Incorporating Multi-Stakeholder Perspectives in Evaluating and Auditing of Health Chatbots Driven by Large Language Models

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

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

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

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

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

ABSTRACT

Recent large language models (LLMs) offer the potential to support public health interventions by monitoring populations at scale through open-ended conversations. When leveraging LLMs in complex health contexts, it is crucial to evaluate and audit the LLM-infused system by incorporating the perspectives of multiple stakeholders, such as developers, public health officials, community health workers, and care recipients. However, we have a limited understanding of how different stakeholders perceive LLM-driven chatbots in collaborative health work and care. Reflecting on our two studies on CareCall—an LLM-driven voice chatbot that aims to support socially isolated individuals via check-up phone calls, we provide insights into how researchers can better account for different stakeholders' needs in the design and deployment of LLM-driven chatbots in public health contexts.

1 INTRODUCTION

Chatbots have been proposed as effective tools for scaling abilities to provide informational and emotional support around health [15, 30], thereby facilitating population-level health interventions. Recent large language models (LLMs) brought breakthroughs in chatbots, especially in performing free-form conversations in open-ended topics (*e.g.*, MindfulDiary [12], Chacha [23]). Such systems can be beneficial for public health interventions in providing empathetic interactions for populations going through difficult health experiences [18] while offloading the burden of public health authorities in monitoring people at scale. LLM-driven chatbots also offer the potential to be effective at eliciting disclosure about broader aspects of personal health, which presents challenges in sensitive health domains [4, 28].

The introduction of LLM-driven chatbots offers great potential to enhance public health work and care, but experiences may be uneven across stakeholders. For example, the adoption of new technology could add tasks to public health workers, despite it being beneficial for care recipients [9, 25]. Conversely, although technology could reduce the burdens of public health workers by automating the collection of personal health information from populations, technology 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 [17, 18]. To holistically evaluate and audit LLM-driven systems in collaborative health contexts, it is crucial to gather perspectives from different stakeholders. However, few studies have explored the use of LLM-driven chatbots in population-level health interventions in real-world settings, limiting an understanding of multi-stakeholder perspectives around such systems in complex health contexts.

In this encore submission, we revisit and expand on our two studies on CareCall[1], a commercial LLM-driven voice chatbot that aims to support socially isolated individuals via check-up phone calls. In **Study 1** [9], we examined the benefits and challenges of leveraging CareCall for public health intervention. Through focus group observations and interviews with 34 people from three stakeholder groups, including users, teleoperators, and developers, we found that CareCall offered a holistic understanding of care recipients while offloading the public health workload and helped mitigate the loneliness that individuals were experiencing. However, our findings highlight that traits of LLM-driven chatbots led to challenges in supporting public and personal health needs. In **Study 2** [10], we investigated how the integration of long-term memory (LTM) might impact user interactions with and perceptions of an LLM-driven chatbot. Through the analysis of 1,252 call logs and interviews with nine users, we found that LTM enhanced health disclosure and fostered positive perceptions of the chatbot by offering familiarity. However, we also observed challenges in promoting self-disclosure through LTM, particularly around addressing chronic health conditions and privacy concerns.

Reflecting on the findings from the two studies, we suggest the need to carefully balance users' needs for topic diversity with public health needs for targeted health data collection when designing evaluation metrics for LLM-driven chatbots for public health monitoring. We also highlight the need for improved auditing processes to clarify LLM-driven chatbots' capabilities and limitations, ensuring that these systems align with more ethically aligned and responsive to different stakeholder needs. We further propose that LTM's ability to generate more thoughtful follow-up questions on sensitive health topics should be carefully considered in the auditing process. Lastly, we call for rigorous auditing of LTM, considering the tensions between users' privacy needs and public health monitoring goals.

1st HEAL Workshop at CHI Conference on Human Factors in Computing Systems, May 12, Honolulu, HI, USA
2024.

[1] CareCall is a service provided by NAVER Cloud. See https://guide.ncloud-docs.com/docs/en/clovacarecall-overview for more detail.

2 RELATED WORK

2.1 Scaffolding Open-Ended Conversations with Large Language Models

Recent LLMs have brought breakthroughs in chatbots, especially those that perform open-ended conversations, thanks to their capabilities in generating coherent and contextual responses through in-context learning [2, 22]. LLM-driven chatbots receive the current dialog history (*i.e.*, list of messages exchanged between the user and the agent) as model input and infer the agent's following response accordingly [22]. 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. Such an intuitive approach to scaffolding an open-ended chatbot has motivated the development of LLM-driven chatbots both by practitioners (*e.g.*, ChatGPT [19], Bard [6], Pi [7]) and researchers (*e.g.*, [12, 23, 24, 32]).

