Evaluating and Auditing LLM-Driven Chatbots for Psychiatric Patients in Clinical Mental Health Settings

Taewan Kim*
KAIST
Republic of Korea
taewan@kaist.ac.kr

Seolyeong Bae†
GIST
Republic of Korea
peixueying@gmail.com

Hyun Ah Kim
NAVER Cloud
Republic of Korea
hyunah.kim@navercorp.com

Su-woo Lee
Wonkwang Univ. Hospital
Republic of Korea
aiesw@naver.com

Hwajung Hong
KAIST
Republic of Korea
hwajung@kaist.ac.kr

Chanmo Yang
Wonkwang Univ. Hospital
Republic of Korea
ychanmo@wku.ac.kr

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

ABSTRACT

Large Language Models (LLMs) present promising opportunities within mental health domains, yet integrating LLM-infused systems in clinical settings still raises concerns regarding LLM's safety and controllability. In this encore paper, we reflect on our CHI 2024 paper on MindfulDiary, an LLM-driven conversational diary app that utilizes LLMs to help psychiatric patients document their daily experiences through conversation. Developed in collaboration with mental health professionals, the app was deployed in a four-week field study with 28 patients diagnosed with major depressive disorder and five psychiatrists. We particularly expand on the process of preliminary testing and auditing MindfulDiary's LLM-driven chatbot before and during its deployment to individuals with mental health issues, offering insights into integrating LLM-driven chatbots in clinical mental health settings.

1 INTRODUCTION

Recent progress in Natural Language Processing through large language models (LLMs) has created new possibilities for developing chatbots capable of engaging in more natural conversations [2, 3, 8, 17, 25]. Particularly, the advantages of LLM-driven chatbots, such as their conversational flexibility and adaptability [16, 21], have become notable in challenging areas like mental health support. For example, GPT-3.5 has shown empathetic characteristics by recognizing emotions and providing emotionally supportive responses, especially in healthcare environments [20]. These models sometimes exhibit capabilities in empathy-related tasks, demonstrating encouraging outcomes in comparison with human performance [1, 5, 19].

Despite their opportunities, the inherent unpredictability in LLMs' text generation necessitates caution to mitigate the risk of unintended or inaccurate replies [6, 8, 10, 24]. For example, a mental health therapy chatbot using GPT-2 sometimes created nonsensical sentences or tended to generate more negative than positive responses [24]. Similarly, Replika, an app powered by LLMs for enhancing mental well-being, has been known to produce inappropriate content and show problematic conversational patterns [12]. Thus, for the effective use of LLM-driven chatbots in clinical and mental health settings, it's crucial that their development involves collaboration with domain experts to ensure the relevance and safety of their interactions.

In this paper, we share a case study of MindfulDiary [9], focusing on the evaluation measures during the development. MindfulDiary is a mobile application designed to assist psychiatric patients by documenting their daily experiences in a clinical mental health setting through conversations with a chatbot driven by LLMs. To conduct a four-week field deployment study with psychiatric patients as participants, we engaged in an iterative design process. This process encompassed various approaches to ensure the safety of the LLM's output throughout its development, including a prompt workshop with mental health professionals (MHPs), the generation of synthetic dialogue data from evidence-based patient personas, and a pilot study utilizing a Human-in-the-Loop (HITL) approach. Reflecting on these measures, we provide insights into integrating MHPs' perspectives into the development and evaluation of LLM-driven applications in the mental health context.

2 BACKGROUND: MINDFULDIARY

2.1 Motivation of MindfulDiary

Journals act as a documented reflection of an individual's previous experiences, emotions, and thoughts, facilitating authentic expression [22, 23]. Research has highlighted the benefits of journaling within clinical settings, where journals often record patients' daily lives, symptoms, and additional relevant information that is difficult to obtain during short clinical visits [7, 27]. Nonetheless, reflecting on past emotions and thoughts through writing can be intricate, as individuals vary in their capacity to recognize, comprehend,

*Taewan Kim conducted this work as a research intern at NAVER AI Lab.
†Seolyeong Bae conducted this work as an engineering intern at NAVER Cloud.

