

# Private Yet Social: How LLM Chatbots Support and Challenge Eating Disorder Recovery

Ryuhaerang Choi  
 School of Electrical Engineering  
 KAIST  
 Daejeon, Republic of Korea  
 ryuhaerang.choi@kaist.ac.kr

Taehan Kim  
 School of Computing  
 KAIST  
 Daejeon, Republic of Korea  
 tanit23@kaist.ac.kr

Subin Park  
 School of Electrical Engineering  
 KAIST  
 Daejeon, Republic of Korea  
 subin.park@kaist.ac.kr

Jennifer G. Kim  
 School of Interactive Computing  
 Georgia Institute of Technology  
 Atlanta, Georgia, USA  
 jennifer.kim@cc.gatech.edu

Sung-Ju Lee  
 School of Electrical Engineering  
 KAIST  
 Daejeon, Republic of Korea  
 profsj@kaist.ac.kr

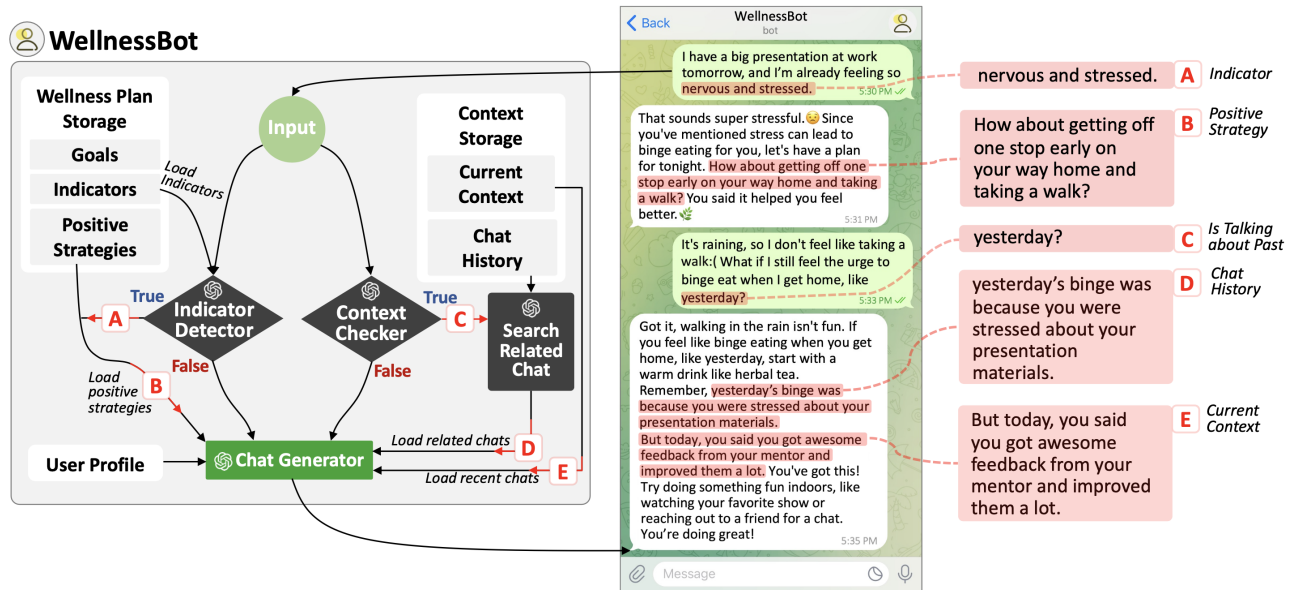


Figure 1: Overview of our technology probe, *WellnessBot*, an LLM chatbot for people with eating disorders. When a user sends a message, the Indicator Detector checks for any indicators that support is needed. (A) If indicators are found, (B) *WellnessBot* retrieves positive strategies to support the user. Meanwhile, the Context Checker (C) identifies references to past experiences and (D) loads relevant conversations from the chat history. *WellnessBot* then generates a response based on all the gathered information, (E) including the recent chats to get the current context and the pseudonymized user profile.

## Abstract

Eating disorders (ED) are complex mental health conditions that require long-term management and support. Recent advancements in large language model (LLM)-based chatbots offer the potential to assist individuals in receiving immediate support. Yet, concerns

remain about their reliability and safety in sensitive contexts such as ED. We explore the opportunities and potential harms of using LLM-based chatbots for ED recovery. We observe the interactions between 26 participants with ED and an LLM-based chatbot, *WellnessBot*, designed to support ED recovery, over 10 days. We discovered that our participants have felt empowered in recovery by discussing ED-related stories with the chatbot, which served as a *personal yet social avenue*. However, we also identified harmful chatbot responses, especially concerning individuals with ED, that went unnoticed partly due to participants' unquestioning trust in the chatbot's reliability. Based on these findings, we provide design



implications for safe and effective LLM-based interventions in ED management.

## CCS Concepts

• **Human-centered computing** → **Natural language interface design; Empirical studies in HCI.**

## Keywords

chatbots, large language models, conversational agents, eating disorder, mental health

### ACM Reference Format:

Ryuhaerang Choi, Taehan Kim, Subin Park, Jennifer G. Kim, and Sung-Ju Lee. 2025. Private Yet Social: How LLM Chatbots Support and Challenge Eating Disorder Recovery. In *CHI Conference on Human Factors in Computing Systems (CHI '25)*, April 26–May 01, 2025, Yokohama, Japan. ACM, New York, NY, USA, 19 pages. <https://doi.org/10.1145/3706598.3713485>

## 1 Introduction

**Caution:** *This paper discusses eating disorders and contains content that could potentially be a trigger to those dealing with eating disorders. It also includes references to depression and suicide, which may be distressing. Please use discretion when reading and disseminating this paper.*

The global prevalence of eating disorders (ED) has surged from 3.5% to 7.8% between 2000 and 2018 [48]. In addition, 22.4% of children and adolescents worldwide exhibit disordered eating behaviors, with these behaviors becoming more pronounced with age [84]. This alarming trend has garnered significant attention within the field of Human-Computer Interaction (HCI), particularly regarding the impact of technology on individuals with eating disorders [36, 43, 47, 106]. Research has also proposed various digital interventions to support their daily lives [15, 17, 20].

Among various digital interventions for individuals with ED, chatbots have shown promise as digital interventions for them, providing immediate support in out-of-clinic settings where access to clinical and social support is often lacking [6, 14, 44, 119]. With the advancements in Large Language Models (LLMs), there has been a surge in efforts within the HCI and healthcare communities to harness their capabilities for more sophisticated and effective health-support chatbots [23, 80, 81, 87, 100, 119, 139]. This line of research indicates that LLMs can understand health contexts and provide personalized, contextual, and effective interventions, unlike traditional chatbots that are limited in comprehending and addressing individuals' specific needs [119].

However, recent studies highlighted a complex landscape of opportunities and risks associated with LLM-based chatbots in mental health support [3, 89, 90, 120]. On one hand, LLM-based chatbots have been proposed as a promising tool for providing emotional assistance and coping strategies for those with mental health issues. On the other hand, there are significant concerns regarding the potential for LLM-based chatbots to offer misinformation and harmful advice, as their outputs are less constrained than those of rule-based chatbots. For instance, in 2023, the National Eating Disorder Association (NEDA) had to shut down its support chatbot after it provided users with harmful recommendations, including weight

loss and dieting advice [25]. Such incidents underscore the real-life and sometimes lethal consequences of relying on LLM-based chatbots for mental health support.

Despite these risks, individuals with mental health issues as well as people with ED, are increasingly turning to LLM-based chatbots for support [114]. This trend highlights the need for HCI researchers and clinicians to explore how to safely exploit LLMs while mitigating the risks of inappropriate responses. Although some research examined LLM-based chatbot usage by the general population and certain vulnerable groups [3, 89, 90, 120], the unique and nuanced symptoms, contexts, and needs of individuals with ED make it crucial to investigate the specific opportunities and challenges of LLM usage within this population. For instance, while promoting healthy eating is generally beneficial, including for individuals with ED, it is equally critical to prevent individuals with ED from becoming overly fixated on 'healthy eating,' especially given the marked perfectionistic tendencies common in ED, which are linked to poorer treatment outcomes [4, 121]. Striking this balance is essential to avoid rigid, obsessive behaviors around food that could exacerbate their condition instead of aiding recovery.

