ChatGPT for Us: Preserving Data Privacy in ChatGPT via Dialogue Text Ambiguation to Expand Mental Health Care Delivery

A. Ovalle, M. Beikzadeh, S. Fazeli, P. Teimouri, K.W. Chang, and M. Sarrafzadeh

***Abstract*— Large language models have been useful in expanding mental health care delivery. ChatGPT, in particular, has gained popularity for its ability to generate human-like dialogue. However, industries with data-sensitive regulation, such as healthcare, face challenges in using ChatGPT due to privacy and data-ownership concerns. To enable its utilization, we propose a text ambiguation framework that preserves user privacy. We ground this in the task of addressing stress prompted by user-provided texts to demonstrate the viability and helpfulness of privacy-preserved generations and find that recommendations are able to be moderately helpful and relevant, even if original user text is not used.**

***Clinical Relevance*— This establishes a mechanism for how to use ChatGPT in data-sensitive health applications while preserving data-privacy.**

I. Introduction

Language technologies have proven useful in improving mental health outcomes according to scholarly literature [1], [2]. ChatGPT has disrupted several domains with its human-like dialogue capabilities, but the health domain requires privacy-preserving techniques. Therefore, this work proposes a text ambiguation framework to enable the use of ChatGPT in data-sensitive domains, focusing on reducing stress related to economic instability[1].

II. Methods

We propose an interactive and privacy-preserving ambiguation framework for text-based recommendations. The framework takes user input (e.g. texts, mobile diary) to populate a masked query with relevant details before eliciting a recommendation from ChatGPT. Illustrated in Figure 1, ChatGPT is provided a masked dialogue question (MDQ), filled by inferred subject matter for the context and therefore - most importantly- does **not** directly pass any user data. Moving forward, P and NP describe MDQs filled with either the inferred social context or original user text, respectively. In order to test how well our ambiguation framework produces recommendations, we use a subset of the Dreaddit dataset[2] consisting of posts describing economic instability, food insecurity, and housing insecurity (N=110). Each data category is used to fill the MDQ's context. We evaluate each recommendation across a counselor trainee rubric[3]. Accordingly, we assess positive relationship building, relevance, practicality, and overall perceived helpfulness of the NP responses (Likert Scale, 1-5). We also assess text similarity across NP and P recommendations.

All authors are with UCLA. {anaelia,sfazeli,kwchang,majid}@cs.ucla.edu, {mehrabbzapril,parshanteimouri}@gmail.com.

[1]https://health.gov/healthypeople/priority-areas/social-determinants-health

[2]http://www.cs.columbia.edu/ eturcan/data/dreaddit.zip

[3]https://www.utoledo.edu/hhs/counselor-education

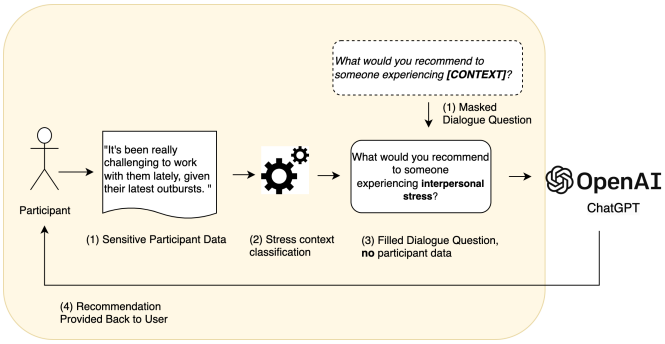

Fig. 1. Our framework for utilizing ChatGPT while preserving user-privacy.

III. Results & Discussion

We measured cosine similarity after calculating TF-IDF on P versus NP responses and found an average score of 0.25, indicating some similarity between responses. Upon review, we observed higher levels of positive relationship building and nuance in NP responses. For instance, NP responses included language that validated and reflected feelings back to the user, as well as expressed more nuanced empathetic statements (e.g. P: "I'm sorry you're experiencing stress from housing instability", NP:"It must be really hard to experience this."). As expected, NP messages are organically more expressive than P messages. We assessed our examples among 3 author annotators and discovered that ChatGPT responses contained pertinent information related to the original Reddit post, even though the original message was not given to ChatGPT (mean Krippendorf α=0.30). The scores for positive relationship building, relevance, practicality, and helpfulness averaged to 3.78, 2.69, 2.52, and 2.41. Despite differences in P vs NP inputs, our findings suggest that ChatGPT may still be helpful even when not provided the original user text. We plan to expand how the context is inferred, although this serves as a good starting point to determine the helpfulness of NP recommendations. We acknowledge that our findings are task-specific and encourage future work across several data-sensitive domains. Nonetheless, this work provides direction in navigating ChatGPT operationalization barriers for mental health applications.

References

[1] S. D'alfonso, O. Santesteban-Echarri, S. Rice, G. Wadley, R. Lederman, C. Miles, J. Gleeson, and M. Alvarez-Jimenez, "Artificial intelligence-assisted online social therapy for youth mental health," *Frontiers in psychology*, vol. 8, p. 796, 2017.

[2] A. Ovalle, O. Goldstein, M. Kachuee, E. S. Wu, C. Hong, I. W. Holloway, and M. Sarrafzadeh, "Leveraging social media activity and machine learning for hiv and substance abuse risk assessment: development and validation study," *Journal of Medical Internet Research*, vol. 23, no. 4, p. e22042, 2021.