Being in the early stage of real-world adoption and academic research, however, designing LLM-driven chatbots involves significant challenges. As LLMs generate the most probable output based on a complex structure of neural networks (*transformers* [27]), it is not explainable how an LLM *'reads'* the model input written in natural language [16]. Therefore, it is challenging for chatbots 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 [2]. One known method to steer the conversations to converge towards desired scenarios is to put ideal conversation examples in the model input together [2]. 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 [2, 29].

2.2 Augmenting Large Language Model-Driven Chatbots with Long-Term Memory

Most exemplary LLM-driven chatbots, represented by ChatGPT [19] and Bard [6], did not suppose repetitive interaction scenarios in their early versions, resulting in each session not informing the following ones because they were mainly designed for single-shot tasks such as code generation and reasoning. Recently, OpenAI started testing the ability for ChatGPT to automatically store information from conversations with individual users to enhance future conversations, which is currently available to a select group of users [20]. However, since the memory capability is in its early stage, we lack an understanding of user perceptions and reactions to this new feature.

Augmenting LLMs to "remember" past information—often referred to as "long-term memory" [1, 31, 33–35]—presents significant challenges for two main reasons. First, LLMs can receive input text only within a limited context window (input size). Thus, it is not feasible to include the entire conversation session history in the model input for longer-term interactions. One common approach is to include summarized information of the conversation history instead of a raw knowledge base (*e.g.*, [1, 14, 31]). Second, designing how chatbots should refer to stored information back in conversation involves complex considerations. For example, Cox *et al.* [5] found that the phrasing style of user messages in past conversations impacts the perceived intelligence of and engagement with chatbots as well as privacy concerns around them, suggesting the importance of careful LTM design. Motivated by the gap in understanding the utility of LTM in LLM-driven chatbots, we explore the case of CareCall, a rare example of an LLM-driven chatbot that automatically stores and updates key information from previous conversations to support public health monitoring.

3 BACKGROUND: CLOVA CARECALL

3.1 Motivation and Deployment of CareCall

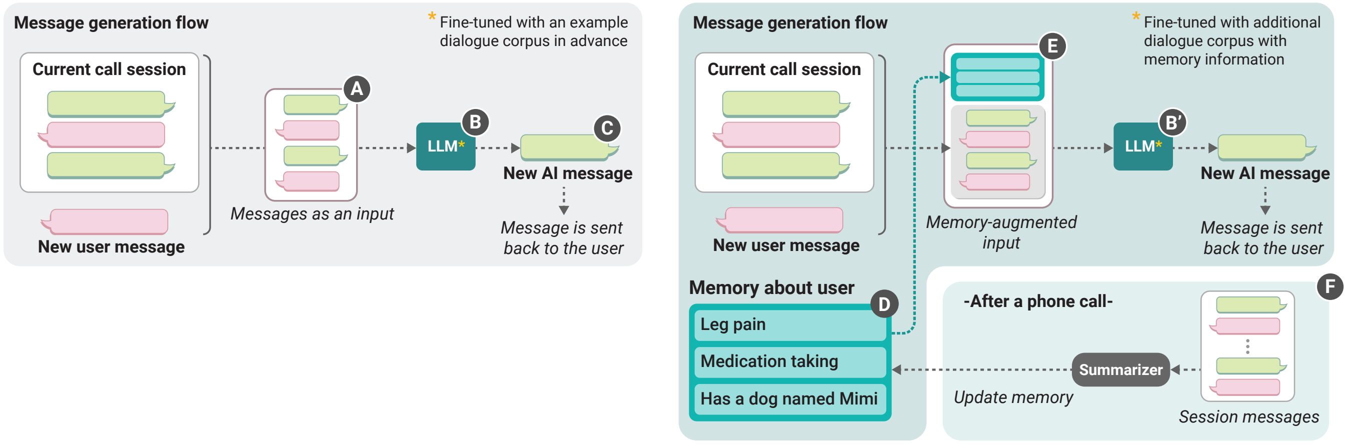
CareCall is an LLM-driven voice chatbot designed for socially isolated individuals [3]. The chatbot calls the users weekly and engages in an open-ended conversation for about 2 to 3 minutes to check in with their health and overall wellbeing and provide emotional support. First rolled out in a local municipality in South Korea in November 2021 [3] and expanded to others over time, as of October 2022, CareCall serves around 6,000 individuals across different municipalities in Korea, targeting middle-aged (40s to 60s) and older adults (60s or older) living alone. The system aims to address the rising number of lonely deaths among low-SES populations. Public officers in most deployment areas monitored call recordings to identify and act upon any signs of declining health or missed calls.

3.2 CareCall and Long-Term Memory

CareCall's initial setup did not include LTM, operating from November 2021 to September 2022. Powered by an LLM called HyperCLOVA [11] (Ⓑ in Figure 1a, CareCall feeds the current dialogue history into the LLM (Ⓑ in Figure 1a) to generate a response (Ⓒ in Figure 1a) that naturally continues the conversation.

