S. Shyam Sundar and Jinyoung Kim. 2019. Machine Heuristic: When We Trust Computers More than Humans with Our Personal Information. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–9. https://doi.org/10.1145/3290605.3300768

[71] Alice Thudt, Dominikus Baur, Samuel Huron, and Sheelagh Carpendale. 2016. Visual Mementos: Reflecting Memories with Personal Data. *IEEE Transactions on Visualization and Computer Graphics* 22, 1 (Jan. 2016), 369–378. https://doi.org/10.1109/TVCG.2015.2467831

[72] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. arXiv:2302.13971 [cs.CL]

[73] David L. Vogel and Stephen R. Wester. 2003. To seek help or not to seek help: The risks of self-disclosure. *Journal of Counseling Psychology* 50, 3 (July 2003), 351–361. https://doi.org/10.1037/0022-0167.50.3.351

[74] Lu Wang, Munif Ishad Mujib, Jake Williams, George Demiris, and Jina Huh-Yoo. 2021. An Evaluation of Generative Pre-Training Model-based Therapy Chatbot for Caregivers. https://doi.org/10.48550/ARXIV.2107.13115

[75] Weizhi Wang, Li Dong, Hao Cheng, Xiaodong Liu, Xifeng Yan, Jianfeng Gao, and Furu Wei. 2023. Augmenting Language Models with Long-Term Memory. arXiv:2306.07174 [cs.CL]

[76] Jing Wei, Sungdong Kim, Hyunhoon Jung, and Young-Ho Kim. 2024. Leveraging Large Language Models to Power Chatbots for Collecting User Self-Reported Data. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW1, Article 87 (apr 2024), 35 pages. https://doi.org/10.1145/3637364

[77] Jing Xu, Arthur Szlam, and Jason Weston. 2021. Beyond Goldfish Memory: Long-Term Open-Domain Conversation. http://arxiv.org/abs/2107.07567 arXiv:2107.07567 [cs].

[78] Xinchao Xu, Zhibin Gou, Wenquan Wu, Zheng-Yu Niu, Hua Wu, Haifeng Wang, and Shihang Wang. 2022. Long Time No See! Open-Domain Conversation with Long-Term Persona Memory. In *Findings of the Association for Computational Linguistics: ACL 2022*. Association for Computational Linguistics, Dublin, Ireland, 2639–2650. https://doi.org/10.18653/v1/2022.findings-acl.207

[79] Deepika Yadav, Prerna Malik, Kirti Dabas, and Pushpendra Singh. 2019. Feedpal: Understanding Opportunities for Chatbots in Breastfeeding Education of Women in India. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 170 (nov 2019), 30 pages. https://doi.org/10.1145/3359272

[80] Yonhap. 2017. Lonely deaths of middle-aged, youth brackets stand out amid single-person households - The Korea Herald. Retrieved Sep 14, 2023 from https://www.koreaherald.com/view.php?ud=20171207000623

[81] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A Survey of Large Language Models. arXiv:2303.18223 [cs.CL]

[82] Zhonghua Zheng, Lizi Liao, Yang Deng, and Liqiang Nie. 2023. Building Emotional Support Chatbots in the Era of LLMs. arXiv:2308.11584 [cs.CL]

[83] Wanjun Zhong, Lianghong Guo, Qiqi Gao, and Yanlin Wang. 2023. MemoryBank: Enhancing Large Language Models with Long-Term Memory. *arXiv preprint arXiv:2305.10250* (2023).