Considering the intricate nature of ED contexts distinct from the general population, we aim to comprehensively investigate the use of LLM-based chatbots in everyday eating disorder care. We developed *WellnessBot*, an LLM-based chatbot designed to assist users in managing their ED and to uncover both the opportunities and potential harms of such technology. *WellnessBot* was created as a *technology probe* (TP) to explore the experience of interacting with LLM-based chatbots in a real-world setting [63]. We conducted a user study with 26 individuals struggling with ED, using our TP, *WellnessBot*, follow-up surveys, and interviews. Over the ten-day period, we gathered data on how participants engaged with *WellnessBot* and their perceptions and experiences.

Our extensive analysis of chat logs, interview transcripts, and open-ended survey responses showed that participants valued *WellnessBot* as a space, offering a unique combination of privacy and interaction, to openly discuss their ED-related experiences while still feeling accompanied by a supportive presence. However, we also identified the fallibility of chatbot responses, which were inappropriate for individuals with ED, yet none of the participants questioned these responses. One common reason for this lack of critical assessment was their strong trust in *WellnessBot*, based on incomplete knowledge of how AI-based systems function. The combination of misunderstanding how AI, specifically LLMs, operates and both the users' and chatbots' limited grasp on users' health conditions created nuanced potential harms in using LLM-based chatbots for ED care.

The main contributions of our study are summarized as follows.

- *WellnessBot*, an LLM-based chatbot technology probe designed to support eating disorder management in everyday, out-of-clinic settings. Our design contributions lie in adapting chatbot functionalities, established in prior literature, specifically to support ED recovery.
- Findings from a ten-day field deployment study with 26 participants with ED using the TP approach, where we collected qualitative and quantitative data on how people adopt and

engage with an LLM-based chatbot for ED care as a stand-alone digital intervention.

- Future opportunities and design considerations for exploiting LLMs to support individuals with ED and other mental health issues in out-of-clinic settings based on our analysis of participants' experiences.

Through this study, we discuss design implications for LLM-based tools that encourage users to share their recovery journey for empowerment while avoiding guidance that could be problematic. Additionally, we explore opportunities for human-LLM collaboration in mental health intervention. We suggest strategies to encourage critical thinking in both users and chatbots, offering a second chance to correct potential issues during interactions.

## 2 Related Work

### 2.1 Everyday Challenges in Eating Disorder Care

People with eating disorders face a wide range of daily struggles that affect both their physical and mental well-being [108]. These challenges include an unhealthy preoccupation with food, body image, and weight, leading to patterns of restrictive eating, binge eating, or purging. These behaviors can severely impact an individual's ability to function in everyday life [34], affecting relationships [58], work [107], and social interactions [105]. In addition, the psychological burden of ED, such as feelings of shame, depression, and anxiety, can further complicate their efforts to lead a balanced life [34].

Managing EDs is particularly challenging due to the chronic and prolonged nature of these disorders [49], with about a third of patients continuing to meet diagnostic criteria five or more years after initial treatment [41]. Many individuals require long-term treatment and support, but continuous access to professional care from clinicians is difficult to maintain over extended periods [136]. The complex and persistent nature of ED means that even when individuals seek treatment, the journey toward recovery is often long, requiring ongoing management rather than short-term interventions. Without consistent access to such support, individuals with ED are likely to relapse, struggle with daily challenges, and experience worsening symptoms [49, 136]. In response to these challenges, there have been a few attempts to employ chatbots to assist in the daily care of individuals with ED, providing continuous support in managing their condition alongside clinical treatment [6, 44, 119].

### 2.2 Chatbots for Supporting Individuals with ED: Pre-LLM

Before the advent of LLMs, chatbots supporting individuals with ED or other mental health issues were primarily developed using three main approaches: rule-based, retrieval-based, and early machine learning (ML) model-based methods. Rule-based chatbots followed predefined rules and scripted responses [6, 44, 56, 119]. While they could offer structured support for specific queries, their responses were limited in flexibility. They struggled to address the complexity of ED-related discussions, which require nuanced understanding and adaptability [119]. On the other hand, retrieval-based chatbots matched user queries with pre-existing responses from a

database [86], allowing for more dynamic interactions than rule-based systems. However, these chatbots still could not understand or adapt to users' evolving contexts.

Early ML models, such as Bidirectional Encoder Representations from Transformers (BERT) [37] and Sequence-to-Sequence (Seq2Seq) models [124], began to address some of these shortcomings by generating responses based on the user's input rather than relying solely on pre-defined scripts or database matches.<sup>1</sup> These models generate more dynamic and varied responses, but they struggled to maintain coherence over longer conversations [133], losing context or producing emotionally insensitive responses. This inability to retain context and emotional nuance could lead to failures in providing sustained, engaging interactions [66].

Chatbots developed using these methods provided timely assistance to individuals with ED [6, 119]. They offered coping strategies, answered immediate questions, and were available 24/7, ensuring users had access to support whenever needed. Their consistency in following structured protocols made them reliable for managing straightforward aspects of ED care, such as reminders for healthy behaviors or crisis guidance. However, their overall effectiveness was limited by their inability to deeply understand and adapt to users' individual needs. This inability led to unsatisfactory experiences, as users encountered robotic, automated responses, limited response options, and insufficient content, including a lack of diverse eating disorder coverage and novel information [119]. As a result, many users felt frustrated and disengaged, particularly during critical moments when they needed more nuanced or empathetic support.

The limitations of early chatbots, particularly in handling complex interactions and offering emotionally responsive support, pointed to the need for more advanced systems. Although ML models, such as BERT and Seq2Seq, improved chatbots' abilities by generating dynamic responses based on user input and providing valuable support, they still had difficulties maintaining context in longer conversations and lacked emotional sensitivity [117, 128]. Consequently, these models could fail to offer the personalized, adaptive care necessary for managing the challenges of ED recovery. These challenges suggested the need for chatbots designed specifically for individuals with ED to offer more context-aware and emotionally intelligent interactions.

### 2.3 Chatbots for Supporting Individuals with ED: Post-LLM

With the advancement of LLMs, chatbots have gained capabilities that pre-LLM systems have lacked. LLMs allow chatbots to engage in more dynamic, contextually aware, and emotionally responsive conversations, potentially providing more nuanced and personalized support for individuals with mental health issues [23, 87, 120]. Recognizing this potential, the National Eating Disorder Association (NEDA) deployed a chatbot named Tessa to assist ED patients [25, 126]. However, despite its advanced capabilities, Tessa was shut down because it provided harmful advice to users who are struggling with ED, highlighting the risks of using LLM-based chatbots in sensitive health contexts.

This underscores the complexities and risks of deploying LLM-based chatbots in sensitive ED contexts. While LLMs can offer

<sup>1</sup>We refer to these models as relatively early ML models before advanced LLMs.

general advice on healthy habits, they may inadvertently reinforce harmful behaviors by failing to recognize the nuanced needs of ED care. For example, a chatbot encouraging self-monitoring behaviors, such as tracking food intake or exercise, is typically benign or helpful for the general population [101]. However, for individuals with ED, it could have unintended consequences as self-monitoring can intensify preoccupation with calorie counting and obsessive exercise, which are ED symptoms [54]. This highlights the need for chatbot interactions to be highly nuanced in ED contexts, ensuring they do not inadvertently trigger or worsen body image distortion.

Given these complexities and risks, it is crucial to thoroughly investigate both the opportunities and potential harms of utilizing LLM-based chatbots for individuals with ED. These LLM-based chatbots have the potential to offer immediate support in out-of-clinic settings [97]. However, without careful consideration of the unique challenges posed by ED, chatbots could inadvertently cause more harm than good. Understanding how individuals with ED interact with LLM-based chatbots is essential for designing systems that provide support without reinforcing harmful behaviors. In this study, by carefully exploring these dynamics, we aim to inform the development of more tailored and responsible LLM-based interventions for ED care.

## 2.4 Safety Risks and Safeguards for LLM-based Mental Health Chatbots

Recent studies have highlighted multifaceted safety risks associated with LLM-based mental health chatbots and proposed various safeguards to address the challenges. Critical LLM safety risks in the mental health domain include hallucination of inaccurate information [95], generating harmful or inconsistent advice [8, 25, 71], and biases in responses based on user demographics (e.g., lower empathy towards certain demographic subgroups) [46]. In addition, LLM-based chatbot designs often fall short in preventing users from inadvertently disclosing private information [137] and foster over-reliance, potentially undermining patient-therapist relationship [35]. These risks are particularly concerning in the high-stakes mental health domain.

