Understanding the Benefits and Challenges of Deploying Conversational AI Leveraging Large Language Models for Public Health Intervention

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

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

Hyunhoon Jung
NAVER CLOUD
Seongnam, Gyeonggi
Republic of Korea
hyunhoon.j@navercorp.com

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

ABSTRACT

Recent large language models (LLMs) have advanced the quality of open-ended conversations with chatbots. Although LLM-driven chatbots have the potential to support public health interventions by monitoring populations at scale through empathetic interactions, their use in real-world settings is underexplored. We thus examine the case of CareCall, an open-domain chatbot that aims to support socially isolated individuals via check-up phone calls and monitoring by teleoperators. Through focus group observations and interviews with 34 people from three stakeholder groups, including the users, the teleoperators, and the developers, we found CareCall offered a holistic understanding of each individual while offloading the public health workload and helped mitigate loneliness and emotional burdens. However, our findings highlight that traits of LLM-driven chatbots led to challenges in supporting public and personal health needs. We discuss considerations of designing and deploying LLM-driven chatbots for public health intervention, including tensions among stakeholders around system expectations.

CCS CONCEPTS

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

KEYWORDS

Chatbot, Large language model, Open-domain dialog system, Public health, Check-up calls, Social isolation

ACM Reference Format:
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 (CHI '23), April 23–28, 2023, Hamburg, Germany*. ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3544548.3581503

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

CHI '23, April 23–28, 2023, Hamburg, Germany

ACM ISBN 978-1-4503-9421-5/23/04.
https://doi.org/10.1145/3544548.3581503

1 INTRODUCTION

Technology has increasingly been used to help monitor populations for public health understanding and intervention. In the HCI and CSCW communities, a range of systems, including chatbots [27, 81] and mobile apps [44, 45] have been proposed and examined to support public health monitoring and intervention at scale. Prior work suggests that such systems can help offload parts of the labor of public health workers by automating some aspects of care, such as answering frequent questions and identifying public resources [4, 69, 81], allowing them to focus more on care-driven tasks like monitoring the wellbeing of individuals [27].

Advances in artificial intelligence (AI) and natural language processing (NLP) technologies open up a promising avenue for supporting population-level health interventions. In particular, chatbots have been proposed as effective tools for scaling abilities to provide informational and emotional support around health [42, 76]. Traditional chatbots rely on *task-oriented* flows, which use conversational rules to respond to specific prompts, such as answering questions. However, recent advances in large language models (referred to as LLMs hereinafter) have brought breakthroughs of *open-domain* dialog systems, which perform free-form conversations in open-ended topics with an overarching goal of providing empathy (*e.g.*, [24, 75, 86]) [23]. Such systems can be beneficial for public health interventions in providing empathetic interactions for populations going through difficult health experiences [44] and reaching out to broader populations who have been underserved. However, few studies have explored how LLM-based chatbots can be leveraged in population-level health interventions in real-world settings, limiting understanding of the benefits and drawbacks of free-form conversations towards addressing public health needs.

To understand the benefits and challenges of deploying conversational AI leveraging LLMs for public health, we explore the case of **CLOVA CareCall** (*c.f.*, [10]; referred to as CareCall hereinafter for brevity), a conversational AI that aims to help support socially isolated individuals via check-up phone calls as a public health intervention. As an open-domain chatbot, CareCall both collects data about the individuals' general health and serves as a conversational partner to mitigate their loneliness by generating human-like questions and answers on the fly. As of May 2022, CareCall had been deployed to 20 municipalities in South Korea for between 2 and 12 months, with the aim of monitoring socially isolated individuals, including middle-aged and older adults living alone. Being a rare example of an LLM-driven chatbot deployed in a real-world setting in

public health contexts, CareCall is a useful case for understanding the role of LLM-driven chatbots in public health intervention.

We observed focus group workshops with 14 CareCall users and conducted interviews with 20 people from three groups of the main stakeholders around the CareCall system, including five users, five teleoperators who monitored the users' conversation logs, and 10 developers who designed and implemented the system as well as communicated with local governments. In total, we report on insights from 34 people who interacted with different aspects of CareCall. From the study, we identified the benefits and challenges in leveraging CareCall in public health interventions. The teleoperators valued that the LLM-driven chatbot helped them gain a holistic understanding of each individual through open-ended conversations while offloading their workload. The users perceived that the open-ended nature of the dialog helped mitigate loneliness by asking caring questions about their health and covering conversation topics beyond health, such as asking about hobbies and interests. However, stakeholders often had different needs around LLM-driven chatbots towards their goals and different expectations of their capabilities. While the municipal authorities desired to incorporate specific health questions and customize conversations to different target groups, the developers faced challenges in accommodating those needs due to the uncertainty in control and the resource-intensive nature of customizing LLM-based chatbots. In addition, the open-ended nature of conversations led the users to expect the system to be able to support social services out of its scope, placing an additional burden on teleoperators. Further, the users felt that the system was impersonal because it lacked follow-ups on past conversations around personal health, as LLM-driven chatbots struggle to incorporate long-term memory, which led to challenges in providing emotional support. Based on the findings, we discuss opportunities for improving LLM-driven chatbots to provide greater emotional support. We also suggest the need for designing resources and processes that help different stakeholders negotiate the tradeoffs between open-domain and task-oriented chatbots. Lastly, we discuss the need and challenges in scaling LLM-driven chatbots to support diverse public health needs.

The key contributions of this work are twofold:

- Understanding of the benefits and challenges in leveraging LLM-driven chatbots in public health interventions through interviews and focus group observations with 34 people who engaged with, managed, and developed CareCall. While CareCall offered emotional benefits, particularly around supporting broader conversation topics, it also had challenges in providing emotional support due to its limited personalization and lack of long-term memory. We also observed tensions around the open-ended nature of LLM-driven chatbots, which made it challenging for the developers to manage expectations around the emergency and social service needs of the users. Municipal authorities further wished to integrate specific health monitoring questions or customize to different target groups, which were hard to meet due to inherent characteristics of LLM-driven chatbots.
- Implications for further research and implementation of chatbots for public health interventions, particularly around (1) improving emotional support through implementing a long-term memory in public health chatbots, (2) designing resources and processes that help communicate the respective strengths and weaknesses of task-oriented and open-domain chatbots to help multiple stakeholders in public health contexts negotiate those tradeoffs, and (3) designing mechanisms to help target populations or care professionals contribute to dialog datasets to scale chatbots to diverse public health needs.

2 RELATED WORK

In this section, we first review the HCI literature on public health work and caregiving technology for individuals living alone. We then examine prior work on LLMs and open-domain dialog systems.

2.1 HCI in Public Health Work

The HCI community has offered insights into the use of technology by different stakeholders involved in public health work, including government officers, community health workers, and care recipients. One major line of the research on technology interventions in public health settings has focused on automating aspects of care that public health workers typically have to provide manually, such as answering common questions [81] and identifying public resources [4, 69]. For example, Pendse *et al.* [56] highlighted that institutional limitations often interfere with providing support through helpline systems, suggesting that automating some aspects of these systems could help care recipients better navigate the barriers. Relevant to our work, technology is often used to automate the collection of personal health information from care recipients, to reduce the burden of public health authorities in monitoring people at scale. For example, Ismail and Kumar found that health workers often perceive collecting such data to be mundane and redundant, and technology offloading that burden could enable workers to focus on more care-driven tasks [27]. A range of systems, including chatbots [27, 81] and mobile apps [44, 45], have been proposed and examined to support care recipients in self-reporting aspects of their health and well-being to public and community health infrastructures. Beyond logistical advantages, a benefit of these automated approaches is that care recipients may feel more comfortable disclosing sensitive information to a digital system rather than a human, such as a positive test result [44, 45, 81]. However, a core concern is that these systems may not be as empathetic or unable to provide emotional support to people going through difficult health experiences in the same way direct communication with a human would [44, 45]. Researchers reiterate that these systems should thus not fully replace public health workers in collection roles but aim to be complementary support [60, 81].

Although the introduction of technology can reduce the burdens of aspects of public health work, those experiences may be uneven across stakeholders. For example, in reflecting on years of deploying FeedFinder, Simpson *et al.* highlighted the uncompensated maintenance and communication labor the service required, despite it being beneficial for care recipients [69]. Further, research often does not capture the attitudes of the people on the front lines of using these technologies, such as community health workers, to understand the technology's benefits and tradeoffs [26]. In studying CareCall, we thus gathered perspectives from as many stakeholders as possible to offer a holistic understanding of the system's use.

2.2 Caregiving Technology for Individuals Living Alone

Individuals living alone tend to be vulnerable to various health concerns, particularly with aging [52]. There is a greater risk of social isolation and loneliness when living alone, which is closely linked to negative health outcomes such as dementia, depression, heart disease, and stroke [18]. In addition, a lack of social contacts limits one's ability to receive help in emergency situations [33]. Research on caregiving technologies has aimed to support these individuals (*e.g.*, [15, 37, 51, 62, 64, 74]). One subset of these systems is often referred to as *telecare* systems, which seek to mediate care among individuals living alone, formal and informal caregivers, and emergency services [37, 62]. Another subset of caregiving technologies—including CareNet [15], Digital Family Portraits [51, 64], and SHel [74]—have aimed to support family members or other care network members in maintaining awareness of the older adults' daily activities through environmental sensors and ambient displays [51, 64, 74]. Field studies have suggested that such systems can alleviate the loneliness of individuals living alone and provide peace of mind for their informal caregivers [15, 64].

A core concern is that existing technologies have predominantly targeted individuals who have readily accessible social contacts, such as informal caregivers [15, 51, 64, 74]. However, studies have pointed out that compared to high socioeconomic status (SES) individuals, low-SES individuals living alone tend to have fewer social contacts that they can reach out to in emergency situations [1, 78], reflecting important differences in how to approach designing technology to support this more vulnerable population [70]. Thus, many of the existing technologies might not fit the lived realities of individuals living alone who have fewer social contacts. Veinot *et al.* [73] argue for the need to study and design population-level interventions, which may be delivered by public health officers [73]. While such at-scale interventions could provide necessary help for vulnerable populations such as low-SES individuals living alone, a key challenge is the immense public resources required for operating such interventions at scale.

New advances in AI opened up new opportunities to facilitate at-scale health interventions for vulnerable populations by automating some aspects of care, such as regularly collecting health information from individuals. Not only can AI-driven technology alleviate public health workers' burden on delivering interventions, but its scalability can also help reach out to broader populations who have been underserved. However, relatively few studies have explored how AI-driven systems can be leveraged in health interventions for vulnerable populations. Motivated by this gap, we explore the benefits and challenges of deploying AI-driven check-up calls with low-SES individuals living alone.