A DETAILED DESCRIPTION OF LONG-TERM MEMORY IN CARECALL

In this section, we provide a more detailed description of the design process, design rationales, technical implementation, and evaluation of LTM in CareCall. The selection of LTM topics in CareCall involved the following process. First, a set of call logs with 100 users (721 sessions) was classified using Positive-Neutral-Negative labels, designed to assess user satisfaction with conversational agents [12]. Positive labels were assigned when users laughed during conversations, expressed gratitude, or provided detailed responses more than three times in a session. Negative labels were given when users showed anger or frustration or disregarded the agent (e.g., not answering the agent's questions, answering open-ended questions with simple "yes" or "no", and terminating calls without explanation). Neutral labels were applied when neither Positive nor Negative characteristics were observed. Through the qualitative analysis of call logs labeled with "Positive", the CareCall developers found that users most frequently spoke about health, meals, and sleep in significant detail. The analysis also showed that users enjoyed having conversations about their family, pets, and visited places although these topics surfaced less frequently. Further, the analysis surfaced that users frequently showed frustration when CareCall did not remember mentions about their health status (*e.g.*, experienced back pain, went through surgery) from previous conversations. Second, a set of logs encompassing 102 sessions was collected from check-up calls between public health officers and users and analyzed after obtaining informed consent from both parties. The analysis of these human-human call logs similarly showed that public health officers often remembered and followed up on health-related topics, such as the status of individuals' specific health conditions and whether they were regularly seeing doctors or taking prescribed medications. Third, public health officers who had been involved in the deployment of CareCall gave feedback that whether individuals went out or stayed at home all day was a critical factor in preventing lonely deaths. Combining these insights, the CareCall developers incorporated five topics in the LTM feature, including Health, Meals, Sleep, Visited Places, and Pets. In the initial phase, family (e.g., whether they have family members or details about family members) was also included as part of the LTM topics, but the developers decided to exclude it later because of the sensitivity of the topic, particularly for individuals living alone. Hobbies and interests were also part of the LTM topics in the initial phase, but the developers ended up excluding them because they observed several instances where the target users did not like getting such questions. (*e.g.*, 'I live from hand to mouth. I don't have the luxury to have hobbies.')

Another key question in the design of CareCall's LTM feature was determining what level of detail the system needs to remember about a person. The extent to which conversational partners can share and remember details about each other often depends on the closeness of the relationship [21]. For instance, one might feel awkward if close friends who have known each other for a long time could not recall essential background information, such as whether they have a child. Conversely, one might feel uncomfortable if acquaintances knew overly specific information, like the exact name of the daycare center that their grandson goes to. Therefore, it is essential to maintain an appropriate level of detail about a person, in line with the closeness of the relationship, during conversations. To ensure a suitable level of closeness in the context of public health intervention, CareCall's LTM feature was designed to mimic a casual friend who can share general information about an individual's background and interests, creating a comfortable and appropriate conversational environment.

With the five topics and the level of closeness in mind, the example dialog corpus was created through the data augmentation process, which is described in [2]. Trained workers collaboratively worked with the LLM, flagging and fixing errors in both chatbots' responses as well as summaries for each session that the LLM subsequently utilized. The process of creating the example dialog corpus adhered to a specific priority sequence: (1) Health, (2) Meals and Sleep, and (3) Visited Places and Pets. In the final version of the example dialog corpus, the distribution of topics being stored in summaries was carefully considered, resulting in the following proportions within five consecutive sessions: Health (72.6%), Meals (75.7%), Sleep (52.7%), Visited Places (28.7%), and Pets (5.2%).

A key aspect of CareCall's LTM feature was to keep the memory up-to-date when information about a person changes over time. Updating the memory with new relevant information is particularly important in the context of CareCall, given the need for monitoring peoples' constantly changing health statuses for public health intervention. CareCall's LTM feature leveraged a dynamic memory management mechanism[2] that finds and eliminates the existing information that contradicts new information from memory to ensure that the chatbot keeps track of and brings up the latest information about users while conversing through multiple sessions. For example, if a person mentions that they have the flu, CareCall stores the information in its memory and might ask about it later. But then, if they later mention that they recovered, the system deletes the information and does not ask about it in the future. In addition, the system updates their memory if users correct certain information that CareCall remembered about themselves wrong. For example, if a user says that their dog's name is Coco, not Leo, the system updates the information.

CareCall's LTM feature was formally evaluated both automatically and by humans. The automatic evaluation of the LTM feature suggests that CareCall with LTM achieved better performance in terms of perplexity (how well a language model understands and predicts the next words in a sentence), F1 Score (how well a model performs in classification tasks), and Distinct-1/2 (diversity of generated text) compared to the one without LTM, particularly as the sessions progressed (please see Bae *et al.* [2] for more details about the experiment results). Human evaluations showed that CareCall with LTM was perceived to have a better ability to remember, humanness, and engagingness compared to the one without LTM [2].