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

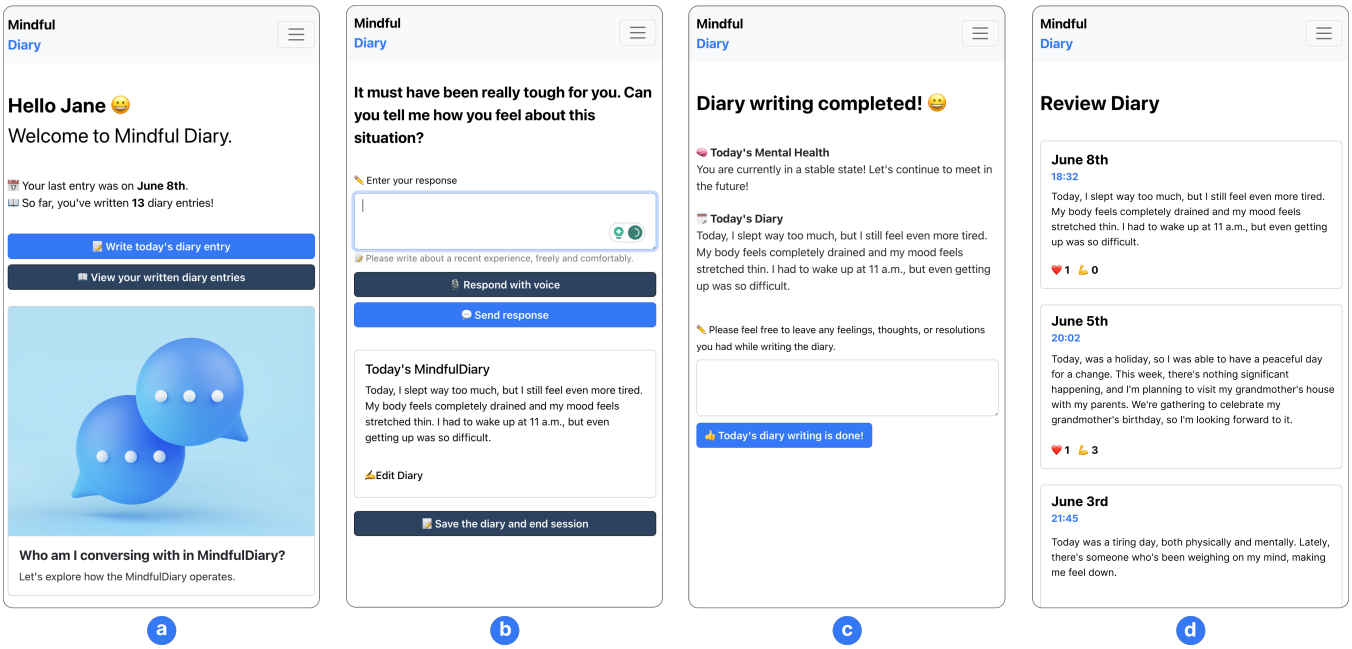

Figure 1: Main screens of the MindfulDiary app. (a) The main screen, (b) the journaling screen, (c) the summary screen shown when the user submitted the journal dialogue, and (d) the review screen displaying the user's past journal.

and articulate their feelings [18]. Furthermore, patients undergoing psychotherapy face challenges in creating narratives and comprehending their experience [4, 15].

To address these challenges, we present a case of collaborative design, development, and evaluation of an LLM-infused conversational AI system. **MindfulDiary** [9] is designed to facilitate the self-reflection of patients and communication with MHPs. The MindfulDiary system consists of (1) a mobile conversational AI with which patients can converse about their daily experiences and thoughts and (2) a web dashboard that allows MHPs to review their patients' dialogues with the LLM. In this paper, we focus on the development of the patient interface, where the LLM interacts with the patient.

2.2 LLM-driven Chatbot in MindfulDiary

The primary function of MindfulDiary is to engage users in dialogues that assist them in documenting their daily experiences. To achieve this, the chatbot within MindfulDiary is designed to generate context-sensitive prompts that encourage users to share more comprehensive details about their activities, feelings, and emotions (Figure 1). The conversational model of the chatbot was informed by findings from psychiatric research [13, 14], encompassing essential methods for conducting clinical interviews. Additionally, we enriched the chatbot's design with the practical insights gained from focus-group discussions with practicing psychiatrists.