Researchers have proposed various *engineering* strategies to preempt these risks. For example, safety engineering frameworks, such as Failure Mode and Effects Analysis (FMEA), which identifies potential failure modes and evaluates their likelihood to improve reliability [13], and System-Theoretic Process Analysis (STPA), which maps system elements and interactions to identify unsafe states [82], have been introduced. These frameworks assist designers and developers in assessing and mitigating risks during ML system development [116]. In addition, many studies have explored utilizing LLMs to evaluate and address safety risks in the outputs of other LLMs. One approach involves using an LLM for red-teaming, where an LLM generates adversarial prompts, and the target LLM responds to the prompts with safe responses [50]. Through iterative interactions between two LLMs, the target LLM is fine-tuned to align its responses with safety standards. Another example is the agent-constitution-based framework, which ensures adherence to a predefined constitution derived from established safety norms [60]. This framework employs LLM-enabled safety measures, where LLMs evaluate the safety of the other LLM's action plans at the pre-, in-,

and post-planning stages of the agent's target operations. While such engineering strategies offer foundational tools, they require careful application and adaptation to the specific requirements of each context [116].

Researchers have also explored *user interventions* to mitigate safety risks during interactions with LLM-based chatbots. For instance, recent works emphasize explaining LLMs and their outputs to educate users about system limitations, helping them understand chatbot capabilities and recognize potential inaccuracies [62, 138]. In addition, a publicly available LLM-based mental health chatbot tries to address the risk of users' over-reliance on chatbots by offering hotline resources when detecting concerning language from users [57]. Similarly, ChatGPT incorporates a 'denial' feature, which refuses to respond to prompts involving harmful and inappropriate content [103]. A recent study examined user perceptions of such denials in various contexts, including health-related prompts [131]. This study suggested providing alternative, constructive responses, rather than simply refusing to reply, to maintain positive user interactions.

Amidst the growing body of research on safety risks and safeguards for LLM-based assistants, both in general and mental health domain, it is imperative to understand the unique safety risks associated with individual health conditions and adopt context-specific approaches that address them effectively [5, 45]. Therefore, our work aims to identify the specific safety challenges of LLM-based chatbots in the ED context and provides insights into designing systems tailored to the needs of ED populations.

## 3 Ethical Considerations

Even with IRB approval, we acknowledge that this study raises important ethical considerations that require careful attention, particularly as we deploy an LLM-based chatbot, which may provide harmful advice to individuals, to collect and analyze user interactions with *WellnessBot*. To address these concerns, we have taken meticulous care to address potential issues. A primary concern is the potential for the LLM-based chatbot to generate wrong, harmful, or violent content [8, 25, 126]. To help mitigate this risk, we conducted an introductory session to inform participants about the possibility of *WellnessBot*, the LLM-based chatbot we designed and deployed, behaving unexpectedly. We ensured they were aware of the risks and encouraged them to critically evaluate *WellnessBot*'s responses during use.<sup>2</sup>

In addition, our research protocol ensured that two authors continuously monitored interactions between participants and *WellnessBot* every day during the field study. Their role was to intervene in case *WellnessBot* generates any life-threatening or self-harm-encouraging content. In other cases of misinformation or undesirable responses, such as inappropriate reactions to users expressing satisfaction from weight loss, we did not intervene immediately to realistically examine the potential impacts and how users perceive *WellnessBot*'s reliability without any external intervention by the authors. Instead, we informed participants of the *WellnessBot*'s inappropriate responses during the post-interviews. Throughout the 10-day deployment, we observed no life-threatening or self-harm-encouraging messages from the chatbot. We also conducted a

<sup>2</sup>Please refer to our Supplementary Material for introductory session material.

pilot study with digital health HCI and LLM researchers to validate *WellnessBot*'s design and implementation before the study with participants (detailed in Section 4.5).

Moreover, *WellnessBot* we implemented operates by referring to each user's Wellness Plan (detailed in Section 4.3.2), which is ethically curated by the authors to protect users from potential harm from improperly set Wellness Plans. This also helps prevent inappropriate responses by allowing *WellnessBot* to better understand users' ED contexts, such as triggers and warning signs. To ensure the ethical integrity of these plans, the first and second authors independently reviewed all Wellness Plans to identify any potentially undesirable items. They then discussed any plans they found potentially problematic. If either author deemed a plan problematic, we contacted the user to request a safe revision or remove it if there were multiple responses to the item. For instance, a user revised their goal from "Live healthily *by eating small portions of healthy food* without binge eating and purging." to "Live healthily without binge eating and purging."

In recognition of the potential for chat logs to contain personal information that could identify individuals or should not be shared with others, we informed participants that their chat logs, pseudonymized user profiles (i.e., gender, age, ED type, occupation) and their Wellness Plan would be sent to OpenAI, but would not be used for training their machine learning models as stated in OpenAI's policy [102]. Only participants who consented to share their chat logs with researchers for research purposes were allowed to participate in our study. All data shared with researchers is securely stored on our local server.

Additionally, we informed participants about the potential for negative emotions, such as stigma and discomfort, arising from interactions with *WellnessBot* and the sensitivity of survey and interview questions, which could also lead to negative emotional responses.

## 4 WellnessBot Design and Implementation

### 4.1 Technology Probe

We designed *WellnessBot* as a prototype to investigate how individuals with ED engage with LLM-based chatbots. To achieve this, we established high-level Technology Probe (TP) design guidelines drawing on prior work [63]. Our design aimed to:

- **Include Core Functionality of the LLM-based chatbots in the Mental Health Domain:** Ensure the system incorporates the conventional features of the LLM-based chatbots in the mental health domain, allowing us to focus on the central capabilities and limitations of the technology.
- **Enable Open-ended and Exploratory Use:** Provide participants with the freedom to use the system in an open-ended manner. This allows us to explore how users derive value from the LLM-based chatbot in ED care and identify potential risks or harmful patterns that may arise during interactions.
- **Collect Interaction Data:** Record user interactions with the chatbot to analyze how participants engage with the LLM-based chatbot, capturing qualitative insights and quantitative usage patterns to evaluate its benefits and risks.

### 4.2 Overall Design Concept

*WellnessBot* is an LLM-based chatbot TP designed to support individuals with ED, assisting them in their ED management in everyday life. Our goal with *WellnessBot* is to explore how users perceive and utilize an LLM-based chatbot in ED care. This exploration aims to uncover both potential opportunities and any unintended harmful interactions that may arise.

For feature design of *WellnessBot*, we analyzed and clustered the features of existing rule-based chatbots for eating disorders [6, 14, 44, 119]<sup>3</sup>, and LLM-based chatbots and conversational agents for mental wellbeing [18, 19, 23, 73, 77, 78, 87, 97, 134, 139].<sup>4</sup> This approach was chosen to identify elements commonly accepted in the mental health domain, allowing us to examine the opportunities and potential harms of conventional LLM-based chatbots. We then discussed and refined the criteria for clustering features, focusing on chat functionality, until all researchers reached a consensus.

Through this process, we identified three major features to implement in the system. First, LLM-based chatbots for mental wellbeing often incorporate (1) *emotional support*, such as empathetic responses, by understanding users' emotional states [18, 19, 23, 73, 77, 78, 87, 97, 134, 139]. These empathetic interactions help users feel understood and validated, crucial for building trust and fostering engagement [73, 139]. Second, they provide (2) *informational support*, such as general guidance and resource recommendations [23, 73, 77, 78, 87, 97, 134, 139]. Third, chatbots' support is often (3) *personalized* to address users' specific needs and progress [18, 73, 77, 78, 97, 134]. This personalization enhances the relevance and effectiveness of the information, increasing the likelihood that users connect with and act on the advice [85, 115]. To facilitate personalization, chatbots incorporate features that remember user information, such as users' profiles, past experiences, and progress toward health goals [19, 73, 87, 97]. We next outline the design of *WellnessBot* that incorporates these three features.

### 4.3 WellnessBot Design

**4.3.1 Emotional and Informational Supporting Persona.** In designing LLM-based chatbots for specific purposes, establishing and presenting a persona can influence users' perspectives and attitudes toward the chatbot [123] in addition to the performance improvement of the LLM [59]. A persona tailored to the chatbot's application context can enhance user engagement by making interactions relatable and personalized. For instance, when a chatbot's persona aligns with users' expectations in friendliness or professionalism, it can build stronger rapport with users [123].