LTM was integrated into the existing deployments in September 2022 to improve its ability to offer familiarity with users. In this version, a summarizer driven by an LLM (Ⓕ in Figure 1b; [1]) generates summaries relevant to the five LTM topics each call. The memory management layer (Ⓓ in Figure 1b) stores and updates the summary sentences upon each call (*e.g.*, Removing the "Regular visit to a clinic due to leg pain" status after a user reports that they have completed the treatment). Unlike CareCall without LTM, the stored information from previous sessions is included in the model input (Ⓔ in Figure 1b), providing cues for the chatbot to refer to (*e.g.*, "*How is your leg feeling?*").

CareCall's LTM stores summarized information about five topics: (1) *Health* (*e.g.*, what health issues they have, what type of clinical care they are seeking), (2) *Meals* (*e.g.*, whether and why users are having difficulty eating), (3) *Sleep* (*e.g.*, what difficulty they are experiencing related to sleep), (4) *Pets* (*e.g.*, what kind of pets they have, what they do with their pets), and (5) *Visited Places* (*e.g.*, what places users visit frequently). LTM remembers any noteworthy information that comes up during conversations relevant to the five LTM topics, including both positive and negative health experiences. For example, when users mention that they have been seeing a doctor for leg pain, CareCall would ask LTM-triggered questions in later sessions, such as *"You mentioned having knee joint issues last time. Are you still seeing the doctor?"*



(a) CareCall without LTM

(b) CareCall with LTM

Figure 1: Architecture of the two different versions of CareCall chatbots, an open-ended dialogue system powered by an LLM called HyperCLOVA [11]. See Jo *et al.* [10] for the full descriptions of the architecture.

4 REFLECTION ON TWO STUDIES

From the two studies, we learned that the evaluation of LLM-driven systems for health contexts should be seen as a multi-dimensional process, focusing on how these chatbots might meet the diverse needs of various stakeholders. Based on the lessons learned across the two studies, we highlight the importance of comprehensive evaluation and auditing to address different stakeholders' needs around LLM-driven chatbots in public health contexts.

4.1 Tensions between Multi-Stakeholder Needs in the Evaluation of LLM-Driven Chatbots

Across the two studies, we learned that the design of LLM-driven chatbots, particularly long-term memory, significantly influenced on user engagement and interactions with chatbots. In Study 1, users found the chatbot impersonal or robotic when it could not remember or follow up on users' personal health history due to the lack of long-term memory. On the other hand, in Study 2, users perceived CareCall as personal and caring as LTM offered familiarity with users, fostering positive reactions and eliciting greater health disclosure. This contrast suggests the pivotal role of chatbots' ability to demonstrate emotional support in shaping user experiences with LLM-driven health chatbots.

Across the two studies, users highlighted how CareCall could help mitigate their loneliness, particularly by supporting conversations on diverse topics, including hobbies and interests. This finding 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 lives. Yet, overemphasis on topic diversity could diverge from the primary goals of public health authorities. Public health workers primarily focus on collecting specific health-related data, which helps them figure out whom they need to prioritize monitoring and sending support, given their limited resources [9, 10]. Conversely, when developers focus on supporting diverse topics in the design of LLM-driven chatbots, there is a risk that the conversations may deviate significantly from collecting data pertinent to the principal goals of public health workers [10]. This dichotomy underscores the complexity of developing balanced evaluation methods that cater to both individual users' needs for conversations on diverse topics and the focused informational needs of public health initiatives. In designing evaluation metrics for complex public health contexts, it is important to navigate this delicate balance between the personal needs for topic diversity and public health needs for targeted health information.

4.2 Promoting Communication Transparency Around LLM-driven Chatbots' Capabilities and Limitations through Auditing

In both studies, we found that some inherent traits of LLM-driven chatbots, such as uncertainty in control, led to challenges in supporting different stakeholders' needs in public health interventions. Expectation management about open-domain, LLM-driven chatbots can be challenging, particularly in public health settings. We posit that 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. This mismatch of expectations can result in both governments and users feeling let down when desired features or support remains unmet.

Interestingly, we found that governments and users had some informational needs that might be more effectively met by 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.

Enhancing the auditing process for LLM-driven health chatbots is critical to ensure clear communication to all stakeholders regarding these chatbots' capabilities and limitations. Designing resources that transparently communicate the capabilities and limitations of open-domain and task-oriented chatbots could help different stakeholders identify technology that best suits their needs. In addition, engaging different stakeholders in the conversation prior to the development or deployment of an LLM-driven chatbot for public health could reveal potential conflicts and misconceptions, allowing LLM-driven chatbots to be more responsive to different stakeholder needs.