Consequently, we structured the dialogue into three distinct phases: *Rapport Building*, *Exploration*, and *Wrap-up*. The **Rapport Building** phase serves as an introductory ice-breaker, focusing on light-hearted conversations about the user's day. During this phase, the assistant shares small pieces of information to promote user openness, drawing on evidence that chatbot self-disclosure fosters greater user engagement [11]. The primary aim here is to cultivate a comfortable atmosphere for users to share their experiences. Moving into the **Exploration** phase, the focus broadens to a deeper dive into the user's daily life, emotions, and thoughts, using a blend of open and closed questions to keep the user involved in the conversation. Open questions are designed to encourage a more expressive sharing of feelings and perspectives, whereas closed questions aim to gather precise details about the user's experiences [13, 14]. The dialogue then shifts to the **Wrap-up** phase, which aims to conclude the session while ensuring the user has thoroughly expressed their experiences. In addition to these main phases, we introduced a **Sensitive Topic** phase designed to address highly delicate issues, such as self-harm or thoughts of suicide. Activation of this phase triggers immediate alerts to psychiatrists.

It is recognized that lengthy and complex input prompts can lead to reduced task performance in LLMs [3], as they may neglect essential underlying concepts [26]. To ensure the LLM adheres to our specific conversational design, we constructed the dialogue system of MindfulDiary as a state machine. This approach involves using distinct, concise, and clear input prompts for each conversation phase rather than a single, comprehensive prompt that encompasses instructions for all stages (see Figure 2).

3 APPROACHES TO EVALUATING MINDFULDIARY'S LLM-DRIVEN CHATBOT

In the development of MindfulDiary, we designed and conducted diverse evaluation measures in collaboration with MHPs before individuals with mental health issues directly interacted with MindfulDiary. Especially for LLMs, given their inherent limitations in predicting and controlling the behavior and output of the LLM, it

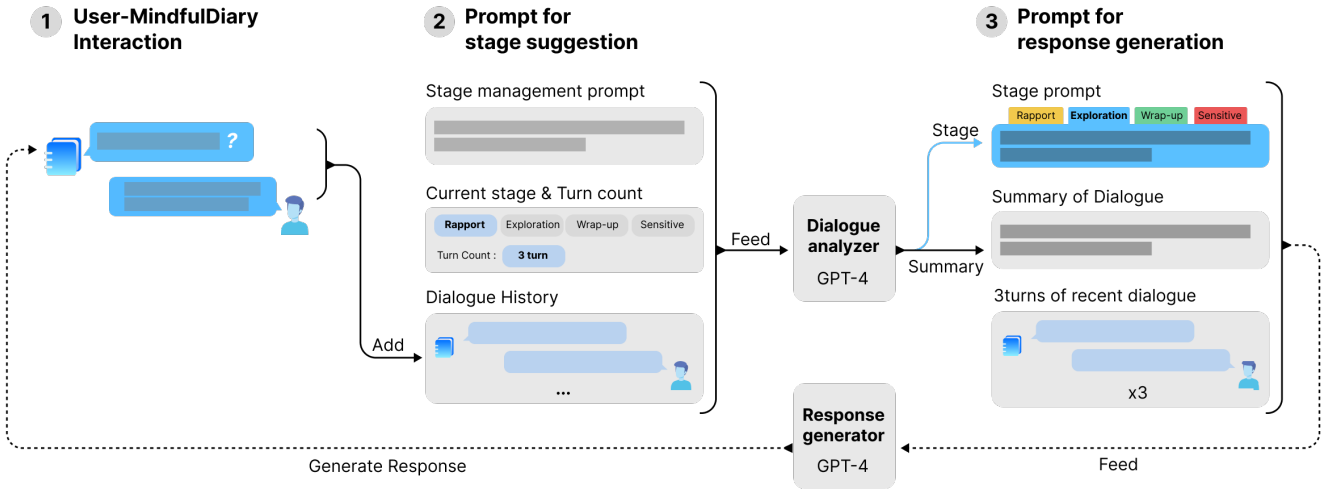

Figure 2: Structure of MindfulDiary's conversational pipeline. (1) Users respond to MindfulDiary's messages. (2) The recent turn count, current state, and whole user-MindfulDiary dialogue are fed into Dialogue Analyzer. (3) Using the output of Dialogue Analyzer, a designated state prompt, a summary of dialogue (containing overall dialogue context), and the latest three conversation turns are fed into Response Generator. The resulting response is then displayed to the user. Both the Dialogue Analyzer and Response Generator operate based on the GPT-4 LLM.

was crucial to explore how we could systematically and effectively simulate the patients' utterances and evaluate the model's outputs and patterns. In the following sections, we detail the approach taken to evaluate the MindfulDiary system.