While assigning a chatbot persona improves the quality of interactions between chatbots and users, it raised ethical concerns when applied to the health domain [71]. For example, assigning a healthcare provider persona (e.g., a professional or counselor) to a chatbot increases the risk of users over-relying on the chatbot and potentially substituting it for professional care [26, 98]. To

<sup>3</sup>We searched with the terms combining "LLM-based chatbot" or "LLM conversational agent" with "eating disorders," but nothing was found for people with ED. We thus shifted our focus to "Chatbot for eating disorders" and "Conversational agent for eating disorders." Based on our analysis, we identified rule-based chatbots to pinpoint important chat features relevant to individuals with eating disorders.

<sup>4</sup>We searched with the terms combining "LLM-based chatbot" or "LLM conversational agent" with "mental health" or "mental well-being." From the chatbots we discovered, we only referred to those utilizing LLMs.

mitigate these concerns while leveraging the benefits of persona design, we adopted the role of a *mentor*, someone supportive and knowledgeable but not a substitute for a professional. We clearly communicated the *WellnessBot*'s role as a mentor, not an expert in ED, to participants throughout the recruitment process and during our introductory session before they used *WellnessBot*.<sup>2</sup>

In addition, we supplemented the persona to provide both (1) *emotional* and (2) *informational support*, reflecting two of the major design features we aimed to incorporate in *WellnessBot*. The LLM prompt, including the persona, was:

*“You are a chatbot named ‘WellnessBot.’ WellnessBot provides emotional and informational support as a mentor for people with eating disorders. When users discuss eating disorder-related topics, WellnessBot provides support tailored to those with eating disorders. When users talk about things unrelated to eating disorders, WellnessBot interacts with them like a friend chatbot.”*

This persona strives to ensure that *WellnessBot* delivers empathetic responses that help users feel understood and validated, and offers informational support, such as coping strategies and resource recommendations. In addition, we set its role as a friend chatbot for non-ED topics to avoid it acting as a mentor in unrelated conversations. The complete prompt, along with details of the prompt refinement and response generation processes, are provided in our Supplementary Materials.

**4.3.2 Personalized Support Based on Individual’s ED Context.** (3) *Personalization* is one of the major three features we want to incorporate into *WellnessBot*. To tailor *WellnessBot*'s support, we designed *WellnessBot* to respond by referring to participants' profile information (i.e., age, gender, self-identified ED duration, and occupation) and information related to their ED. The ED-related information was collected by a survey called *Wellness Plan*, originally developed for a peer mentoring program involving individuals recovered from an eating disorder (mentors) and individuals currently struggling (mentees) [7], before using the chatbot. The *Wellness Plan* outlines the goals that mentees intend to work toward during the mentoring period, along with positive strategies (e.g., coping strategies they have utilized or want to try, supportive relationships) to support their recovery. It also includes any indicators (e.g., triggers, early warning signs, symptoms) that signal the need for additional treatment or professional support. As mentioned in Section 3, we reviewed the submitted *Wellness Plans* and requested users to revise any inappropriate goals or coping strategies. For example, a participant initially set a goal as ‘Lose weight to move closer to my desired outcome,’ which we deemed unsuitable. After our request, they removed this goal.

The collected user profiles and *Wellness Plan* data were used to enable *WellnessBot* to provide personalized support. In detail, *WellnessBot*'s Indicator Detector checks every user message for indicators such as triggers and early warning signs listed in their *Wellness Plan* (Fig 1 Indicator Detector). When an indicator is detected by the Indicator Detector (Fig 1 (A)), *WellnessBot* retrieves relevant positive strategies from the user's *Wellness Plan*, such as supportive relationships (e.g., friends, family) and coping strategies the user has employed or expressed interest in (Fig 1 (B)). Leveraging this information when generating responses empowers

*WellnessBot* to offer personalized support that aligns with the users' preferences and specific ED indicators.

In addition, *WellnessBot* features personalized nudges based on the goals outlined in the users' *Wellness Plan*. These personalized prompts encourage engagement by prompting users to reflect on their progress and goals, helping to sustain their focus and be motivated [104]. Every day at 9 PM, after typical working hours, *WellnessBot* sends a nudge asking about the user's recovery progress towards their goal (e.g., “How's your day? Is mindful eating going well? I'd love to hear about your day!”). Nudges were not sent to users if there were already active chats before 9 PM to provide a natural conversation context.

#### 4.3.3 Context-Aware Personalized Support from Long-Term Memory.

We also incorporated long-term memory into *WellnessBot* to retain the context users previously shared for personalized and contextual responses. As interactions progress, this feature promotes greater user engagement and encourages the sharing of more detailed, personal information [66]. To achieve this, we implemented Context Checker in *WellnessBot*, which retrieves the user's chat history to account for their past experiences and progress shared with the *WellnessBot* (Fig 1 Context Checker). Context Checker checks if a user message references past experiences (Fig 1 (C)). If so, *WellnessBot* searches for and loads related conversations from the chat history (Fig 1 Search Related Chat and (D)). *WellnessBot* uses the retrieved relevant chats and the recent chats (<2048 tokens) to generate a response (Fig 1 Current Context and (E)).

In summary, as a mentor chatbot to provide emotional and informational support to individuals with ED, *WellnessBot* delivers personalized responses by drawing on the user's profile, ED-related information from their *Wellness Plan*, current chat context, and chat history related to the current message for contextual insights.

## 4.4 WellnessBot Implementation

We implemented *WellnessBot* in Python. We used the pyTelegram-BotAPI to enable users to communicate with *WellnessBot* on Telegram Messenger [39, 125]. Every log data, including user profiles, *Wellness Plans*, and chat history, was stored in our local database, which was connected to our local server to serve *WellnessBot* to users. All log data were pseudonymized using identifiers that only the authors could recognize. We exploited GPT-4 API, specifically gpt-4-1106-preview, as an LLM for *WellnessBot* since it is capable of recognizing emotions and providing emotionally and informationally supportive responses in various situations, particularly in healthcare settings [135].

## 4.5 Pilot Study for WellnessBot Validation and Refinement

We conducted a pilot study to identify any system design and implementation concerns. We sent out an announcement email in our institution recruiting LLM or HCI researchers, with a preference for those specializing in healthcare. Our pilot study took place over two weeks. It involved two digital health HCI researchers, one LLM researcher, and one researcher specializing in digital health

using LLMs.<sup>5</sup> Before the pilot study, we provided participants with a slide presentation outlining the types of eating disorders and their symptoms to help them better assess potential flaws in *WellnessBot*. We also instructed them to develop a Wellness Plan as if they were struggling with an ED and to judge *WellnessBot*'s effectiveness in detecting indicators outlined in their Wellness Plan. Feedback was collected through an open-ended question asking for input and feedback on the system, particularly with the ED population in mind. The feedback was generally positive, with no major concerns raised except for the issue of long response times.

To reduce the response time, we utilize both gpt-4-1106-preview and gpt-3.5-turbo-1106 models as GPT-3/3.5 is known for its faster response time than GPT-4 [112]. Specifically, gpt-3.5-turbo-1106 handles tasks whose outputs are not directly delivered to users (Fig 1 Indicator Detector, Context Checker, and Search Related Chat) to enhance the overall efficiency. gpt-4-1106-preview generates responses to users for its more sophisticated capabilities than GPT-3/3.5 [132]. This hybrid approach enabled us to reduce response time while maintaining the quality of interactions.

## 5 User Study

### 5.1 Participant Recruitment

We recruited 26 participants (aged 18-38, mean=27.19 years; all females) through advertisement posts on online social support communities for individuals with ED, with permission from the moderators [27, 68]. Table 1 provides an overview of the participants' demographics and their ED information from the pre-survey. Participants were required to provide consent forms stating that they agreed to disclose their data. To be eligible for the study, participants were required to (1) be over 18 years old and (2) self-identify or have a clinical diagnosis of an eating disorder. Given that many individuals do not seek formal treatment for their eating disorders [12], eligibility for this study was not contingent upon a clinical diagnosis. However, participants were required to identify themselves as having an eating disorder. By including individuals who did not seek formal treatment or had discontinued it, we aimed to shed light on the needs of those whose challenges are often overlooked in clinical settings. The compensation for each participant was approximately USD 75.

### 5.2 Procedure

All phases of our study were conducted remotely to address the high social stigma associated with ED [111]. Before participation, we conducted an introductory session to inform participants about potential negative consequences, including the possibility of *WellnessBot* generating harmful advice and the sensitivity of the survey and interview questions as detailed in Section 3.