2.3 Large Language Models

The area of NLP has shown remarkable achievements with the advance in language models. Language models aim to generate coherent follow-up text to inputs, trained on human-generated textual data (*e.g.*, a corpus) such as Wikipedia contents or social media posts [9, 43]. With the underlying knowledge about the probabilistic relationship among adjacent words in the language corpus, the *pre-trained* models can be retargeted to more specific NLP tasks—such as machine translation (*e.g.*, [82]), sentiment classification (*e.g.*, [50]), and question answering (*e.g.*, [59])—through *fine-tuning* with task-specific datasets [9, 43].

While the early language models with millions of parameters (*e.g.*, BERT [16]) required additional fine-tuning steps to perform a specific task, recent *large* language models (*e.g.*, GPT-3 [9], HyperCLOVA [29], PaLM [12], OPT [85]) with a larger number of parameters (*e.g.*, 13–175B for GPT-3, 82B for HyperCLOVA), have enabled a new paradigm of *in-context learning* [9, 43]. In in-context learning, models understand input text written in human language, which is called *a prompt*, and generate the following text that coherently follows the prompt. For example, if given a prompt like '`Classify the food into categories. Apple→Fruit; Onion→Vegetable; Milk→`' as an input, an LLM is likely to infer the following text, '`Dairy`.' While the nature of the task is still the text continuation, the model understands the latent concept of food classification in the input prompt. In the similar vein, prompts can be composed in a variety of ways to transform LLMs to solve diverse problems. Motivated by such capability of LLMs, NLP and HCI researchers have leveraged LLMs in various problem spaces, including but not limited to creative writing (*e.g.*, [13, 39]), information extraction (*e.g.*, [32, 54]), and writing programming code (*e.g.*, [11]). Among many application domains, our work focuses on the open-domain dialog systems driven by LLMs.

2.4 Supporting Open-Ended Conversations with Large Language Models

Designing AIs that converse with humans coherently and engagingly has been an active research topic in the areas of NLP, Machine Learning, and HCI. Depending on the goal of the interaction, conversational AIs are usually designed as either *task-oriented* or *open-domain* dialog systems [23]. Task-oriented dialog systems are designed for a specific goal (*e.g.*, booking a flight ticket) with pre-defined information schema (*e.g.*, slots to fill such as destination, date, and preferred airlines). Within the HCI community, task-oriented dialog systems have recently been proposed with the goal of promoting mental health. Specifically, studies have designed chatbots for eliciting self-disclosure [40, 41, 55] or increasing self-compassion by taking care of chatbots that experience distress [30, 38]. Relevant to our work, Yeonheebot performs conversations with older adults to mitigate their depression and anxiety [65]. However, as rule-based or hybrid (*e.g.*, combining rules and intent-based response retrieval) chatbots with pre-defined conversation flows, prior systems were limited in supporting serendipitous topics that users might bring up during conversations [38]. Conversely, open-domain dialog systems are intended to perform free-form conversations in open-ended topics ranging from daily life (*e.g.*, [84]) to movies (*e.g.*, [49]), with an overarching goal of providing empathy and enhancing feelings of social belonging (*e.g.*, [24, 75, 86]) [23].

Research has often discussed that designing quality open-domain dialog systems is more challenging than designing task-oriented dialog systems [20, 23]. Technically, it is relatively straightforward to define the 'quality' of the task-oriented dialogs because there exist clear user goals and information slots that the agent should ask the user about [19, 23]. Conversely, guidelines for open-domain

dialog systems are less fixed. Huang *et al.* suggest that open-dialog systems should aim to (1) understand the *semantics* of what the user said, (2) behave *consistently* with their predefined persona, conversation history, and speaking style, and (3) interact with the user *emotionally* [23]. However, these multidimensional goals made it hard to define an objective quality metric for a chatbot's responses. State-of-the-art neural network models have not satisfied these goals simultaneously due to the complexity of multi-turn reasoning of the conversational context and infeasibility of automated evaluations to improve model quality [23].

Recent LLMs, however, have brought breakthroughs in open-domain dialog systems thanks to their capabilities in generating coherent and contextual responses through in-context learning [3, 63]. LLM-based chatbots[1] receive the current dialog history (i.e., list of turns of the user and the agent) in a prompt and infer the agent's following response accordingly [63]. The in-context learning inherently covers the multi-turn reasoning of the conversational context, generating responses that are generally aware of and specific to the context. Since research on LLM-driven chatbots is still sparse and in the early stage, there still exist limitations and challenges in designing LLM-driven chatbots, which are mainly resulting from the inherent characteristics of LLMs. As language models generate the most probable output based on a complex structure of neural networks (called *transformers* [72]), it is not explainable how an LLM *'reads'* the input prompts written in natural language [43]. In the context of chatbots, it is therefore challenging to anticipate how an LLM would process the history of dialog and what response it would generate. Since LLMs have learned a tremendous amount of human-generated text, there is always a risk that the conversation flow might follow directions unintended or unaccounted for by the chatbot designer [3]. For example, from a study with a mental therapy chatbot built with GPT-2, Wang *et al.* found that the chatbot was likely to provide more negative comments than the human therapists would [75]. Also, there exists a possibility that the unethical or biased phrases ingrained in the models' pre-training datasets might be exposed in the model's output, causing the chatbot to say socially biased [6, 7, 21, 67, 68] or toxic [22] messages. One known method to steer the conversations to converge towards desired scenarios is to put ideal conversation examples in the prompt together [3]. Although such an in-context learning approach helps steer the model output, it is still challenging to perfectly control the model to say or not to say specific phrases [3, 75].

Given the aforementioned challenges and risks of leveraging LLMs for open-ended chatbots, CareCall presents a unique example of an LLM-based open-ended chatbot being deployed in a real-world setting as a public health intervention. By identifying the benefits and challenges from focus group observations and interviews with users, teleoperators, and developers who engaged with different aspects of CareCall, we extend the line of health and AI research for care work and public health interventions.

3 STUDY CONTEXT: CLOVA CARECALL

In this section, we cover background information about CareCall as an example of LLM-driven chatbot deployed as a public health intervention. This background is based on what we learned from interviews with the CareCall developers and the literature on the underlying technology (*c.f.*, [3, 29]). Our contribution treats CareCall as a case study for considering the utility and limitations of LLM-based chatbots for public health, building on these prior studies that contribute the novel implementation of CareCall.

3.1 Motivation and Deployment of CareCall

CareCall is a conversational AI system designed for socially isolated individuals in South Korea [10]. Motivated by the recent Act on the Prevention and Management of Lonely Death in South Korea [34], CareCall is aimed at providing individuals with emotional support and regularly checking their health status.

Figure 1 describes a brief overview of the system architecture and the interaction between the two stakeholders of CareCall. The CareCall chatbot (Ⓐ in Figure 1) regularly (e.g., once or twice a week) calls the users and leads an open-ended conversation about daily life for about 2–3 minutes, in a female voice. After each call, the dashboard (Ⓑ in Figure 1) automatically extracts (1) five health metrics, including meals, sleep, general health, going out, and exercise, as one of three statuses (Positive/Negative/Unknown), and (2) emergency alerts (e.g., dizziness, chest pain, high fever, difficulty in breathing) from the dialogs using user state detection classifiers. The summary of each user's status is displayed on a web dashboard for social workers. On the dashboard, social workers can access the call recordings as well as the five health metrics and emergency alerts of the individuals whom they are in charge of.

CareCall first started to roll out in Haeundae-gu in Busan in November 2021 [10]. As of May 2022, CareCall was being deployed to 20 out of 226 municipalities in South Korea as a pilot project with the intent to scale up in the future. In this study, we specifically focused on Seoul where CareCall was deployed to 301 individuals from June 2022 to August 2022 as part of the pilot project. Each municipality's government had slightly different criteria for the target users (*i.e.*, people who receive the calls) in terms of the age group or chronic health conditions, though sharing the overarching characteristic of social isolation. CareCall was deployed to older adults living alone in most of the municipalities, but in a few cases, it was deployed to middle-aged adults, individuals with early dementia, or healthy older adults. In Seoul, where our study is focused, CareCall was deployed to middle-aged (40s to 60s) adults who were living alone and were predominantly (87%) recipients of the National Basic Livelihood Security (below 50% of median household income). The deployment with such a population was motivated by the highest proportion of solitary deaths among all age groups in Seoul [83]. The CareCall pilot project participants in Seoul were recommended by public officers who were providing social care services to these individuals. Most of the CareCall project participants in Seoul were receiving regular check-up calls from different types of public officers, including social welfare officers, public health officers, and emergency response officers. Introduction of CareCall did not replace their existing check-up calls from humans but rather increased the frequency of check-up calls, partially due to the short-term nature of the pilot project. The pilot deployment of CareCall across all municipalities obtained participants' informed

[1]Throughout the paper, we use the term *chatbot* as synonymous with *conversational AI* or *dialog system* for brevity.

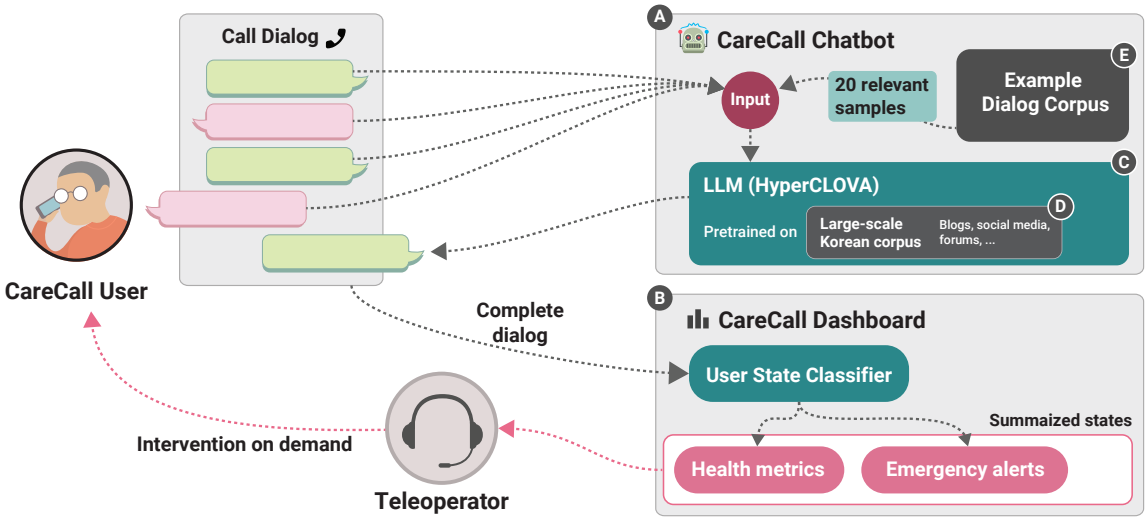

Figure 1: System architecture of CareCall, describing Ⓐ a chatbot conversing with users and Ⓑ a dashboard for teleoperators.

consent prior to their voluntary enrollment, which included collecting health information through conversations with the AI system. Note that the scope of our study was conducting interviews and observations of different stakeholders related to the CareCall pilot project; thus, the development and the pilot deployment of CareCall were outside our scope of the study.