3.1 Prompt Design Workshop

As a first step to incorporate expert evaluation and review throughout the system's development, we conducted a review of the terms and expressions contained within the prompts for LLMs. For this purpose, we held prompt workshops with MHPs. Understanding that the appropriateness of the LLM behavior hinges on the suitability of these prompts, we engaged MHPs to utilize their expertise. They were briefed on the importance of prompts for LLM behavior and the fundamentals of prompt engineering, after which they reviewed and provided feedback on the early version of prompts.

Their expertise was crucial in assessing the prompts' suitability for clinical settings, offering a platform to evaluate terminology and strategies effectively. Specifically, they provided feedback from a procedural perspective. For example, they highlighted the importance of offering messages of encouragement before the chatbot presented its first question, acknowledging the user's decision to engage with the app. They also provided feedback on enhancing the instructions with examples from their experience. For instance, they suggested adding more descriptive sentences to the term 'greetings' in the existing prompts with specific question samples, such as "Have you eaten lunch?" or "How was the weather today?", to start the conversation. They mentioned that providing such specific and easy-to-answer questions allows patients to begin sharing their stories easily.

3.2 Interaction-Based Evaluation with MHPs

After incorporating feedback from MHPs into the prompts, we proceeded with the evaluation of MindfulDiary's chatbot, involving five clinical psychologists and three psychiatrists. In this evaluation, MHPs directly interacted with our LLM pipeline, simulating potential patient utterances based on their clinical experience to assess the model's responses and identify any problematic outputs and scenarios. However, during these experiments, we realized that providing one response, as if simulating a patient interface, made it challenging for MHPs to understand patterns or tendencies in the LLM's behavior due to its characteristic of generating varied responses each time. Consequently, professionals attempted to grasp the LLM's behavioral patterns by repeatedly inputting the same utterances they were focusing on. This turned out to be a time-consuming and inefficient task. To address this and facilitate a better understanding of the model's behavior, we designed a test module capable of generating five responses for the same input. This allowed professionals to more effectively observe the model's response patterns and identify problematic cases. The module also enabled experts to label sentences that the model should not generate and provide a text input field for qualitative feedback.

Through this process, we gained a detailed understanding of the scenarios that concerned MHPs and the potential utterances patients might express in such contexts. Notably, MHPs evaluated messages that contained implicit or explicit references to thoughts or actions of suicide or self-harm. The input data gathered from MHPs during this phase proved to be invaluable for our development, diverging significantly from the researcher's initial expectations in terms of content, expression intensity, and input categories. Additionally, this input data served as a test set for researchers to examine the effects of prompt revision.

Category	Details
Background	*Minji is a Korean teenager living in Seoul with her parents and younger brother. She has always been a high achiever, excelling in academics and sports, but lately, she has been struggling to keep up with her responsibilities.*
Symptoms	*Minji suffers from bipolar depression, which manifests as extreme mood swings. Sometimes she feels elated and invincible, while at other times she is overwhelmed by sadness and hopelessness. She struggles to maintain stable relationships with family and friends.*
Treatment	*Minji's parents have taken her to see a psychiatrist who has prescribed medication along with therapy sessions. While the medication helps stabilize her moods, the therapy sessions are helping her develop coping mechanisms for dealing with her depression.*

Table 1: Evidence-based patient persona named Minji created via GPT-4

3.3 Evaluation Using Synthetic Patient Dialogue

Evaluation through direct interaction with the LLM pipeline was valuable for understanding the model's behavior by simulating specific scenarios that MHPs aimed to explore. However, the process of creating conversations one by one with a limited number of experts was insufficient to generate a comprehensive number of cases. Additionally, there were limitations in reviewing dialogues at a session level rather than at a sentence level.

To overcome this challenge, we generated synthetic dialogues using LLMs and evidence-based patient personas. This step involved close collaboration with psychiatrists to ensure that the created personas accurately reflected real clinical experiences. These patient personas were designed with three primary symptoms (bipolar disorder, self-harm, and strong suicidal impulses) across three age categories (adolescent, adult, elderly), resulting in a framework of nine distinct characteristics. Initially, we used the LLM to create synthetic personas, complete with patient names, ages, backgrounds, symptoms, and treatment information, to ensure a diverse range of scenarios. Table 1 provides an example of such a persona.