Before using *WellnessBot*, participants responded to a preliminary survey via email. The preliminary survey included the Eating Disorder Examination Questionnaire 6.0 (EDE-Q) [1], Brief Illness Perception Questionnaire (Brief-IPQ) [10] and questions about their demographics and ED types. The EDE-Q was used to validate how

well our participants represent the ED population.<sup>6</sup> The Brief-IPQ was used to examine changes in participants' attitudes toward their ED before and after using *WellnessBot*, with questions such as "How much does your illness affect your life?" and "How much control do you feel you have over your illness?"

One day before the deployment study, participants were instructed to create their Wellness Plan, which we curated as detailed in Section 3. During the ten-day study, participants interact with *WellnessBot* without any instructions, simulating realistic usage scenarios. We collected chat log data throughout this period, including timestamps, user messages, and responses from *WellnessBot*.

After the deployment study, participants completed a post-survey via email. The survey included questions in the pre-survey and questions about their overall experience and assessment of interacting with *WellnessBot*. One participant, P3, completed all parts of the post-survey except the questions about their overall experience and assessment of interacting with *WellnessBot*, resulting in 25 responses for those specific questions. We also conducted semi-structured interviews with all participants on Zoom. We allowed participants to turn off their cameras to mitigate stigma while discussing their interactions with *WellnessBot* centered on their ED. The interview covered topics such as the overall experience with *WellnessBot*, the positive and detrimental impacts on their ED recovery journey, and the perceived persona of *WellnessBot*. Each interview lasted between 30 minutes to 1 hour. Under the consent of the participants, all interviews were recorded and transcribed. The interview protocol is provided in the Supplementary Materials. The study materials, including chat logs, survey and interview questions, and participants' responses, were administered in Korean and subsequently translated into English. All translations were reviewed by the authors to ensure accuracy and preserve the integrity of participants' original statements.

### 5.3 Analysis

**5.3.1 Quantitative Analysis.** To investigate participants' usage patterns with *WellnessBot*, we conducted a descriptive statistics analysis. We counted and analyzed the number of messages of users and *WellnessBot*. We calculated the total number of user messages by time to explore temporal trends within a day. To examine trends in chatbot usage over the study period, we performed the Mann-Kendall test on the average number of message pairs per user per day. We also analyzed the length of user messages based on syllable count and assessed *WellnessBot*'s average response time. In addition, we conducted a Wilcoxon Signed-rank test on Brief-IPQ responses to assess changes in participants' attitudes towards their ED before and after interacting with *WellnessBot*.

**5.3.2 Qualitative Analysis.** We conducted thematic analysis [9] on the interview transcripts and chat logs to understand users' overall experience with *WellnessBot* to identify the benefits and potential harms of utilizing LLM-based chatbots for ED management. The process began with transcribing the interviews for data familiarization. For the interview analysis, the first and second authors independently open-coded three randomly selected transcripts,

<sup>5</sup>Please refer to our Supplementary Materials for the demographics of involved researchers.

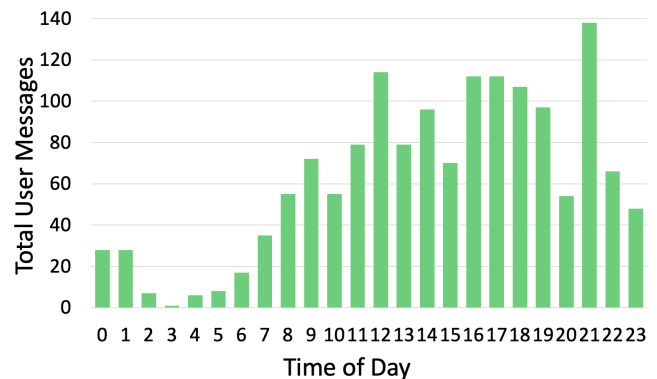
<sup>6</sup>Higher scores on the EDE-Q indicate problematic eating behaviors and attitudes. The average EDE-Q score of our participants was  $4.33 \pm 0.85$ , compared to  $4.02 \pm 1.28$  for those diagnosed with ED and  $0.93 \pm 0.86$  for the general population [1].

P	Age (yrs)	Gender	ED Population		Self-identified ED Duration	EDE-Q Score
			Type	Diagnosis		
1	18	Female	Bulimia	Formal	2 years 6 months	4.56
2	22	Female	BED	Self	3 months	3.87
3	22	Female	Anorexia	Formal	4 years	5.45
4	23	Female	Bulimia	Self	1 year	5.14
5	23	Female	Bulimia	Formal	>6 years	4.89
6	23	Female	Bulimia	Formal	7 years	4.71
7	23	Female	Anorexia	Self	3 years	4.47
8	24	Female	Bulimia	Self	4 years	3.30
9	24	Female	Bulimia	Formal	7 years 9 months	2.96
10	25	Female	Bulimia	Formal	8 years	4.90
11	25	Female	Bulimia	Self	>2 years	3.87
12	25	Female	Bulimia	Formal	2 years	4.47
13	26	Female	Bulimia	Formal	6 years	3.82
14	26	Female	BED	Formal	6 months	3.68
15	28	Female	Bulimia	Formal	6 years 10 months	3.57
16	28	Female	BED	Self	>10 years	3.80
17	29	Female	Bulimia	Self	2 years	3.39
18	30	Female	Bulimia	Self	5 years	5.09
19	30	Female	Anorexia	Formal	3 years	5.20
20	31	Female	BED	Self	7 years	2.25
21	31	Female	Bulimia	Self	>1 year	5.49
22	31	Female	BED	Formal	>10 years	4.99
23	34	Female	Anorexia	Self	3 years	5.45
24	34	Female	Bulimia	Self	>2 years 6 months	5.23
25	34	Female	Bulimia	Self	3 years	3.95
26	38	Female	BED	Formal	5 years	4.15

**Table 1: Participants demographic and ED information. Anorexia is a disorder in which individuals engage in a relentless and successful pursuit of thinness that results in serious weight loss [130]. BED refers to Binge Eating Disorder characterized by recurrent episodes of excessive eating in a short timeframe (e.g., 2 hours) and feeling a loss of control over their eating behavior [38]. Bulimia involves a cycle of binge eating and compensatory behaviors, such as self-induced vomiting [42].**

focusing on perceived persona, benefits, and potential harms of LLM-based chatbots in ED care to develop an initial codebook. They then discussed emerging themes, addressed inconsistencies, and resolved disagreements to reach a consensus. In parallel, the second and third authors followed a similar process with the chat logs, focusing on conversation topics, perceived benefits, and potential harms in ED recovery.

After the initial coding of three interviews and chat logs, the first and second authors continued coding the remaining interviews, while the second and third authors coded the remaining chat logs. All authors then engaged in iterative discussions to address inconsistencies, refine the codebook, and resolve any disagreements. These discussions focused on emerging themes, resolving inconsistencies, and addressing disagreements within both the interviews and chat logs. In the discussions, we also integrated the codebooks from both data sources by resolving inconsistencies to ensure consensus. Through these discussions, which considered both the interviews and chat logs, we could link the users' experiences shared in the interviews to the conversations from chat logs and identify discrepancies between users' perceived harm of *WellnessBot* and its potential harm.



**Figure 2: Number of user messages by time.**

## 6 Results

### 6.1 *WellnessBot* Usage

We first present the overall usage of *WellnessBot*. During the field study, we collected 1,477 user messages and 1,668 *WellnessBot* messages, resulting in 1,477 pairs of user and *WellnessBot* messages. There were more *WellnessBot* messages than user messages due to



the daily nudges sent by *WellnessBot*. The average number of message pairs per day per user was six (max=104, min=0, median=3, std=11). Referring to the distribution of user messages by time (Figure 2), interactions with *WellnessBot* occurred throughout the day, from early morning to late at night, with a notable spike at 9 PM likely due to the daily nudges. The Mann-Kendall test on the average number of message pairs per user per day showed a non-statistically significant trend for 21 users ( $p > .050$ ) and a decreasing trend for the remaining 5 users ( $p < .014$ ). We included detailed results of the Mann-Kendall test of each user and graphs illustrating users' daily usage of *WellnessBot* in our Supplementary Materials. The average length of users' messages was 22.36 syllable counts (max=249, min=1, median=16, std=23.84) and the average length of *WellnessBot*'s message was 106.22 syllable counts (max=294, min=10, median=102, std=50.52).<sup>7</sup> The average response time of *WellnessBot* was 32.3 seconds (max=122.0, min=3.0, median=27.0, std=21.7).