Each municipality's government handled the teleoperating tasks of CareCall in different ways. For example, some governments had their social welfare officers in charge of the teleoperating, as an aspect of their social care work, while others hired part-time workers for the teleoperating tasks specifically for the CareCall pilot project. The government of Seoul hired 14 part-time social workers for the teleoperating tasks for the CareCall pilot project through a social enterprise that employs retired individuals over the age of 55 (referred to as *teleoperators* hereinafter for brevity). In Seoul, the teleoperators' protocols required them to monitor the call recordings for negative health signals (e.g., skipping meals, poor sleep) or emergency alerts on the dashboard. If they found any health issues from the call recordings, they were asked to share with their team and reach out to the person to check if everything is okay. If they noticed anything noteworthy from the manual check-up calls, they were asked to write a report to escalate to those who provide social care services in their municipalities alongside the deployment. Other municipalities used similar protocols for the teleoperating tasks of CareCall, though public officers' workflows slightly differed because they were often in a position to directly connect to social services or healthcare services.

3.2 Design of CareCall Chatbot

The CareCall chatbot was designed as an open-ended dialog system powered by an LLM (Ⓒ in Figure 1) called HyperCLOVA [29] which has 82B parameters trained on a Korean corpus of 561.8B tokens (Ⓓ in Figure 1). The training corpus includes blog posts, online forums, news articles, comments, and online Q&As [29]. At each conversation turn, the chatbot generates a response by putting 20 relevant example dialogs along with the current dialog history to the LLM. These example dialogs are sampled on the fly from a large-scale dialog corpus[2] (Ⓔ in Figure 1) generated with a data augmentation technique, where a machine learning model generates synthetic dialogs from a small set of human-written dialogs, and crowdworkers flag and fix errors in the synthetic dataset [3].

Since the example dialogs in an input significantly affect the flow of the conversation [9], the example dialog corpus was inspected to ensure consistency with a specific **agent persona**—*an AI chatbot that calls the user* in a polite and respectful tone and manner—and **system policies** such as the agent should not accept the user's commands that are unsupported by the system (e.g., "*I'll play a song.*" or "*I'll call your daughter.*"). Such a policy was imposed because CareCall's conversation was over a phone call and it did not support many of the task-oriented dialogs that are commonly supported in smart speakers like Alexa or Siri. (Bae *et al.* [3] provides more detailed description of the supported dialogs.) As an additional effort to better steer conversations, the underlying LLM was also *fine-tuned* (*c.f.*, Section 2.3) on the *undesirable* phrases that violated the persona (*e.g.*, the agent acts as if it was a child of the user or speaks impolitely) or system policies in a way which decreased the probability of them being selected [3, 77].

4 METHODOLOGY

To understand the benefits and challenges of LLM-based chatbots as a public health intervention, we observed focus group workshop sessions with 14 CareCall users and interviewed 20 people from three groups of the main stakeholders around the CareCall system: The **users** of CareCall (N = 5), the **teleoperators** who monitored the users' conversations with CareCall (N = 5), and the **developers** of CareCall system (N = 10). We conducted multi-stakeholder interviews because stakeholder groups often had insights into the perspectives or opinions of other stakeholders by virtue of their frequent interactions. For example, teleoperators had insights into

[2]A subset of the corpus is available at https://github.com/naver-ai/carecall-corpus.

how users interact with CareCall and what perspectives they have toward the system through their frequent interactions with users for follow-ups on any health issues. Similarly, UX designers had insights about the perspectives of users and municipal authorities as they conducted formative work with both stakeholders to design and iterate on the system. Business managers also had insights about the perspectives of municipal authorities as they frequently interacted with them to gain feedback on the design and deployment of the system. The quality manager similarly had insights about the real-world usage of CareCall because they were monitoring CareCall logs as part of their work. Together, these interviews aimed to provide a holistic perspective on experiences creating and using such a system. Since our study was conducted in a corporate setting without its own IRB, we submitted our study protocol and obtained IRB approval from an outside public entity that conducts ethical oversight for research. The interview study was approved by the public institutional review board affiliated with the Ministry of Health and Welfare of South Korea. The observation of the focus group workshops was classified as exempt by the guidelines from the Ministry of Health and Welfare of South Korea. In total, we report on insights from 34 people who interacted with different aspects of CareCall including the users (240 total minutes of focus group observation and 230 total minutes of individual interviews), teleoperators (250 total minutes of individual interviews), and developers (430 total minutes of individual interviews). For clarity, we did not have access to nor did we review CareCall users' conversation logs. All interviewees, including teleoperators and developers, did not pull specific conversation logs during the interview sessions, and their perspectives drew from their holistic experiences working with CareCall and its users rather than recalling or reviewing any particular conversation or CareCall user.

4.1 Observation of Focus Group Workshops with CareCall Users

We observed six focus group workshop sessions with 14 CareCall users for four hours in total. The focus group workshops were held by the Seoul Metropolitan Government from mid-July to mid-August of 2022. The workshop participants were middle-aged adults living alone who were participating in the CareCall pilot project in Seoul and had used CareCall for at least two months, having missed no more than a week of calls. The goal of the workshop was to understand the users' perspectives on using CareCall in their daily life and, broadly, to brainstorm ideas about AI-powered public health interventions for middle-aged individuals living alone. The workshop participants included 7 individuals in their 50s and 7 individuals in their 60s (12 males and 2 females) (Table 1a). We did not collect further demographic information on each workshop participant because we were passive observers of the focus group; thus, in this paper, we refer to them as *focus group participants.*

During the workshop, the participants were asked about aspects of CareCall that they liked or did not like and what characteristics they might value in AI-based check-up calls like CareCall. Each session lasted for 40 minutes, with 3 to 6 individuals participating. Note that our research team did not organize or facilitate the focus group workshops. We only took observational notes of the workshops to gain broader perspectives from CareCall users, which was pre-approved by the workshop organizers at the Seoul Metropolitan Government and was made aware to the participants. Through these observations, we sought to better understand what benefits and challenges users perceive when using conversational AI leveraging a large language model as part of public health intervention. We opted for focus group observation because the municipality aimed to protect the privacy of the participants in the public health deployment of CareCall, and therefore understandably did not want to provide us with contact information for the participants. However, the municipality gave us the opportunity to hear how the perspectives of CareCall users contrasted to one another and to recruit interviewees directly through the focus groups. Together with the interview data, the findings from the focus group observation helped deepen our understanding of the users' lived experiences.

4.2 Multi-Stakeholder Interviews

We conducted 1:1 semi-structured interviews with 20 participants from the three groups of stakeholders via Zoom conference calls (N = 8) or in person (N = 12) based on their availability to travel. To compensate for their time and efforts, we offered each participant 50,000 KRW (approximately 38.5 USD as of July 2022) as a gift card.

Interviews with Users. We recruited five CareCall users (P1–5; Table 1a) from the focus group workshops we observed by distributing flyers. Since all CareCall user interviewees were recruited among the participants of the CareCall pilot project in Seoul, they shared demographic characteristics: middle-aged adults who were living alone and were low-SES. The CareCall user interviewees included 2 individuals in their 50s and 3 individuals in their 60s (4 males and 1 female). They had been using CareCall twice a week for two months at the time of the study. We met each interviewee in person in a private meeting room, and each interview lasted for about 60 minutes. The interview questions covered (1) prior experience with receiving regular check-up calls from municipalities or as part of community services; (2) perception of AI phone calls both before and after using CareCall; (3) good and bad experiences with CareCall conversations; and (4) perspectives around AI phone calls in general towards their health care and companionship.

Interviews with Teleoperators. We recruited five teleoperators (T1–5; Table 1b) by distributing flyers to a social enterprise for senior employment that was in charge of the teleoperating task of CareCall in Seoul. Participants had been working as teleoperators for 16 hours a week for about two months at the time of the study. The teleoperator interviewees included 3 individuals in their 50s and 2 individuals in their 60s (1 male and 4 females), with all having relevant experiences such as customer support, social services, or psychological therapy. Each teleoperator was in charge of monitoring 20 to 28 individuals via CareCall. Each interview lasted for about 60 minutes. The interview questions focused on (1) the participants' thoughts on the role and the impact of CareCall on their teleoperating task and broader public health work and (2) their interactions with the users whom they were in charge of.

Interviews with Developers. We recruited ten IT professionals (D1–10; Table 1c) who participated in the design and development of CareCall through a mailing list at NAVER, the vendor of CareCall. With regards to the role in the CareCall development team, the

participants consisted of four UX designers, three machine learning engineers, two business managers, and one quality manager. The developer interviewees' ages ranged from 25 to 51 (5 males and 5 females). The UX designers were in charge of designing the conversation flows and conducting user studies. The machine learning engineers were in charge of improving the language model used for predicting responses and detecting user status. The business managers were in charge of coordinating with municipalities. The quality manager was in charge of monitoring the product quality. Most of the development team members had been involved in this project for about a year at the time of the study, with a few having been involved for about 2 to 3 months. All team members were managing the design and deployment of CareCall across multiple municipalities rather than just Seoul.

Each interview lasted for 40 to 60 minutes. The interviews generally covered the participants' experiences in the development process, including challenges they encountered in designing or implementing aspects of CareCall and communicating with other members and stakeholders. We also focused on different aspects depending on the role of the participants. For instance, we specifically asked UX designers about the rationales and challenges of the conversation design of CareCall. For machine learning engineers, we focused on their thoughts on the unique characteristics and challenges of designing an LLM-based chatbot and how they addressed the challenges.

Table 1: Demographic of the CareCall user interviewees and the focus group participants (a), teleoperator interviewees (b), and developer interviewees (c).