[2] Bae *et al.* [2] provides a more detailed description of the novel memory management mechanism

B FINAL CODEBOOK FOR QUANTITATIVE ANALYSIS OF CALL LOGS

Category	Code	Code description	Example
Meals	*Meals-simple*	Whether a person has eaten, or whether they have an appetite	"Yeah, I just had lunch." "I don't really feel like eating."
	Meals-detail	What kind of food a person had, or why they have difficulty eating	"I had Kimchi Stew for lunch." "I've not been feeling well, so I lost my appetite."
Sleep	*Sleep-simple*	Whether a person sleeps well	"Yeah, I'm having a good sleep."
	Sleep-detail	Details on sleep quality, including why a person has difficulty sleeping	"I have a hard time falling asleep." "I often wake up in the middle of the night." "I couldn't sleep well because of the back pain."
Health	*Health-simple*	Whether a person is experiencing any health issues that likely warrant clinical care/attention	"I was a bit under the weather yesterday, but now I'm doing better." "I'm feeling well."
	Health-detail	Details about one's health issues, including physical/mental discomforts, pain, or clinical symptoms.	"I have a toothache." "I have leg pain, so I can't walk around much." "I have been terribly ill over the last few days." "My diabetes got worse"
	Health-attitude	Thoughts, feelings, and attitudes toward their health and clinical treatment	"I make regular visits to my doctor but it isn't really working." "I think I'll just have to live with my leg pain." "I'm too old to get better."
Clinical	*Clinical-simple*	Whether a person is currently seeking clinical care or taking medications regularly, and whether they recently saw a doctor	"Yeah, I've been seeing a doctor." "(Are you taking your pills regularly?) Yes."
	Clinical-detail	Details about clinical care being sought (*e.g.*, treatment, medical specialists)	"I'm taking pills for high blood pressure." "I have a slipped disc." "I just went to see a dentist." "I got pain relief shots on my back."
Activity	*Activity-simple*	Whether a person engages in any social or physical activities	"I'm just staying home today." "I'm just chilling at home." "I'm just watching TV."
	Social-detail	Details about one's social activities, including people, place, and jobs	"I just came back from the community center." "I'm at my daughter's place." "I'm working as a babysitter."
	Physical-detail	Details about one's physical activities, including chores and exercises, reasons if not engaged in any	"I'm doing some laundry now." "I'm taking some aerobic classes." "I can't do any exercise 'cause I have back pain." "I'm not doing any exercise. Just in my wheelchair all day."
Wellbeing	*Wellbeing-simple*	How a person is doing in general	"I'm doing okay." "I'm a bit tired today."
	Wellbeing-detail	Details about one's general wellbeing in both physical and mental aspects	"I usually feel a bit tired in the morning." "I think I need some rest today." "Life is not enjoyable living alone. Don't have anyone around."
Hobbies	*Hobbies*	Descriptions of one's interests, preferences, and hobbies	"I like to drink coffee with sugar." "I like to read detective fiction." "I'm transcribing the Bible."
Positive Reactions	*Appreciation*	A user thanks the agent explicitly. Conventional expressions of thanks in farewell are not included.	"It was really nice of you to say that." "Thanks for looking out for me." "Thank you for checking in with me."
	Anthromorphization	A user interacts with the agent as if it were a human being.	"I'm doing well. How are you doing?" "By the way, what's your name?" "You have a great day." "I'll cook noodles for you one day."
Negative Reactions	*Negative Feedback*	A user expresses frustration to the agent.	"You're not a human. You're what, an answering machine?" "Stop playing the recording." "It's definitely not a human voice. You're just playing a recorded message, right?"
	Disregard	A user does not respond to the agent's question, or gives yes/no answers to open questions, or hangs up the call without explanation.	"(What did you have for lunch?) Yeah." "(Do you have any plans for today?)" [hung up]