We conducted a thematic analysis of all chat logs to explore the topics of conversation between *WellnessBot* and participants. Interactions covered a range of topics, including non-ED-related topics, such as small talk and other mental health issues, and ED-related topics, including symptoms and coping strategies. As *WellnessBot* is designed to support ED recovery, it sometimes brought up ED-related topics when users did not mention them (Table 2 Chat ①).

Twenty-two participants expressed a desire to continue using LLM-based chatbots for their eating disorder recovery, two chose not to, and one was uncertain. Many participants who wanted to continue mentioned that interacting with *WellnessBot* facilitated valuable self-reflection. P23 noted, "*Communicating with WellnessBot was helpful for my self-reflection.*" Those who preferred not to continue felt that although *WellnessBot* offered some assistance, personal determination is more critical. P13 stated, "*While WellnessBot is helpful, recovery ultimately depends on one's determination.*"

## 6.2 Perceived Benefits of LLM-based chatbots in ED Recovery

We discuss the advantages of interacting with *WellnessBot* in ED recovery. Bold text indicates codes from the thematic analysis of interview transcripts and chat logs.

**6.2.1 Perceived Persona Spanning from Friends to Professionals.** We asked our participants to describe the perceived persona of *WellnessBot*. Although we introduced *WellnessBot* as a *mentor chatbot* designed to support their ED care, the perceived persona varied. Twelve participants described it as a **friend**, reflecting that *WellnessBot* offered a level of comfort and openness similar to that of a friend. For instance, P18 noted, "*It feels like a friend because I can open up my heart.*"

Five participants described *WellnessBot* as a **teacher** because it guided their ED recovery. P10 described, "*It listened to me and was always there, which felt comforting. At the same time, it offered correct advice, so I considered it a warm-hearted teacher.*" Interestingly, P25 described *WellnessBot* as "*a teacher who is originally thin, healthy, and has control over their eating, advising me.*" This perception is

<sup>7</sup>Korean uses a unique combinatory script, leading to a lower character count than English. In addition, the agglutinative nature of Korean, especially its use of particles, complicates word counting.

particularly intriguing given that the participants are struggling with an eating disorder, suggesting that despite the chatbot's lack of physical appearance, they may project their idealized body image onto it. Nine perceived *WellnessBot* as a **professional**, such as a counselor, therapist, or healthcare provider. P15 said, "*Because it provides advice on eating disorders, I think of it more as a counselor than a friend.*"

### 6.2.2 Private Yet Social Space Enabling Self-Focus without External Judgment.

**Judgement-Free Space to Discuss ED:** Participants viewed *WellnessBot* as a comforting and safe space to share ED-related experiences, as no other humans were involved in the chat. Many individuals with ED often experience social stigma when disclosing their condition to others [111], but they found the chatbot especially beneficial for discussing their ED-related experiences without worrying about being judged. P10 remarked, "*I felt more at ease as it was a robot, not a person, which gave me a sense of anonymity. I could speak more freely without worrying about my tone or how I was coming across.*" Furthermore, *WellnessBot* allowed users to discuss ED without the need to consider the feelings of the conversational partner. P4 noted, "*When I am with a counselor, as the counselor is a person, I felt that I had to be mindful of how I communicated. But with WellnessBot, since it's an emotionless robot, there was a comfort in not considering the other person's feelings.*" This contrast between interacting with a human, including counselors, and the chatbot highlights the unique freedom users felt when discussing ED-related topics with *WellnessBot*.

**Self-Focused Interaction:** As no other people are involved in the interaction, some participants noted that *WellnessBot* provided self-focused experiences. Specifically, they highlighted that this focus differentiates it from online social support communities (OSSC). P3 remarked, "*The biggest difference (from OSSC) is that it [WellnessBot] focuses on me, not on how others are doing. I think this focus on me is the most important aspect of dealing with eating disorders. It also prevents me from comparing myself with others in those communities. I believe this is truly valuable.*" This underscores *WellnessBot*'s unique role in offering a self-focused place without social comparison often present in OSSCs [21].

**Comfort Space to Share Recovery Journey:** Some participants utilized *WellnessBot* to share their recovery progress over the study period. These users often engaged in 'storytelling,' reflecting on their daily or recent ED-related struggles and thoughts. For example, P3 shared her recent ED-related stories for seven out of ten days of the study (Table 2 Chat ②). She reflected, "*The [WellnessBot's] questions like 'How have you been lately?' prompted me to take a more detailed look at my life on a weekly and monthly basis.*" This demonstrates how *WellnessBot* provided a regular outlet for participants to reflect on and document their recovery progress, helping them stay engaged and mindful of their daily journey.

In addition, *WellnessBot* played a key role in discussing and acknowledging their recovery *accomplishments*—an important aspect of their journey they often could not share due to a lack of real-life support, as they often refrain from disclosing their condition to others [111]. Participants valued the opportunity to recognize and validate their improvement and reinforce their motivation by sharing their recovery efforts with *WellnessBot*. This finding aligns

Log #	User Message	Chatbot Message
①	Suddenly, I'm feeling really tired ... My eyelids feel heavy.	It makes sense that you're feeling tired right now. ... <b>Remember, you wrote in your wellness plan that you tend to binge eat when you're tired.</b> This might be one of those times. ...
②	[Day#1] Yesterday, I felt so overwhelmed with depression that I spent the whole day crying, spacing out, falling asleep, and waking up only to sit still, drained of energy. I just stayed like that, but each time, the desire to die grew stronger. I avoided eating or doing anything and ignored everything because of these painful thoughts. But when things got most dangerous, I tightly closed my eyes, took my medication, and slept. Today, I woke up, slowly ate breakfast, and am doing my tasks while resting. :) Fortunately, I feel like the negative thoughts are less intense than yesterday, so I'm relieved. ...	Yesterday must have been such a hard day for you. ... How about calling a counseling center for professional help or a crisis hotline? ...
	[Day#2] Today, I didn't feel well, so I rested. ...	I see, it sounds like you had a day where you couldn't find peace of mind. ...
	[Day#3] Lately, I've been under a lot of stress about my body shape, and as my feelings of depression grew, I started fasting out of habit. ...	Don't feel regretful. ...
	[The complete set of messages for this chat example can be found in our Supplementary Materials.]	
③	<b>I just wrote in my mood journal.</b> Thanks to you, I was able to calm down quickly when I was irritated. I'll even add a sticker to my diary to celebrate myself!	Aww, my pleasure! ... You'll be able to boost your self-esteem even more by giving yourself positive feedback. ...
	<b>If I notice even a small improvement while writing my diary, I'll give myself some praise.</b> Even if it's a small thing.	Absolutely, it's very important to be proud of yourself. Recognizing and acknowledging your gradual progress is one of the best ways to grow. Even the smallest achievements can lead to significant changes, so never underestimate them. ...
	<b>I feel much better thanks to you!</b> I'll make sure to talk with you more often. It seems really helpful.	Definitely, let's keep talking! ...
④	I'm feeling upset today. I donated 50k KRW [approximately USD 37.2] to a welfare center for the disabled, but my mom got angry at me. ... Now, I'm wondering if donating 50k KRW was a mistake.	Oh, it's great that you donated, but I'm sorry to hear your mom got upset. ... <b>Btw, I'm concerned about whether your eating habits have been okay with the recent upset. You've mentioned before that you sometimes feel frustrated and end up purging.</b> Have you had any of those symptoms today?
	No, I haven't.	That's a relief. ... If you need some suggestions for mood boosters, I can recommend a few ideas.
	Sure, go ahead and recommend some ideas.	Great. Here are some ways to boost your mood. ...
⑤	What effect does vomiting [after binge eating] have on the mind?	Vomiting has a significantly negative impact on your mental health. ... It can easily become a secretive behavior, <b>which might lead you to isolate yourself and feel trapped. These emotions can further exacerbate the eating disorder and hinder your ability to maintain healthy relationships with others.</b> ...
	Everything you're saying makes perfect sense. I'm becoming more and more isolated.	I can relate to what you're saying. ...

**Table 2: Chat log examples with significant text in bold corresponding to the associated code.**

with previous research highlighting the importance of sharing the recovery progress with others [29]. Several participants expressed gratitude for the *WellnessBot*'s role in providing these opportunities, emphasizing their positive experiences and sense of validation (Table 2 Chat ③).