(a) CareCall Users

Alias	Age	Gender
P1	68	Male
P2	59	Male
P3	64	Male
P4	61	Female
P5	54	Male
Focus group participants	50-59	5 males, 1 female
	60-69	7 males, 1 female

(b) Teleoperators

Alias	Age	Gender	Relevant Experience
T1	49	Female	Customer support & social services
T2	51	Female	Customer support & social services
T3	61	Female	Social services
T4	55	Female	Customer support
T5	53	Male	Psychological therapy

(c) CareCall developers

Alias	Age	Gender	Role
D1	30	Female	Business manager
D2	31	Female	UX designer
D3	33	Female	UX designer
D4	51	Male	Business manager
D5	32	Male	Machine Learning engineer
D6	33	Female	UX designer
D7	30	Male	Machine Learning engineer
D8	50	Male	Quality Manager
D9	25	Female	Machine Learning engineer
D10	25	Male	UX designer

4.3 Data Analysis

All interview sessions were audio-recorded and transcribed later. Observational field notes for the focus group workshop sessions were created to capture broader CareCall users' perspectives. We used thematic analysis [8] to qualitatively analyze both interview transcripts and observational notes. The first author open-coded the interview transcripts and the observational notes simultaneously using a spreadsheet, going through several rounds of iterations. Analyzing different data sources together allowed us to verify that the perspectives were present among participants recruited through different techniques. The full research team then discussed and identified patterns and themes through multiple rounds of peer-debriefing meetings. From this coding, we surfaced the main theme about the benefits and challenges around the lack of conversational control in LLM chatbots, which we organized our results around. The final codebook contained 10 parent codes (automation of health monitoring work, performing specific tasks, customizing to different target groups, connecting to social services, emergency management, inappropriate responses, personalization, conversation topics, emotional support, emotional burden) and 24 child codes.

4.4 Limitations

In our study, we specifically focused on the context of Seoul where CareCall was deployed with low-SES middle-aged individuals living alone. Our findings might not represent all target populations' experiences with LLM-driven check-up calls. For example, as explained in Section 3.1, CareCall was deployed in municipalities that have different characteristics of the populations in terms of age groups or health conditions, including older adults living alone in Busan and people with early dementia in Ilsan. These populations likely have different health and companionship needs as well as different perspectives toward LLM-driven chatbots. Similarly, chatbots could be deployed in different social care settings. The teleoperating tasks of the Seoul sample were handled by part-time workers specifically hired for the CareCall pilot project by the Seoul Metropolitan Government. Social welfare officers took the teleoperating tasks as an aspect of their social care work in other municipalities, and therefore, our findings might not generalize to different social care contexts where LLM-driven chatbots could be deployed with different monitoring goals.

Participants' experiences may change as they engage with LLM-based chatbots in the longer term. At the time of the study, the users

and the teleoperators had been engaging with CareCall for two months, being aware that the pilot project would end in a month. Experiences of both users and teleoperators may change if they engage with the system in the longer term. For example, they might become to better understand the capabilities and the limitations of the system so that they can interact with the system in a more informed way; or, their engagement may decrease as they get tired of it over time. Future research on a longitudinal deployment of LLM-driven chatbots for public health interventions would help understand how users' engagement change in the long term.

Our study sample has a skew toward experiences of socially isolated males in their 50s and 60s, which may have impacted the findings. Females who live alone and are younger or older might have different perspectives towards LLM-driven chatbots for social isolation intervention, and their interactions with the system might also be different. Further, our focus on the users who used CareCall regularly (*e.g.*, missed fewer than two calls per week) among the pilot sample may have resulted in participants having a more positive attitude towards the chatbot leveraged in the public health intervention. CareCall users who have occasionally or frequently missed calls or non-users who had dropped out of the intervention might have different, more critical attitudes or perspectives around LLM-driven chatbots. In addition, our interview data overrepresents developers (N = 10) in comparison to teleoperators (N = 5) or users (N = 5). To address this issue, we sought to gain additional insights into the end-user perspectives through the accounts of other stakeholders. However, the end users' original accounts might have been filtered through the lens of these other stakeholders, who have power over the users in how the intervention is ultimately designed and enacted. We also supplemented the end-user perspectives with focus group observations, but this method offered less direct engagement with the users. Therefore, while we have made efforts to represent the perspectives of the socially isolated individuals who used CareCall, our results may not fully capture their lived experiences or their concerns with the technology.

5 FINDINGS

Through the qualitative analysis of interviews and observational notes, we surfaced the lived experiences of the multiple stakeholders who engaged with, managed, and developed a public health intervention leveraging an LLM. In this section, we present the findings of the study, focusing on the benefits and challenges multiple stakeholders—the users, the teleoperators, and the developers—experienced. Note that we blend multiple stakeholders' responses in the findings because stakeholder groups often had insights into the perspectives of others by virtue of their frequent interactions.

5.1 Benefits of Leveraging an LLM-driven Chatbot in Public Health Interventions

Overall, the teleoperators and the users perceived the benefits of leveraging an LLM-driven chatbot in public health intervention. The teleoperators valued that CareCall helped them gain a holistic understanding of each individual through open-ended conversations while offloading their workload. The users perceived the benefits of mitigating loneliness and emotional burdens.

5.1.1 Providing a Holistic Understanding of the Individuals While Offloading Workload. The teleoperators taking care of the CareCall users valued that the system provided a holistic understanding of the individuals through open-ended conversations while offloading their workload. As explained in the background, the dashboard provided a summary of health metrics and emergency alerts so that teleoperators could focus on monitoring and reaching out to cases that need their attention. Teleoperators perceived that the care work process supported by CareCall offloaded a significant amount of workload. T2 said: *"If I were to call all the 26 individuals by myself twice a week, I don't know if I could take on that job. It would be both mentally and physically exhausting to ask the same questions over and over again to that many people."* Based on her previous experience in customer support call centers, T2 assumed that human check-up calls are likely to become redundant and inefficient: *"Human phone calls are likely to get sidetracked. We'll ask questions to check what we need to know, but they'll probably mention other things, too; the phone call might end up being super long, like 30 minutes. That's not feasible given the time frame."* T2, therefore, appreciated that CareCall could manage some of the more redundant aspects of monitoring, allowing them to focus on monitoring individuals who need care the most.

Despite the reduced workload, teleoperators felt that CareCall's open-ended conversations provided rich contextual information to help them gain a holistic understanding of each user's circumstances, which might have been difficult with rule-based dialog systems based on pre-defined scenarios. T5 stated: *"I think I have a pretty good understanding of each person's circumstances at this point because I've been monitoring the call recordings."* T4 noted that the conversation between the CareCall agent and the users surfaced broader aspects of the users' life which were useful for understanding how they are doing: *"Some users are leading a satisfying life, typically people who have jobs, regularly go to a community welfare center, and have friends to meet; I'm not too worried about them. I'm more worried about those who are mostly lying in bed all day and have depression."* This information helped them figure out whom they needed to prioritize monitoring. T4 further stated: *"I mostly focus on monitoring the individuals that I'm concerned about. I got to learn about those individuals over time by monitoring the call recordings."* T5 similarly appreciated: *"CareCall works like a patrol who leads the way and tells us how things are going. I found it really useful to have such information."* The teleoperators further mentioned that thanks to CareCall, they had found cases where some serious health issues might have occurred to the users. T1 and T4 mentioned that they had found users mentioning they had been hospitalized through the conversation logs. Both T1 and T4 were able to then reach out to the users, asking why they were hospitalized and sending emotional support.

5.1.2 Mitigating Loneliness and Emotional Burden. Both CareCall users and teleoperators highlighted how CareCall could help manage people's loneliness and the emotional burdens. The teleoperators mentioned that many of the users had a strong desire to have more conversation opportunities. T1 said, *"There were a few people who cried when I called them. They said they wouldn't have spoken*

a word if I didn't call her that day." T5 similarly noted, *"There are a lot of people who feel terribly lonely. When we called them, the person thanked me, saying that I was the only one who had called them recently."* Teleoperators had observed several instances where users looked forward to receiving the scheduled check-up calls from CareCall. T3 noted, *"I think getting regular check-up calls makes them feel like someone is thinking about them. I noticed some of them looked forward to getting the scheduled calls."* T1 also noted, *"Some people are really looking forward to getting the calls. I notice that they want to talk as much as possible to AI."* T5 further mentioned that some users regularly said *'Thank you'* during the call, which led them to think that the individuals might have received emotional support from CareCall. Teleoperators further perceived that the users enjoyed CareCall's support for diverse conversation topics. T1 mentioned: *"People occasionally talk about their hobbies in detail, for example, paper crafts. Then the AI responded, 'It would be great to showcase your art one day!' I noticed the user was surprised that AI could talk about such things."*

Likewise, the users appreciated receiving check-up calls from CareCall. A focus group participant stated, *"I like getting the AI calls. I feel pretty lonely living alone, so it's nice to have someone to talk to, even though it's a machine."* Another focus group participant similarly said, *"I barely have anyone to talk to after losing my job last year. I feel so empty and lonely. I like that it asks about my health."* Specifically, the users appreciated that the system asked caring questions about their health. A focus group participant noted, *"It was nice to get a phone call checking in with me, asking why I couldn't sleep well last night."* P5 similarly said, *"I feel thankful when they [CareCall] ask caring questions as if they were my wife."*

The users also valued that CareCall covered broader conversation topics beyond health. Specifically, they appreciated that they were able to talk about their hobbies. P5 enjoyed having conversations about his habits in sketching with CareCall: *"When it asked what I was doing, I said I was drawing something. It then responded, 'That sounds fun! I want to learn how to draw too.' I really liked it when it said that. I wanted to talk more about my work."* Other users desired that they could engage in more detailed conversations about cultural life. During the focus group workshops, many participants mentioned their wish that CareCall could recommend movies, TV shows, books, and music or ask about what foods they like. P2 further envisioned that AI could give personalized recommendations based on the conversation data: *"If AI collects a lot of data about us, they might be able to know what sports I am interested in or what kind of art I like. Then it might be reflected in the conversations."*

Furthermore, the CareCall users valued a lack of emotional burden when receiving check-up calls from an AI compared to receiving phone calls from a human. A couple of users noted that they sometimes felt emotionally burdened when contacted by humans. While CareCall was not aimed at replacing other social experiences, a focus group participant said that they might feel more comfortable getting AI calls than getting phone calls from humans: *"My friends might suggest going out for dinner or something when they call me. I sometimes don't want to because of my depression, but I feel uncomfortable turning them down. But I don't need to feel that way to AI."* Another focus group participant similarly mentioned, *"Sometimes I feel more comfortable talking to the AI because it's not a human and doesn't have feelings."* Some participants similarly mentioned the emotional burden that they felt when receiving check-up calls from public health officers. P3 stated: *"I know that some public health officers are checking up on me because I have chronic conditions and live alone. But I feel like they are pretty perfunctory because they only ask one or two questions, and that's it. I would rather prefer getting AI calls."* A focus group participant suggested they might feel emotionally burdened about adding more work to public health officers: *"Sometimes I get phone calls from a public health officer during the weekend. I guess they had too much work during the week, so they had to call me over the weekend. I felt sorry for them. I don't have to feel that way when getting AI calls."*

5.2 Challenges in Leveraging an LLM-driven Chatbot in Public Health Interventions

Despite the benefits, we observed various challenges in leveraging CareCall for public health interventions. In this section, we first describe the inherent challenges of LLMs in uncertainty in control that the developers faced. Next, we illustrate the challenges in leveraging an LLM-driven chatbot, specifically around tailoring it to public health needs and supporting personal health needs.