Subsequently, we generated dialogues between virtual patients, driven by the LLM and our conversational pipeline. We sent these synthetic dialogues to MHPs in the form of spreadsheets, allowing them to provide feedback on the entire session and specific lines where issues were identified. This approach enabled us to gather feedback on a significantly larger number of conversations compared to earlier methods.

3.4 Pilot Test Under the Oversight of MHPs

After reviewing previous evaluation processes, we conducted our initial pilot test with patients using MindfulDiary. However, we anticipated the risk of patients, encountering the system for the first time, might use the app in unintended ways under uncontrolled situations. Therefore, it was necessary first to understand how real patients would interact with the system and observe any characteristic patterns and utterances under the supervision and control of experts. To this end, we developed a monitoring module and protocol that allowed MHPs to intervene in the interactions between the patient and MindfulDiary, pre-screening and approving the LLM outputs in real-time before they were delivered to the patient. The pilot study involved five patients who were currently admitted to a university hospital but were about to be discharged soon.

The psychiatrists could preview and validate the generated output from the LLM before it was delivered to the patient, thereby ensuring safeguards against potential hazards. Additionally, during this user evaluation process, experts could label problematic outputs for improvement, similar to the expert evaluation process. Through this verification with patients, we were able to identify behavioral characteristics that had not been expected before. For example, the speed at which patients interacted with the system was significantly slower than we had anticipated, and in some cases, conversations ended without completing key stages like 'exploration.' We also discovered a pattern of behavior where patients would provide very long responses to a single question, leading to excessively lengthy LLM outputs that made it difficult to generate understandable follow-up prompt questions. No critical scenarios related to patient safety were identified. This verification process served as an important checkpoint in assessing whether the app was ready to be deployed to patients in the real world.

4 REFLECTION ON MINDFULDIARY DEVELOPMENT

During the development of MindfulDiary, we were able to learn important lessons about evaluating LLM-driven chatbots within the context of mental health. We realized the importance of contemplating how to reflect the knowledge and expertise of MHPs from various perspectives throughout the development process. The evaluation method of MindfulDiary, including the prompt workshop, the interaction processes between the LLM and experts simulating patients, and the generation of synthetic patients based on expert knowledge, were all conducted by the same professionals. However, these activities provided a unique environment that allowed them to evaluate the system from multiple angles in terms of safety and relevance to mental health. Notably, HCI researchers played a crucial role in designing and implementing protocols and systems that effectively incorporate the domain expertise of professionals from various perspectives.

We also recognized the need to gradually expand the scope and content of evaluation to provide safe LLM output for individuals with mental health issues. We underwent four stages of expert-involved evaluation before reaching the stage where potential users interacted with the LLM. Although not covered in this paper, a real-time monitoring protocol and system for expert intervention in problematic situations was also developed and operated during

the deployment. Particularly noteworthy is that during the initial evaluation phases, professionals simulated patient interactions to predict and model potential user inputs. However, as we progressed to direct interactions with patients, we encountered unforeseen user inputs and corresponding responses from the LLM. Therefore, it may be important to gradually expand the scope and content of the evaluation to enable researchers to anticipate and respond to these unexpected inputs and scenarios.

In sum, in this paper, we introduced methods for evaluating LLM-driven chatbot systems targeting individuals with mental health issues. We hope for constructive and in-depth discussions on the use of LLMs for vulnerable groups that require careful consideration within the HCI community.

REFERENCES

[1] John W Ayers, Adam Poliak, Mark Dredze, Eric C Leas, Zechariah Zhu, Jessica B Kelley, Dennis J Faix, Aaron M Goodman, Christopher A Longhurst, Michael Hogarth, et al. 2023. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA internal medicine* (2023).

[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. *arXiv preprint arXiv:2205.00176* (2022).

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[4] Daniel A Donnelly and Edward J Murray. 1991. Cognitive and emotional changes in written essays and therapy interviews. *Journal of Social and Clinical psychology* 10, 3 (1991), 334–350.