4.3 Enhancing Chatbots' Sensitivity in Health Domains Through Auditing

Our findings show that LTM significantly improved users' impressions of chatbots by offering familiarity. While those who used CareCall without LTM expressed frustration when the agent was unable to acknowledge their health history mentioned in previous sessions, those who used CareCall with LTM found the chatbot personal and emotionally supportive, frequently conveying excitement and gratitude. Our findings demonstrate the potential of LTM in mitigating the impersonal nature of technology by providing empathetic interactions, which could have a significant impact on how users engage with and perceive chatbots. Empathetic interactions through LTM could be particularly beneficial for supporting individuals who are going through difficult health experiences in the context of public health monitoring.

However, careful considerations are needed when designing LTM for complex health contexts. Through this study, we observed some challenges in repeatedly following up on chronic health conditions that are unlikely to improve (*e.g.*, chronic pain, tooth loss), leading users to perceive the chatbot as inattentive or inconsiderate. However, this frequency might be necessary for public health workers who are in charge of tracking the individuals' health conditions over time. Although remembering information about chronic health conditions is valuable for public health monitoring, designers need to carefully curate and audit LTM-triggered questions concerning such issues. Our study highlights that how stored information in LTM is referenced back to users is critical in the context of public health monitoring, which requires sustained engagement from the population to develop an understanding of their health and well-being. In the auditing process, careful attention should be given to the capability of LLM-driven chatbots to generate more thoughtful follow-up questions on sensitive health topics. The auditing should aim to strike a balance between providing empathetic interactions for individuals and meeting public health workers' needs for tracking specific health conditions.

4.4 Balancing Public Health Utility and Privacy Sensitivity in Memory through Auditing

While our study demonstrated that LTM can successfully encourage engagement and disclosure from users, some users raised privacy concerns as some LTM-triggered questions became overly specific about sensitive health topics. Compared to personal health contexts, users' privacy concerns might be exacerbated in public health monitoring as the collection of sensitive personal health information is typically aimed at achieving public health goals rather than personal benefits. To address privacy concerns, past work has primarily focused on providing users with better control over conversational agents' memory, such as whether and what data they want the agents to *store* [13, 21, 26]. While these measures could help address users' privacy concerns, unlike in personal use scenarios, it could be challenging to implement some of these measures in the context of public health interventions. For example, giving users the ability to keep chatbots from storing their past conversations or to clear their history could lead to losing important health information, potentially affecting public health authorities' ability to provide necessary interventions. For instance, in our study context, if CareCall users had serious health problems that warranted clinical care but opted to delete their conversation history because of their privacy concerns, public health authorities could miss an opportunity to send emergency responders or formal caregivers for support. This finding highlights the need for careful auditing of LTM in LLM-driven chatbots, balancing users' privacy needs and public health monitoring goals.

Prior work on chatbots with memory showed how chatbots that *reference* past conversations significantly impact users' privacy perceptions, suggesting that verbatim or paraphrased references can raise privacy concerns, whereas non-explicit references do not [5]. In our study, although the users had consented that CareCall collecting their health information for public health monitoring and research before they started using the system, some still had privacy concerns as LTM-triggered exchanges made it more apparent to the users that the chatbot was collecting their health information through conversations. In traditional settings, a typical public health worker might observe that a topic appears sensitive and may drop the topic altogether or develop alternative ways of asking related questions [8]; however, we suspect that chatbots lack such foresight. More careful auditing is needed to ensure that public health chatbots are using reference formats that can mitigate users' privacy concerns about sensitive health topics.

ACKNOWLEDGMENTS

This work was supported through a research internship at NAVER AI Lab and by the National Science Foundation under Award IIS-2237389.

REFERENCES

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

[2] 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

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

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

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

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

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

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

[9] 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* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 18, 16 pages. https://doi.org/10.1145/3544548.3581503

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

[11] 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

[12] Taewan Kim, Seolyeong Bae, Hyun Ah Kim, Su woo Lee, Hwajung Hong, Chanmo Yang, and Young-Ho Kim. 2024. MindfulDiary: Harnessing Large Language Model to Support Psychiatric Patients' Journaling. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3613904.3642937

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

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

[15] 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

[16] 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

[17] 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

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

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

[20] OpenAI, Inc. 2024. Memory and new controls for ChatGPT. Retrieved Feb 28, 2024 from https://openai.com/blog/memory-and-new-controls-for-chatgpt

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

[22] 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

[23] Woosuk Seo, Chanmo Yang, and Young-Ho Kim. 2024. ChaCha: Leveraging Large Language Models to Prompt Children to Share Their Emotions about Personal Events. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3613904.3642152

[24] Donghoon Shin, Gary Hsieh, and Young-Ho Kim. 2023. PlanFitting: Tailoring Personalized Exercise Plans with Large Language Models. arXiv:2309.12555 [cs.HC]

[25] 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

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

[27] 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.

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

[29] 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

[30] 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

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

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

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

[34] 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

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