The CareCall developers frequently mentioned the difficulty in controlling the responses that might not be appropriate for public health contexts. In the initial stage of development, the developers were concerned that the system might generate utterances that make promises that non-human agents could not keep because the LLM embedded in CareCall was pre-trained with human-generated text data (*i.e.*, the Korean corpus depicted as Ⓓ in Figure 1). D3 noted that even though the example dialog corpus (Ⓔ in Figure 1) did not include cases making infeasible suggestions, the system still generated responses doing so: *"When the person said they didn't have any plans this weekend, the agent kept saying infeasible things such as 'How about going to a karaoke with me?' or 'Let's go hiking with me!.' That was the most difficult part in the development process."* The CareCall developers were generally concerned that such suggestions might make the users confused. D9 noted that the developers had to encourage the system to disagree if users made similar suggestions: *"The agent shouldn't suggest, for example, playing billiards together because it can't. Also, it shouldn't say 'yes' when a user makes similar suggestions."* The developers were also concerned about the risk of generating impolite utterances, particularly given the vulnerability of the target population. D2 said, *"Recently, we saw that the agent said something rude, like 'Hope you stay healthy not to burden your family,' which made us freak out."* D7 gave a similar example: *"I don't know what exactly happened, but the system might have detected something wrong and said 'Congratulations!' when the person said they didn't feel well."*

The uncertainty in control largely resulted from the inherent characteristics of LLMs. The developers valued that an LLM enabled them to develop an open domain dialog system much faster and easier compared to other rule-based systems. Because an LLM was used as a backbone model to generate utterances, CareCall was able to cover much broader topics of conversations that would not be feasible for rule-based systems. D9 said, *"LLMs are capable*

of generating various kinds of utterances even without manually defining the rules." However, such characteristics made it difficult for the developers to steer the conversations to prevent inappropriate responses. D3 noted that the responses generated by the backbone model tended to be significantly affected by the large-scale corpus used for the initial pre-training, which includes toxic and biased content that might hurt conversations. D9 further described the process of controlling LLMs: *"Language models have a strong ego, so we have to fight with them. When it generates inappropriate responses, we need to see how it came out, rather than fixing the responses themselves, going through many trials and errors. So it's very difficult to develop a system that is perfectly under control."* D2 noted that such a challenge is a distinct characteristic of LLM-driven chatbots from rule-based ones: *"To fix inappropriate responses of rule-based chatbots, all we need to do is just to modify the scenario. But for LLM-driven ones, we have to consider the patterns where the response came out, which is far more difficult to control."* Even though they incorporated additional steps, including the in-context learning with an example dialog corpus and fine-tuning on the undesirable and inappropriate phrases (*c.f.*, Section 3.2), the developers still acknowledged the uncertainty in control of the system.

5.2.1 Tailoring to Public Health Needs. We noticed several mismatches between the municipalities' needs and LLM-driven chatbots' challenges. First, the CareCall developers faced challenges in addressing the municipalities' needs for asking specific health questions during the calls. Since CareCall was introduced as a technology to assist public health work, the municipalities expected that they could integrate specific questions that they were interested in. For example, D3 mentioned: *"Some local government officials asked if we could integrate dementia screening questionnaires into CareCall."* However, CareCall had inherent uncertainty in controlling the dialog flows. D5 stated: *"What we can do is to fine-tune the model with more datasets that ask certain questions so that the probability of asking such questions becomes higher, but we cannot guarantee that. Such tasks are performed just indirectly."* Therefore, the developers could not accommodate the municipalities' requests. D2 indicated: *"We got asked by several local government officials to ensure that our system asks questions about medication adherence or something. But at least for now, we can't guarantee that."*

Due to the resource-intensive nature of customizing LLMs, the CareCall developers also experienced challenges in customizing to different target groups. Municipalities had different target groups with different monitoring needs in mind, such as older adults living alone in Busan, middle-aged living alone in Seoul, healthy older adults in Gwangju, and people with early dementia in Ilsan. D2 indicated: *"The government of Seoul wanted to deploy CareCall with middle-aged adults because this age group had the highest lonely death cases recently."* Similarly, D3 mentioned that the government of Ilsan had reached out to them, indicating the need for regular check-up calls for older adults with early dementia. However, the developers perceived that CareCall might not fit those groups well because the current dialog corpus (Ⓓ in Figure 1) did not simulate conversations regarding these wildly different health needs. For example, D2 was concerned about deploying CareCall with middle-aged adults: *"When someone says that they have a backache, CareCall is likely to say 'It happens as we age.' A response like this might be perfectly fine for someone in their 70s, but might be odd for someone in their 40s."* D2 also mentioned a similar example with people with early dementia: *"When someone says 'I'm so forgetful these days,' we can simply say 'It happens. I also forget about things sometimes.' But we might need to dig deeper into it if the person had early dementia."* The CareCall developers wished to provide more customized conversations to different target populations given their characteristics and needs, but due to the nature of the example-driven response generation of LLM, tailoring to new target groups demanded new sets of example dialog corpus simulating conversations with those groups. D2 stated such tailoring would not be feasible: *"I wish that the system could provide more customized conversations, but it's not feasible. It's almost like making the example datasets from scratch."* Other CareCall developers similarly mentioned the challenges in customizing to middle-aged adults because of the immense resources needed to generate new sample datasets. Generating new sample datasets would require several iterative cycles of collecting patterns of human-bot dialogs with the specific target population in mind, augmenting the example dialogs with LLM, and labeling positive and negative utterances manually.

In addition, the open-ended nature of LLM-driven chatbots made it challenging for CareCall to manage expectations around the emergency and social service needs. The users wished that the system offered a direct connection to emergency services. They predominantly mentioned their anxiety resulting from living alone, getting older, and having chronic conditions. A focus group participant stated: *"I am getting check-up calls from a community welfare center, a community health center, and a church. I am most concerned about dying alone, so I have applied to all kinds of check-up calls."* P1 similarly mentioned their fear of passing out or dying alone due to their health history involving diabetes or stroke. P1 noted, *"I could pass out at any time. The right side of my face is partially paralyzed because of my diabetes (complications)."* P3 also noted, *"I had a stroke last year, which left my right side of the body paralyzed. I'm worried about having a stroke again when alone."* Therefore, many users desired CareCall could detect emergency situations and automatically call emergency services. However, the developers were not confident about the reliability of the emergency detection, making them hesitant to support such a feature. D3 noted: *"We do not want situations where CareCall fails to detect even just a single case after making a contract that CareCall would detect emergencies and call 911. So we've decided that our product is NOT for actively sending help in emergency situations."*

We further noticed that CareCall users expected that the system would help provide access to a variety of social services, but the developers and the teleoperators felt it was out of scope. D4, D10, and T4 observed that the users asked to join the food assistance program as part of social care for underserved populations. Even though CareCall was not targeted at processing such requests, in some municipalities where the users were managed by social welfare officers, they were able to discover the needs and process the requests. D10 described an instance: *"There are food assistance programs for delivering free lunch boxes for low-SES older adults in most of the municipalities. Through monitoring CareCall logs, the*

public health officers were able to find the need and had the user join the program." In contrast, the teleoperators in Seoul felt confused because they did not have the power to accommodate them as part-time workers who were outside the social service department in their municipalities. T4 said *"They ask for lunch box deliveries, but all we can do is just empathize with them and report it to their municipality. We don't have any power to connect to such social services."* Similarly, D3 and D5 also mentioned that some users requested to fix their refrigerators or fans during their phone calls but were concerned about adding unexpected tasks to public health workers who were managing CareCall. D5 elaborated, *"The public health officers were just in charge of checking whether the individuals were doing well; their job was not to check whether a lunch box had been delivered. When CareCall starts to receive such requests, it adds another task for them."* In addition, T1 and T2 indicated that some users also mentioned that they needed escort services to the doctor's office during their phone calls with CareCall. T1 said: *"Some people were desperate to find someone to go with them to the doctor. I felt really bad, but I couldn't help."* Furthermore, T3 and T4 referred to instances where some of the users requested financial assistance in accessing healthcare services. T3 noted: *"There was a person who kept talking about their circumstances to the AI, like 'I am sick. I need to go see a doctor, but I'm short on money. Can I talk to a person who can help me out?' But AI could only say, 'Why don't you see a doctor?' It's a bit frustrating."* Because the teleoperators did not have the power to help with such requests themselves, they typically relayed the requests to the public health officers in their municipalities when receiving them. Despite the users' needs related to social services, the developers were concerned about the potential burden on the public health officers and wanted to keep the system specifically for regular check-up calls that inform the public health workers of concerning cases.