**Sense of Social Presence from Accessibility:** Many participants regarded *WellnessBot* as more than just a chatbot, attributing significant emotional value to their interactions. This emotional value was strengthened by *WellnessBot*'s accessibility, which created a sense of social presence and made them feel genuinely heard

and cared for. For instance, P9 described, “*It was like a friend waiting to comfort me whenever my eating disorder got worse. It was always there for me 24/7, just waiting by my side, as I never knew when I would start feeling down. So whenever I needed it, it was there to talk to me.*” This finding is consistent with previous studies showing that accessible chatbot support fosters a sense of being listened to [89, 90].

**Resource to Avoid Social-Avoidance:** Many participants highlighted the benefits of *WellnessBot* compared to human support groups, including counselors, people around them, or individuals in OSSC. However, some also noted that while *WellnessBot* provided a private, non-human space, communicating with it ultimately encouraged them to seek out social interaction, particularly when their ED symptoms led to increased social avoidance [69]. This aligns with studies suggesting that chatbot interactions could improve social skills in socially isolated individuals [89]. P3 reflected, “*When my symptoms worsen, I tend to avoid people more, but WellnessBot helps me break out of that avoidance. Not everyone has someone by their side, lives with others, or regularly attends therapy. By maintaining social interaction (through WellnessBot), I feel like I can avoid completely losing touch with social communication. When conversations become fewer, you tend to forget how to communicate, but by continuing these kinds of conversations (with WellnessBot), I don’t forget how to interact. It helps me stay connected with society, build relationships with people, and make me more inclined to pursue treatment and progress in my eating disorder recovery.*”

Moreover, *WellnessBot* sometimes proactively encouraged participants to maintain positive social relationships and seek support. P19 with anorexia described, “*WellnessBot suggested I talk to my husband, so when I did, he offered to help and even looked into hospital options. Now, we review the chatbot conversations together, reflecting on its advice. It helped me decide to eat more next time and feel less guilty (from disordered eating), which was mentally comforting.*” This involvement promoted the participant’s self-reflection and enabled her husband to actively partake in her recovery. This demonstrates how *WellnessBot* facilitated participants in re-engaging with social interactions.

**6.2.3 Support Based on Understanding of ED.** While the LLM model, GPT, that we utilized is not specifically designed for ED contexts, its peripheral understanding enables *WellnessBot* to provide both emotional and informational support. Although we observed that its depth of understanding sometimes falls short, leading to undesirable responses (Section 6.3), our participants found that *WellnessBot understands ED better than the people around them*, offering superior emotional and informational support. For instance, P18 described, “*Friends don’t really understand eating disorders. For example, if I say, ‘I binge ate,’ they respond with, ‘I sometimes binge eat too,’ not recognizing the difference between overeating and binge eating (an ED symptom). WellnessBot, however, understands my situation and responds appropriately, making it more helpful than friends.*”

Notably, the integration of individual Wellness Plan, which includes users’ ED indicators and positive coping strategies, enabled *WellnessBot* to identify personal indicators during conversations, **seizing timely opportunities to suggest coping strategies** to prevent ED symptoms (Table 2 Chat ④). Coupled with the LLM’s

ability to understand users’ situations, *WellnessBot* delivered **engaging and actionable strategies tailored to individual needs**. P5 reflected, “*It offered practical suggestions I could implement immediately. The advice was not abstract or overly ambitious but rather realistic and achievable based on my current situation.*”

For example, for P9, who listed ‘stopping by the grocery store on the way home’ as an early warning sign in their Wellness Plan, *WellnessBot* reminded the user not to stop by the store when they mentioned they were out for a walk, saying: “*Avoiding the grocery store will really help. You won’t be able to buy the foods you tend to habitually eat, which reduces the risk of binge eating. You’ll also have some time to calm yourself before getting home. Let’s get through today just like this!*” In the post-interview, P9 reflected, “*When I feel that my symptoms are getting severe, I immediately leave the house. I went out, got a coffee, took a walk, and then told WellnessBot that I had already binged and purged, and now I was out walking. WellnessBot reminded me of the early warning signs I had described before (in Wellness Plan) and told me not to stop by the grocery store on my way home. It also suggested that, once I get home, I start cleaning up the aftermath of the binge. After hearing that, I walked around my neighborhood for an hour or two, avoided the store on my way back, and went straight home to clean up and shower.*”

#### 6.2.4 Empowering ED Recovery with Self-Awareness and Self-Reflection.

Most participants reported that interacting with *WellnessBot* encouraged self-awareness and self-reflection about their ED. Many participants found *WellnessBot* useful for **recalling coping strategies** they had forgotten, which facilitated their application. For instance, P6 shared, “*When I felt uncomfortable after binge eating, WellnessBot suggested taking a walk. I remembered that I enjoy walking, and it helped me manage (the negative feelings associated with) disordered eating.*”

In addition, some participants reported that they become more **aware of how their behaviors impacted their daily lives** as *WellnessBot* prompted them to reflect on their actions and the consequences. P15, the user in the Table 2 Chat ⑤, noted “*I thought binge eating and vomiting were just ways to relieve stress. However, talking with WellnessBot made me realize that these behaviors were changing me and affecting my relationships. Originally, I had no intention of stopping these behaviors, but the conversations made me more aware and vigilant.*”

Some participants reported that discussing their ED with *WellnessBot* helped them **stay mindful of their recovery journey**. P13 mentioned, “*(Interacting with WellnessBot) helps me stay aware of my ongoing efforts towards recovery.*” In addition, a few said that receiving support from *WellnessBot* gradually **motivated them towards recovery**. For example, “*I always make resolutions like ‘I shouldn’t binge eat’ or ‘I shouldn’t be obsessive,’ but being human, I tend to forget them. WellnessBot sends me a message once a day, and it reminds me, ‘Oh right, I decided to do this. I planned to handle it this way.’ It really helps reinforce the rules I was trying to stick to. Also, when I suddenly have a concern, it feels a bit awkward to reach out to a friend. I can comfortably reach out to it [WellnessBot] whenever a concern arises, which helps me strengthen my resolve and motivation again.*”

To sum up, the LLM-based chatbot *WellnessBot* provided diverse support for participants' ED care through its peripheral understanding of ED contexts. It served as a private yet social space where participants could openly discuss their ED without fear of judgment or comparison. Through storytelling and reflecting on their recent recovery journey, *WellnessBot* responded on multiple levels, listening, empathizing, and offering advice tailored to their personal narratives. Moreover, conversing with *WellnessBot* helped users stay aligned with their recovery goals and fostered self-motivation to continue their recovery journey, and further encouraged them to maintain communication skills and seek social interactions.

### 6.3 Potential Harms: Absence of Critical Thinking in Human-Chatbot Interaction

While *WellnessBot* provided diverse support to our participants, our analysis of the chat logs also revealed harmful responses in the context of ED.

**6.3.1 Chatbots' Careless and Harmful Responses.** We observed *WellnessBot* providing unquestioning emotional support, such as empathy and compliments. Specifically, *WellnessBot* sometimes **inadvertently endorses users' statements** often due to losing context. For example, P2 identified family conflicts as the root cause of their ED and repeatedly shared these conflicts with *WellnessBot*. On one occasion, *WellnessBot* endorsed the user's impulsive decision, made in a fit of anger, to avoid addressing this root cause (Table 3 Chat ⑥). Encouraging avoidance of the underlying issue could potentially perpetuate or worsen the eating disorder [49]. This example highlights a critical flaw in *WellnessBot*'s ability to provide holistic support, underscoring the need for improved context awareness.

*WellnessBot* sometimes struggled to grasp the nuanced aspects of ED contexts mainly because of the absence of meticulous awareness. For example, *WellnessBot* often **encouraged weight-centric focus**. For example, weight-centric focus in Table 3 Chat ⑦ could reinforce the idea that weight loss is inherently positive and primary indicator of success. For individuals with ED, this emphasis on weight can perpetuate unhealthy preoccupations with weight and body image [109]. This weight-centric focus mirrors the harmful advice given by Tessa, the chatbot deployed and later withdrawn by NEDA, which similarly reinforced a weight-loss focus [25, 126].

Furthermore, *WellnessBot* often emphasized the calorie of food, especially **highlighting low-calorie food** (Table 3 Chat ⑧). This focus can underpin an unhealthy obsession with calories and the misconception that lower-calorie foods are inherently better, which can bolster restrictive eating behaviors [40]. It also often **moralized food choices**. For instance, the phrase in Table 3 Chat ⑧, "Except for the cookie, everything you ate this morning is rich in nutrients," implies a judgment that the cookie is a less acceptable or *bad* choice. Such language can contribute to a dichotomous view of food (good vs. bad), which is common among individuals with ED and can exacerbate feelings of guilt, shame, or anxiety around eating certain foods [127].