[5] Zohar Elyoseph, Dorit Hadar-Shoval, Kfir Asraf, and Maya Lvovsky. 2023. ChatGPT outperforms humans in emotional awareness evaluations. *Frontiers in Psychology* 14 (2023), 1199058.

[6] Faiza Farhat. 2023. ChatGPT as a complementary mental health resource: a boon or a bane. *Annals of Biomedical Engineering* (2023), 1–4.

[7] Mayara Costa Figueiredo, Yunan Chen, et al. 2020. Patient-generated health data: dimensions, challenges, and open questions. *Foundations and Trends® in Human–Computer Interaction* 13, 3 (2020), 165–297.

[8] 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* (<conf-loc>, <city>Hamburg</city>, <country>Germany</country>, </conf-loc>) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 18, 16 pages. https://doi.org/10.1145/3544548.3581503

[9] 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

[10] 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, 1 (2021), 93. https://doi.org/10.1038/s41746-021-00464-x

[11] 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. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW1, Article 31 (may 2020), 27 pages. https://doi.org/10.1145/3392836

[12] Zilin Ma, Yiyang Mei, and Zhaoyuan Su. 2023. Understanding the Benefits and Challenges of Using Large Language Model-based Conversational Agents for Mental Well-being Support. *arXiv* (2023). https://doi.org/10.48550/arxiv.2307.15810 arXiv:2307.15810

[13] Ekkehard Othmer and Sieglinde C Othmer. 2002. *The clinical interview using DSM-IV-TR: Vol 1: Fundamentals.* American Psychiatric Publishing, Inc.

[14] Ekkehard Othmer and Sieglinde C Othmer. 2002. *The clinical interview using DSM-IV-TR: Vol. 2: The difficult patient.* American Psychiatric Publishing, Inc.

[15] James W Pennebaker and Janel D Seagal. 1999. Forming a story: The health benefits of narrative. *Journal of clinical psychology* 55, 10 (1999), 1243–1254.

[16] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.

[17] 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

[18] Peter Salovey and John D Mayer. 1990. Emotional intelligence. *Imagination, cognition and personality* 9, 3 (1990), 185–211.

[19] Ashish Sharma, Inna W Lin, Adam S Miner, David C Atkins, and Tim Althoff. 2023. Human–AI collaboration enables more empathic conversations in text-based peer-to-peer mental health support. *Nature Machine Intelligence* 5, 1 (2023), 46–57.

[20] Vera Sorin, Danna Brin, Yiftach Barash, Eli Konen, Alexander Charney, Girish Nadkarni, and Eyal Klang. 2023. Large language models (llms) and empathy-a systematic review. *medRxiv* (2023), 2023–08.

[21] Michael J. Tanana, Christina S. Soma, Patty B. Kuo, Nicolas M. Bertagnolli, Aaron Dembe, Brian T. Pace, Vivek Srikumar, David C. Atkins, and Zac E. Imel. 2021. How do you feel? Using natural language processing to automatically rate emotion in psychotherapy. *Behavior Research Methods* 53, 5 (2021), 2069–2082. https://doi.org/10.3758/s13428-020-01531-z

[22] Cheryl Travers. 2011. Unveiling a reflective diary methodology for exploring the lived experiences of stress and coping. *Journal of Vocational Behavior* 79, 1 (2011), 204–216. https://doi.org/10.1016/j.jvb.2010.11.007

[23] Philip M. Ullrich and Susan K. Lutgendorf. 2002. Journaling about stressful events: Effects of cognitive processing and emotional expression. *Annals of Behavioral Medicine* 24, 3 (2002), 244–250. https://doi.org/10.1207/s15324796abm2403_10

[24] 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. *arXiv* (2021). https://doi.org/10.48550/arxiv.2107.13115 arXiv:2107.13115

[25] 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

[26] Tongshuang Wu, Michael Terry, and Carrie Jun Cai. 2022. AI Chains: Transparent and Controllable Human-AI Interaction by Chaining Large Language Model Prompts. 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 385, 22 pages. https://doi.org/10.1145/3491102.3517582

[27] Haining Zhu, Joanna Colgan, Madhu Reddy, and Eun Kyoung Choe. 2016. Sharing patient-generated data in clinical practices: an interview study. In *AMIA Annual Symposium Proceedings*, Vol. 2016. American Medical Informatics Association, 1303.