5.2.2 Supporting Personal Health Needs. We noticed the challenges of LLM-driven chatbots in providing emotional support due to the technical challenges in remembering personal health issues. The teleoperators and the users wished that CareCall would ask personalized questions that consider personal health history. However, due to the technical difficulty in implementing long-term memory in LLM-driven chatbots [79, 80], CareCall could not generate personalized questions and answers that follow up on personal health issues based on past conversations. T5 felt disappointed that the personal health history survey that the teleoperators conducted with the users before rolling out the system was not taken into account to provide personalized conversations: *"One of the individuals that I am in charge of has liver cirrhosis involving ascites. It would have been great if the AI call asked questions like 'Have you seen a doctor to remove the fluid?' based on the pre-survey, but it only asks general questions."* T2, T3, and T5 further mentioned that they felt awkward when the CareCall agents asked inappropriate questions without considering one's current health status. T2 described: *"Some people have severe lower back pain so that they can barely walk. But the AI system kept asking whether they had exercised or whether they had taken a walk. I felt so awkward monitoring such logs."* T5 similarly indicated: *"The person has a chronic condition, so they have already been seeing a doctor. But AI thought that was a new health issue and kept suggesting seeing a doctor."* The users similarly noted that not acknowledging their health issues made the system feel impersonal. A focus group participant said: *"I feel someone understands me and takes care of me when they remember what I've said before. So, when I told them [CareCall] I had a backache, they should have asked questions about that the next time. But they acted as if we had never talked about that."* P3 similarly indicated, *"It would be nice if it could remember that I've seen a doctor and ask follow-up questions. Or, it could at least remember what it has said themselves in the past, like, 'I suggested taking more steps last time. Have you tried it? How did you feel?' Then I could respond, 'Yep, I've tried it as you've suggested. I feel it helped me fall asleep faster."*

The lack of long-term memory of CareCall also limited its ability to provide emotional support to the users. While some users perceived the emotional benefits of the system, others did not partially because of the repetition of general questions and responses across the sessions. For example, they felt that the system always responded in the same way when they mentioned not feeling well. A focus group participant noted, *"It always asks a fixed set of questions like, 'Have you seen a doctor?' when I say I'm not feeling well."* Another focus group participant similarly said: *"When I say something, it always says 'Oh, I see.' I don't feel like we're really communicating."* The repetition of general conversation patterns seemed to interfere with providing emotional support. Some users mentioned feeling like the system was a stranger even after months of engagement. A focus group participant said: *"I've talked to them [CareCall] for a few weeks, but it didn't seem like we got to know each other over time. It always asks the same general questions."* P3 similarly said, *"It's a familiar voice that I've heard for many weeks, but I always feel like talking to a stranger because it never asks specific questions about me. I'd like to talk as if I am talking to an old friend rather than a stranger."* The repetitiveness of the conversations also led the users to feel the conversations were robotic. Several users mentioned that the repetitive utterances felt too machine-like, which decreased their motivation to engage in the conversations. P4 noted, *"I can foresee what it'll ask next or how it'll respond, so I don't get too excited about the conversations."* Another focus group participant also mentioned: *"I don't feel like it really understands how I am doing. It just keeps saying, 'Oh, I see,' so I don't feel it empathizes with me."*

6 DISCUSSION

Our findings from observing focus groups and interviews with multiple stakeholders who created and interacted with CareCall suggest opportunities for leveraging LLM-driven chatbots to support public health interventions. Our findings demonstrated that LLM-driven chatbots have emotional benefits, particularly around supporting broader conversation topics, but also have challenges due to the limited personalization. Based on the findings, we highlight the opportunities for improving emotional support in LLM-driven chatbots. Our findings also pointed to the tensions between multiple stakeholders' needs and the capabilities and limitations of LLM-driven chatbots in public health contexts. We suggest that designing better resources that transparently communicate the respective

capabilities and limitations of open-domain and task-oriented chatbots could help different stakeholders negotiate those tradeoffs. Lastly, we observed tensions around the desire and challenges of scaling LLM-driven chatbots to diverse public health needs. We suggest opportunities for designing mechanisms to help the target populations or care professionals contribute to dialog datasets.

6.1 Improving Emotional Support in LLM-Driven Chatbots

Our findings highlight that technical challenges of LLM-driven chatbots in personalizing responses interfered with providing emotional support. While the users wished that their conversations with CareCall would consider personal health history, the system could not due to the lack of long-term memory[3], which made them feel that the system was impersonal and robotic. Addressing the technical difficulties of implementing long-term memory [79, 80] in LLM-driven chatbots would help resolve part of the challenges in providing conversations that consider personal details such as health history. Future research on investigating how the implementation of long-term memory on chatbots impacts people's perceptions of emotional support would be beneficial.

Accounts from some of the users, such as a user who thought that CareCall would lead some users to reduce their interactions with their social contacts, further point to concern that systems like CareCall might be misapplied to take the place of social support. Prior work highlighted the concern that the introduction of AI technology that supports aging in place could lead to unintended consequences such as reducing human contact with their formal and informal caregivers [25, 37, 66]. For example, if family members know that the older adult is "safe" through AI monitoring technology, they might visit the older adult less frequently. Similarly, if everyday caregiving tasks are replaced by robots at care facilities, older adults might lose the opportunity for caring social interactions. Sharkey *et al.* [66] pointed out that such a reduction in human contact is unethical because it might have a negative impact on the health and wellbeing of the individuals. In addition, recent work argued [35] that LLM-based chatbots are still limited in their conversational abilities to engage in empathetic conversations in sensitive care settings [35]. They further pointed out that LLMs might convey biased perspectives or provide misinformation, which may critically impact the physical and mental health of users [35]. Our study similarly reinforces that technology should not aim to replace the social support that vulnerable populations receive due to technical limitations and potential social consequences, but instead offer an opportunity to increase interaction.

On the other hand, our findings suggest that there is still value in LLM-based chatbots towards other goals, such as supporting conversations on diverse topics. Our findings indicated that the open-ended nature of the conversations helped mitigate loneliness, particularly by supporting broader conversation topics beyond health, such as hobbies and cultural life, which would be challenging to configure rule-based dialog systems to support. Prior work for technology interventions suggested that even surface-level interactions and mere company could help mitigate the loneliness of older adults [14, 58]. In contrast, our study suggests that topic diversity could be one of the key aspects in providing emotional support to individuals who have limited conversation opportunities in their daily life. We highlight the utility of open-domain chatbots in mitigating the loneliness of socially isolated individuals, particularly around supporting diverse conversation topics. Future work on designing LLM-driven chatbots to allow for immersive conversations around specific topics of users' interest can also benefit their abilities to provide emotional support.

[3]In September 2022, after this paper was written, a new version of CareCall with long-term memory [2] was implemented and distributed to the users.

6.2 Tensions between Supporting Informational and Emotional Needs in Public Health Chatbots

Through this study, we found that some of the inherent characteristics of LLM-driven chatbots, such as the uncertainty in control and the resource-intensive nature of customization, led to challenges in supporting different stakeholders' needs in public health interventions. Prior work on chatbots for mental health indicated that expectation management around the system capabilities is challenging but critical [41, 48, 55]. Our findings further highlight that expectation management about open-domain, LLM-driven chatbots can be challenging, particularly in public health settings. From a technical standpoint, open-domain chatbots are radically different from task-oriented chatbots. The primary goal of open-domain chatbots is to support naturalistic conversations on diverse topics, whereas task-oriented chatbots are aimed at performing specific tasks in a closed domain. However, interactions with LLM-driven chatbots performing open-ended conversations are likely to lead various stakeholders in public health interventions to assume that the chatbots can take on the maximal, most flexible set of tasks. Users may assume that the chatbot is a conduit for all things government-related–emergency services, food services, public health care services, financial services, and more. Government agencies can similarly assume that chatbots can take on a whole suite of public health tasks based on the promise of natural conversations. As a consequence, governments may feel disappointed by not being able to get their specific questions answered, and so do the users by not being able to receive the care that they desire.

In the long term, technical advances in better controlling the open-domain chatbots could help address part of this challenge (e.g., ensuring that the chatbot asks specific health questions and supporting direct connections to emergency assistance). However, addressing the larger problems requires understanding multiple stakeholders' needs involved in complex public health settings [36]. Our findings indicated both the governments and the users had some informational needs that could have been better served by more traditional task-oriented systems. For example, task-oriented chatbots can more easily support asking specific health questions that fit governments' needs, such as whether or not a person is adhering to their medication. Task-oriented chatbots could also more reliably respond to a user's request to connect to emergency or social services. In contrast, while open-ended chatbots faced challenges in serving these needs, they demonstrated clear benefits in providing a holistic understanding of care recipients to facilitate care and emotional support through open-ended conversations.

This suggests that, currently, the choice of model puts informational and emotional support in tension with one another.

Prior work on HCI and CSCW has highlighted the challenges in balancing multiple stakeholders' needs when using new technology in complex care settings [28, 57, 61], suggesting the need for mechanisms to assist each stakeholder in voicing and negotiating their needs [5, 28]. When novel and complex technologies like LLM-driven chatbots are introduced in public health interventions, negotiating multiple stakeholders' needs in light of the capabilities and limitations of the system could be even more challenging. Aligned with prior work, our study suggests that when designing one of these open-domain chatbots for public health interventions, it is valuable to have conversations around its capabilities and expectations with multiple stakeholders. Designing resources that transparently communicate the capabilities and limitations of open-domain and task-oriented chatbots could help different stakeholders figure out what type(s) of technology they need and negotiate their needs with each other. In addition, as prior work highlighted [5], it would be beneficial to create opportunities to hear multiple stakeholders' perspectives *before* developing or deploying a system for public health intervention. This opportunity will help developers better recognize what tensions might exist among different stakeholders and what misconceptions they might have toward the system, potentially benefiting the design of conversational prompts to avoid or prevent those.

6.3 Scaling LLM-Driven Chatbots to Diverse Public Health Needs

Our findings surfaced the needs and challenges of LLM-driven chatbots in serving diverse public health needs of different target populations. Prior work has indicated that municipalities frequently have different public health needs from others based on their demographics and organizational capacity [17, 36]. Similarly in our study, we observed that municipalities had different target groups (e.g., older adults living alone, middle-aged adults living alone, and individuals with early dementia) and different ways of handling the teleoperating tasks (e.g., having existing social welfare officers take on the task versus hiring part-time workers). Despite the municipalities' desire for customized conversations based on their needs, CareCall developers found customization infeasible to support due to the immense resources and challenges involved in generating new example datasets. While the open-domain nature and scalability of LLM-driven chatbots make them suitable for addressing the diversity of public health goals that governments might use chatbots for monitoring, when LLM-driven chatbots are deployed in practice, the lack of support for customization could lead to neglecting the specific health needs of different populations and public health monitoring goals.

Efforts to customize LLM chatbots in light of these goals are a valuable direction for future work. However, customizing LLM-driven chatbots to the governments' and end-users' needs involves non-trivial challenges around collecting a relevant dialog corpus. Typically, crowdworkers are often used to take on the task of creating dialog corpus when developing a chatbot; however, they are likely not from the target populations and thus lack a deep understanding of the populations' needs. As a result, even with clear guidelines and training, crowdworkers might find it challenging to create datasets that reflect the populations' needs. Developing mechanisms for the target populations to effectively contribute dialog datasets could help overcome such challenges. Prior work in personal informatics has shown promise for speech interactions for collecting personal health data (*e.g.*, [31, 46, 47]). Relevant to our work, Kim *et al.* [31] have proposed a speech-based smartwatch app to assist older adults in labeling physical activities with a low capture burden. Similar approaches could help target populations in collecting dialog datasets in an accessible way, leading to developing chatbots that are more well-suited for them. However, not all target populations in public health contexts might be reliable to perform such tasks. For example, individuals with dementia might be less reliable in collecting and labeling dialog datasets, depending on their cognitive abilities or motor skills. Furthermore, collecting private data, such as everyday conversations, for machine learning purposes involves privacy concerns [71], particularly with marginalized populations [53]. An alternative approach would be to have experienced social or health care professionals who have a good understanding of the target populations contribute to the dialog datasets. However, this approach involves concerns over adding burdens to already overburdened professionals. Future research is needed to explore ways to help care professionals contribute to the creation of dialog datasets that better suit target populations' needs in chatbot-based interventions.

7 CONCLUSION

Through observing focus groups and interviews with multiple stakeholders who created and interacted CareCall, we found that LLM-driven chatbots can provide emotional benefits, such as supporting broader conversation topics, but also have difficulties providing emotional support due to limited personalization of conversations. We also observed tensions between multiple stakeholders' needs and the capabilities and limitations of LLM-driven chatbots in public health contexts, with municipalities often desiring specific health questions to be asked, with LLMs lacking that level of control. Based on the findings, we highlight that implementation of long-term memory could improve emotional support in LLM-driven chatbots. We further suggest designing better resources and processes that help multiple stakeholders negotiate the respective tradeoffs of open-domain and task-oriented chatbots. Lastly, our work points to a need to explore how to scale LLM-driven chatbots to diverse public health needs, suggesting opportunities for designing mechanisms to help the target populations or care professionals contribute to dialog datasets. In closing, we hope this work can inspire collaborations among the researchers in the HCI, Public health, and NLP communities to design chatbots leveraging large language models for public health intervention.

ACKNOWLEDGMENTS

We thank our participants for their sincere participation. We are also grateful to Sang-houn Ok and HaYeon Kang at NAVER for helping us recruit study participants. Jing Wei gave feedback on the early version of this paper. This work was supported as a research internship at NAVER AI Lab.

REFERENCES

[1] Ingrid Arreola, Zan Morris, Matthew Francisco, Kay Connelly, Kelly Caine, and Ginger White. 2014. From checking on to checking in: designing for low socio-economic status older adults. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, Toronto Ontario Canada, 1933–1936. https://doi.org/10.1145/2556288.2557084

[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*. Association for Computational Linguistics, 3769–3787. https://preview.aclanthology.org/emnlp-22-ingestion/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] Madeline Balaam, Rob Comber, Ed Jenkins, Selina Sutton, and Andrew Garbett. 2015. FeedFinder: A Location-Mapping Mobile Application for Breastfeeding Women. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 1709–1718. https://doi.org/10.1145/2702123.2702328

[5] Vince Bartle, Janice Lyu, Freesoul El Shabazz-Thompson, Yunmin Oh, Angela Anqi Chen, Yu-Jan Chang, Kenneth Holstein, and Nicola Dell. 2022. "A Second Voice": Investigating Opportunities and Challenges for Interactive Voice Assistants to Support Home Health Aides. In *CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–17. https://doi.org/10.1145/3491102.3517683

[6] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 610–623. https://doi.org/10.1145/3442188.3445922

[7] Shikha Bordia and Samuel R. Bowman. 2019. Identifying and Reducing Gender Bias in Word-Level Language Models. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*. Association for Computational Linguistics, Minneapolis, Minnesota, 7–15. https://doi.org/10.18653/v1/N19-3002

[8] 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

[9] 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

[10] Hye-jin Byun. 2022. NAVER launches AI call service aimed at seniors - The Korea Herald. Retrieved Sep 15, 2022 from https://www.koreaherald.com/view.php?ud=20220530000643

[11] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. 2021. Evaluating Large Language Models Trained on Code. https://doi.org/10.48550/ARXIV.2107.03374

[12] 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

[13] John Joon Young Chung, Wooseok Kim, Kang Min Yoo, Hwaran Lee, Eytan Adar, and Minsuk Chang. 2022. TaleBrush: Sketching Stories with Generative Pretrained Language Models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 209, 19 pages. https://doi.org/10.1145/3491102.3501819

[14] Simon Coghlan, Jenny Waycott, Amanda Lazar, and Barbara Barbosa Neves. 2021. Dignity, Autonomy, and Style of Company: Dimensions Older Adults Consider for Robot Companions. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021). https://doi.org/10.1145/3449178

[15] Sunny Consolvo, Peter Roessler, and Brett E. Shelton. 2004. The CareNet Display: Lessons Learned from an In Home Evaluation of an Ambient Display. In *UbiComp 2004: Ubiquitous Computing*, David Hutchison, Takeo Kanade, Josef Kittler, Jon M. Kleinberg, Friedemann Mattern, John C. Mitchell, Moni Naor, Oscar Nierstrasz, C. Pandu Rangan, Bernhard Steffen, Madhu Sudan, Demetri Terzopoulos, Dough Tygar, Moshe Y. Vardi, Gerhard Weikum, Nigel Davies, Elizabeth D. Mynatt, and Itiro Siio (Eds.). Vol. 3205. Springer Berlin Heidelberg, Berlin, Heidelberg, 1–17. https://doi.org/10.1007/978-3-540-30119-6_1 Series Title: Lecture Notes in Computer Science.

[16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. https://doi.org/10.48550/ARXIV.1810.04805

[17] Dieudonne Diasso, Maimouna Halidou Doudou, Mohamed Cheikh Levrak, Holly Dente Sedutto, and Aly Savadogo. 2021. Municipalities' organisational capacity to support the implementation of the Multi-Sector Nutrition Plan in Burkina Faso. *Global Health Action* 14, 1 (2021). https://doi.org/10.1080/16549716.2021.1979279

[18] Nancy J. Donovan and Dan Blazer. 2020. Social Isolation and Loneliness in Older Adults: Review and Commentary of a National Academies Report. *American Journal of Geriatric Psychiatry* 28, 12 (2020), 1233–1244. https://doi.org/10.1016/j.jagp.2020.08.005

[19] Mihail Eric, Rahul Goel, Shachi Paul, Abhishek Sethi, Sanchit Agarwal, Shuyang Gao, Adarsh Kumar, Anuj Goyal, Peter Ku, and Dilek Hakkani-Tur. 2020. MultiWOZ 2.1: A Consolidated Multi-Domain Dialogue Dataset with State Corrections and State Tracking Baselines. In *Proceedings of the 12th Language Resources and Evaluation Conference*. European Language Resources Association, Marseille, France, 422–428. https://aclanthology.org/2020.lrec-1.53

[20] Jianfeng Gao, Michel Galley, and Lihong Li. 2018. Neural Approaches to Conversational AI. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*. Association for Computational Linguistics, Melbourne, Australia, 2–7. https://doi.org/10.18653/v1/P18-5002

[21] Ismael Garrido-Muñoz, Arturo Montejo-Ráez, Fernando Martínez-Santiago, and L. Alfonso Ureña-López. 2021. A Survey on Bias in Deep NLP. *Applied Sciences* 11, 7 (2021). https://doi.org/10.3390/app11073184

[22] Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, Online, 3356–3369. https://doi.org/10.18653/v1/2020.findings-emnlp.301

[23] Minlie Huang, Xiaoyan Zhu, and Jianfeng Gao. 2020. Challenges in Building Intelligent Open-Domain Dialog Systems. *ACM Trans. Inf. Syst.* 38, 3, Article 21 (apr 2020), 32 pages. https://doi.org/10.1145/3383123

[24] Bernd Huber, Daniel McDuff, Chris Brockett, Michel Galley, and Bill Dolan. 2018. Emotional Dialogue Generation Using Image-Grounded Language Models. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3173574.3173851

[25] Lesa Lorenzen Huber, Kalpana Shankar, Kelly Caine, Kay Connelly, L. Jean Camp, Beth Ann Walker, and Lisa Borrero. 2013. How In-Home Technologies Mediate Caregiving Relationships in Later Life. *International Journal of Human-Computer Interaction* 29, 7 (July 2013), 441–455. https://doi.org/10.1080/10447318.2012.715990

[26] 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

[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] Eunkyung Jo, Seora Park, Hyeonseok Bang, Youngeun Hong, Yeni Kim, Jungwon Choi, Bung Nyun Kim, Daniel A. Epstein, and Hwajung Hong. 2022. GeniAuti: Toward Data-Driven Interventions to Challenging Behaviors of Autistic Children through Caregivers' Tracking. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW1 (mar 2022), 1–27. https://doi.org/10.1145/3512939

[29] 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

[30] Taewan Kim, Mintra Ruensuk, and Hwajung Hong. 2020. In Helping a Vulnerable Bot, You Help Yourself: Designing a Social Bot as a Care-Receiver to Promote Mental Health and Reduce Stigma. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–13. https://doi.org/10.1145/3313831.3376743

[31] Young-Ho Kim, Diana Chou, Bongshin Lee, Margaret Danilovich, Amanda Lazar, David E. Conroy, Hernisa Kacorri, and Eun Kyoung Choe. 2022. MyMove: Facilitating Older Adults to Collect In-Situ Activity Labels on a Smartwatch with Speech. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. ACM, New York, NY, USA. https://doi.org/10.1145/3491102.3517457

[32] Young-Ho Kim, Sungdong Kim, Minsuk Chang, and Sang-Woo Lee. 2022. Leveraging Pre-Trained Language Models to Streamline Natural Language Interaction for Self-Tracking. In *NAACL '22 The 2nd Workshop on Bridging Human-Computer Interaction and Natural Language Processing*. arXiv. https://doi.org/10.48550/ARXIV.2205.15503

[33] Eric Klinenberg. 2002. *Heat wave: A social autopsy of disaster in Chicago*. University of Chicago Press.

[34] 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

[35] Diane M. Korngiebel and Sean D. Mooney. 2021. Considering the possibilities and pitfalls of Generative Pre-trained Transformer 3 (GPT-3) in healthcare delivery. *npj Digital Medicine* 4, 93 (jun 2021). https://doi.org/10.1038/s41746-021-00464-x

[36] Sanna Kuoppamäki. 2021. The application and deployment of welfare technology in Swedish Municipal Care: A qualitative study of procurement practices among municipal actors. *BMC Health Services Research* 21, 1 (2021). https://doi.org/10.1186/s12913-021-06944-w

[37] Amanda Lazar, Hilaire J. Thompson, Shih Yin Lin, and George Demiris. 2018. Negotiating relation work with telehealth home care companionship technologies that support aging in place. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018). https://doi.org/10.1145/3274372

[38] 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

[39] Mina Lee, Percy Liang, and Qian Yang. 2022. CoAuthor: Designing a Human-AI Collaborative Writing Dataset for Exploring Language Model Capabilities. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 388, 19 pages. https://doi.org/10.1145/3491102.3502030

[40] 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

[41] 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

[42] 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

[43] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. https://doi.org/10.48550/ARXIV.2107.13586

[44] 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, 26 pages. https://doi.org/10.1145/3555569

[45] 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

[46] Yuhan Luo, Young-Ho Kim, Bongshin Lee, Naeemul Hassan, and Eun Kyoung Choe. 2021. FoodScrap: Promoting Rich Data Capture and Reflective Food Journaling Through Speech Input. In *Designing Interactive Systems Conference 2021* (Virtual Event, USA) *(DIS '21)*. ACM, New York, NY, USA, 606–618. https://doi.org/10.1145/3461778.3462074

[47] Yuhan Luo, Bongshin Lee, and Eun Kyoung Choe. 2020. TandemTrack: Shaping Consistent Exercise Experience by Complementing a Mobile App with a Smart Speaker. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. ACM, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376616

[48] Wookjae Maeng and Joonhwan Lee. 2022. Designing and Evaluating a Chatbot for Survivors of Image-Based Sexual Abuse. In *CHI Conference on Human Factors in Computing Systems*. ACM, New Orleans LA USA, 1–21. https://doi.org/10.1145/3491102.3517629

[49] Nikita Moghe, Siddhartha Arora, Suman Banerjee, and Mitesh M. Khapra. 2018. Towards Exploiting Background Knowledge for Building Conversation Systems. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 2322–2332. https://doi.org/10.18653/v1/D18-1255

[50] Manish Munikar, Sushil Shakya, and Aakash Shrestha. 2019. Fine-grained Sentiment Classification using BERT. In *2019 Artificial Intelligence for Transforming Business and Society (AITB)*, Vol. 1. 1–5. https://doi.org/10.1109/AITB48515.2019.8947435

[51] Elizabeth D. Mynatt, Irfan Essa, and Wendy Rogers. 2000. Increasing the opportunities for aging in place. In *Proceedings on the 2000 conference on Universal Usability - CUU '00*. ACM Press, Arlington, Virginia, United States, 65–71. https://doi.org/10.1145/355460.355475

[52] National Academies of Sciences, Engineering, and Medicine. 2020. *Social Isolation and Loneliness in Older Adults: Opportunities for the Health Care System*. The National Academies Press, Washington, DC. https://doi.org/10.17226/25663

[53] V. Nirmala and A. Rajagopal. 2019. Artifically intelligent physics solver: This AI understands Newtons Law. *Science & Technology Journal* 7, 1 (2019), 22–28. https://doi.org/10.22232/stj.2019.07.01.03

[54] Alberto Olmo, Sarath Sreedharan, and Subbarao Kambhampati. 2021. GPT3-to-plan: Extracting plans from text using GPT-3. In *ICAPS '21 Workshop on Knowledge Engineering for Planning and Scheduling*. https://doi.org/10.48550/ARXIV.2106.07131

[55] 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

[56] Sachin R. Pendse, Amit Sharma, and Aditya Vashistha. 2021. Can i not be suicidal on a sunday?: Understanding technology-mediated pathways to mental health support. *Conference on Human Factors in Computing Systems - Proceedings* (2021). https://doi.org/10.1145/3411764.3445410

[57] Adrienne Pichon, Kayla Schiffer, Emma Horan, Bria Massey, Suzanne Bakken, Lena Mamykina, and Noemie Elhadad. 2021. Divided We Stand: The Collaborative Work of Patients and Providers in an Enigmatic Chronic Disease. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW3 (2021). https://doi.org/10.1145/3434170

[58] Alisha Pradhan, Leah Findlater, and Amanda Lazar. 2019. "Phantom friend" or "just a box with information": personification and ontological categorization of smart speaker-based voice assistants by older adults. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019). https://doi.org/10.1145/3359316

[59] Chen Qu, Liu Yang, Minghui Qiu, W. Bruce Croft, Yongfeng Zhang, and Mohit Iyyer. 2019. BERT with History Answer Embedding for Conversational Question Answering. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Paris, France) *(SIGIR'19)*. Association for Computing Machinery, New York, NY, USA, 1133–1136. https://doi.org/10.1145/3331184.3331341

[60] Divya Ramachandran, John Canny, Prabhu Dutta Das, and Edward Cutrell. 2010. Mobile-izing health workers in rural India. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Atlanta, Georgia, USA) *(CHI '10)*. Association for Computing Machinery, New York, NY, USA, 1889–1898. https://doi.org/10.1145/1753326.1753610

[61] Olivia K. Richards, Adrian Choi, and Gabriela Marcu. 2021. Shared Understanding in Care Coordination for Children's Behavioral Health. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021). https://doi.org/10.1145/3449095

[62] Celia Roberts, Maggie Mort, and Christine Milligan. 2012. Calling for Care: 'Disembodied' Work, Teleoperators and Older People Living at Home. *Sociology* 46, 3 (2012), 490–506. https://doi.org/10.1177/0038038511422551

[63] Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston.

2021. Recipes for Building an Open-Domain Chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Association for Computational Linguistics, Online, 300–325. https://doi.org/10.18653/v1/2021.eacl-main.24

[64] Jim Rowan and Elizabeth D. Mynatt. 2005. Digital Family Portrait Field Trial: Support for Aging in Place. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, Portland Oregon USA, 521–530. https://doi.org/10.1145/1054972.1055044

[65] Hyeyoung Ryu, Soyeon Kim, Dain Kim, Sooan Han, Keeheon Lee, and Younah Kang. 2020. Simple and Steady Interactions Win the Healthy Mentality: Designing a Chatbot Service for the Elderly. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (Oct. 2020), 1–25. https://doi.org/10.1145/3415223

[66] Amanda Sharkey and Noel Sharkey. 2012. Granny and the robots: Ethical issues in robot care for the elderly. *Ethics and Information Technology* 14, 1 (2012), 27–40. https://doi.org/10.1007/s10676-010-9234-6

[67] Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. 2021. Societal Biases in Language Generation: Progress and Challenges. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Online, 4275–4293. https://doi.org/10.18653/v1/2021.acl-long.330

[68] Vered Shwartz and Yejin Choi. 2020. Do Neural Language Models Overcome Reporting Bias?. In *Proceedings of the 28th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, Barcelona, Spain (Online), 6863–6870. https://doi.org/10.18653/v1/2020.coling-main.605

[69] Emma Simpson, Rob Comber, Andrew Garbett, Ed Ian Jenkins, and Madeline Balaam. 2017. Experiences of Delivering a Public Health Data Service. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 6171–6183. https://doi.org/10.1145/3025453.3025881

[70] Elizabeth Stowell, Mercedes C. Lyson, Herman Saksono, Reneé C. Wurth, Holly Jimison, Misha Pavel, and Andrea G. Parker. 2018. Designing and evaluating mhealth interventions for vulnerable populations. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (2018). https://doi.org/10.1145/3173574.3173589

[71] Divy Thakkar, Azra Ismail, Pratyush Kumar, Alex Hanna, Nithya Sambasivan, and Neha Kumar. 2022. When is machine learning data good?: Valuing in public health datafication. *CHI Conference on Human Factors in Computing Systems* (2022). https://doi.org/10.1145/3491102.3501868

[72] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All You Need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Long Beach, California, USA) *(NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 6000–6010.

[73] Tiffany C. Veinot, Jessica S. Ancker, Heather Cole-Lewis, Elizabeth D. Mynatt, Andrea G. Parker, Katie A. Siek, and Lena Mamykina. 2019. Leveling up: On the Potential of Upstream Health Informatics Interventions to Enhance Health Equity. *Medical Care* 57, Suppl 2 (Jun 2019). https://doi.org/10.1097/mlr.0000000000001032

[74] John Vines, Stephen Lindsay, Gary W. Pritchard, Mabel Lie, David Greathead, Patrick Olivier, and Katie Brittain. 2013. Making family care work: Dependence, privacy and remote home monitoring telecare systems. *UbiComp 2013 - Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (2013), 607–616. https://doi.org/10.1145/2493432.2493469

[75] 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

[76] Liuping Wang, Dakuo Wang, Feng Tian, Zhenhui Peng, Xiangmin Fan, Zhan Zhang, Mo Yu, Xiaojuan Ma, and Hongan Wang. 2021. CASS: Towards Building a Social-Support Chatbot for Online Health Community. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–31. https://doi.org/10.1145/3449083

[77] Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. 2019. Neural Text Generation with Unlikelihood Training. https://doi.org/10.48550/ARXIV.1908.04319

[78] Ginger White, Tanya Singh, Kelly Caine, and Kay Connelly. 2015. Limited but satisfied: Low SES older adults experiences of aging in place. In *Proceedings of the 9th International Conference on Pervasive Computing Technologies for Healthcare*. ICST, Istanbul, Turkey. https://doi.org/10.4108/icst.pervasivehealth.2015.259095

[79] Jing Xu, Arthur Szlam, and Jason Weston. 2022. Beyond Goldfish Memory: Long-Term Open-Domain Conversation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Dublin, Ireland, 5180–5197. https://doi.org/10.18653/v1/2022.acl-long.356

[80] 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

[81] 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

[82] Jiacheng Yang, Mingxuan Wang, Hao Zhou, Chengqi Zhao, Weinan Zhang, Yong Yu, and Lei Li. 2020. Towards Making the Most of BERT in Neural Machine Translation. *Proceedings of the AAAI Conference on Artificial Intelligence* 34, 05 (Apr. 2020), 9378–9385. https://doi.org/10.1609/aaai.v34i05.6479

[83] Yonhap. 2017. Lonely deaths of middle-aged, youth brackets stand out amid single-person households - The Korea Herald. Retrieved Sep 15, 2022 from https://www.koreaherald.com/view.php?ud=20171207000623

[84] Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing Dialogue Agents: I have a dog, do you have pets too?. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Melbourne, Australia, 2204–2213. https://doi.org/10.18653/v1/P18-1205

[85] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. OPT: Open Pre-trained Transformer Language Models. https://doi.org/10.48550/ARXIV.2205.01068

[86] Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2018. Emotional Chatting Machine: Emotional Conversation Generation with Internal and External Memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32. https://doi.org/10.1609/aaai.v32i1.11325