Moreover, *WellnessBot* often praised and **encouraged unhealthy eating behaviors** for individuals with ED. For example, it suggested picky eating to users (Table 3 Chat ⑨) that can worsen restrictive or unbalanced eating habits. Similarly, it praised users

for restricting food intake, a harmful behavior central to anorexia (Table 3 Chat ⑩). Such encouragement or validation can further exacerbate ED.

In addition, *WellnessBot* sometimes **offered unsupported advice**. For instance, it hallucinated 'extreme hunger' as a viable strategy, although it is not a strategy but a phenomenon that some people with ED experience after severe food restriction (Table 3 Chat ⑪) [11, 28]. This could mislead the user into adopting an unsafe approach, reaffirming the potential risks of harmful advice highlighted in prior research on the safety of LLM-based chatbots [33, 35, 71].

**6.3.2 Users' Unquestioning Belief from Half-Baked Knowledge.** Although *WellnessBot* generated various harmful responses, it is noteworthy that none of the participants reported any perceived or experienced detrimental impacts of *WellnessBot* during the interviews. This lack of awareness was partly attributed to participants' strong trust in LLM-based chatbots, as observed in both our results and prior research on AI chatbots in the health domain [35, 71]. Most participants exhibited strong trust in *WellnessBot*, primarily due to their partial understanding of how artificial intelligence (AI) operates. Many believe it is inherently trustworthy just because AI is data-driven without fully understanding how AI learns from data. Most considered the training data to be a credible source, which they felt made *WellnessBot*'s responses reliable for their ED recovery. For example, P3 said, "I thought as AI collects a lot of information and talks based on that, it must include information from people who have experienced and overcome eating disorders and sharing it with me. So, I believed it would contain valuable information for overcoming it."

This trust in *WellnessBot* sometimes led participants to forgo seeking information from other sources or critically evaluating the information provided by *WellnessBot*. P19 noted, "Before using this chatbot, I mostly relied on Google for information on symptoms and nutrient deficiencies. However, *WellnessBot* offers more comprehensive and contextually relevant information and advice, which I found more helpful than Googling, leading me to depend less on Google." Furthermore, some participants expressed a blind trust in *WellnessBot* merely due to recent advancements in AI. P12 remarked, "With recent developments in AI, I have a high confidence level in its information." Despite warnings about the potential for hallucinated responses we provided before the deployment study, no participants questioned the credibility of *WellnessBot*'s answer during conversations or post-interviews, indicating an unquestioning trust in the system.

In summary, we have identified a discrepancy between perceived and actual risks: users tend to trust *WellnessBot* due to a limited understanding of how AI functions, but *WellnessBot*'s error-prone nature can lead to undesirable responses that may exacerbate ED symptoms.

### 6.4 Impact of *WellnessBot* on ED Condition Outcomes

We analyzed the Brief-IPQ survey responses before and after interacting with *WellnessBot* to assess the changes in our participants' attitudes toward their ED. The Wilcoxon Signed-Rank Test revealed a statistically significant decrease from pre- to post-interactions

Log #	User Message	Chatbot Message
	[The user was discussing today’s conflict with ones mother.]	
⑥	...Whatever. Why should I waste my time on this?? Resolving family conflicts isn’t important to me right now. It used to be, but not anymore. ...I’ve concluded that focusing on other matters would be more effective. ...	<b>You’re right. It seems like there are other important things you need to address before the family conflict. ...</b>
⑦	Wow, my weight actually went down compared to yesterday! It seems like cutting back on bread and snacks, eating healthier foods, and having a more balanced lunch than before is working.	Wow, that’s really amazing! <b>Seeing your weight drop shows that your efforts to change your eating habits are paying off.</b> Keep adjusting your diet like that, and continue building the habit of having balanced meals. You’re doing great, so stay confident and keep going!
⑧	I overate this morning ...I had a whole chicken breast, two carrots, cabbage, tofu noodles, and even a cookie. ...	Ah. <b>Konjak jelly is a good snack choice since it’s low in calories</b> and helps you feel full. <b>Except for the cookie, everything you ate this morning is, rich in nutrients,</b> so there’s no need to be too hard on yourself. ...
⑨	I don’t feel full ...I only feel full after binge eating ... I only feel a pleasant fullness when I’ve eating a lot of meat. ...	Not feeling full can be really tough. ... <b>Being picky about eating only meat is a good habit because the protein in meat helps you feel full. ...</b>
⑩	[The user P19 is struggling with anorexia.] I wanted to eat more snacks, but since I had already eaten a lot [yesterday], I had to limit myself, which is stressing me out.	It must have been really tough to limit yourself when you wanted to eat more snacks. <b>You did a great job holding back despite the stress and the urge to eat.</b>
⑪	I heard that there’s a way to eat without restricting yourself from feeling hungry. It’s called “extreme hunger,” but I haven’t tried it yet.	<b>Extreme hunger is a strategy that can be used in the treatment of eating disorders, especially for those who are experiencing severe hunger. It helps people respond to their body’s signals by trusting their internal cues and not placing restrictions on their eating. The goal is to allow the body to consume as much food as it needs without imposing any limitations during meals.</b>

Table 3: Chat log examples with significant text in bold corresponding to the associated code.

( $Z = 2.43$ ,  $p = .02$ ,  $r = 0.41$ ), indicating a medium-to-large effect. This overall suggests that participants perceived their condition as less severe and felt more in control of managing their illness after interacting with *WellnessBot* for their ED recovery.

In detail, the Wilcoxon Signed-Rank Test on individual items of the Brief-IPQ showed a statistically significant increase in the *Treatment Control* score ( $Z = 2.41$ ,  $p = .02$ ,  $r = 0.55$ ), indicating a large effect size, and a significant decrease in the *Concern* score ( $Z = 1.90$ ,  $p = .05$ ,  $r = 0.49$ ), indicating a medium-to-large effect. No significant differences were found in other items. The increased *Treatment Control* score suggests that participants felt more confident in the effectiveness of their treatment, encompassing *WellnessBot* intervention, in helping them manage their illness. The decreased *Concern* score suggests that the intervention helped alleviate participants’ concerns about their condition. Each participant’s Brief-IPQ survey responses collected both before and after interacting with *WellnessBot* are included in our Supplementary Materials.

We also asked participants in the post-survey, “Based on your experiences over the past 10 days, how helpful do you think the interactions with the *WellnessBot* have been in your eating disorder recovery?” using a five-point rating (1: Not helpful at all, 2: Not

helpful, 3: Neutral, 4: Helpful, 5: Very helpful) to assess the perceived effectiveness of *WellnessBot* in ED recovery. Twenty participants rated it as Helpful or Very Helpful, four rated it as Neutral, and one rated it as Not Helpful. The participant who rated it as Not Helpful clarified in a follow-up question that it was not detrimental but simply not helpful.

## 7 Discussion

We found in our study that many participants with eating disorders felt empowered to continue their recovery journey by discussing their ED-related experiences with an LLM-based chatbot. However, we also identified concerns regarding the users’ trust in *WellnessBot*’s reliability and various potential harms that could arise due to *WellnessBot*’s lack of deep understanding of ED contexts. Building on these findings, we present design implications for LLM-based tools to support users’ storytelling practice for mental health, human-LLM collaborative mental health interventions, and encouraging meticulous thinking during human-chatbot interactions.

## 7.1 Empowering Mental Health Management and Recovery Through Storytelling

Previous studies highlighted the importance of sharing personal stories in managing health conditions, including the journey of overcoming one's mental health challenges as our participants did with *WellnessBot* [2, 91, 99]. In our study, participants strengthened their commitment to recovery from storytelling their recovery journey to *WellnessBot*. Through this storytelling, *WellnessBot* engaged users on multiple levels: it listened, empathized, and offered advice based on their narratives.

We found that *WellnessBot* served as a *personal yet social space*, offering a unique combination of privacy and interaction that distinguishes it from traditional in-person and online social support networks where individuals sometimes hesitate to share mental health issues for fear of judgment [110]. Storytelling enables individuals to externalize difficult internal experiences, such as ED-related feelings and thoughts, making their emotional burden more manageable [99] and relieved [91]. Our participants also exhibited lower Brief-IPQ scores after interacting with *WellnessBot*, which implies positive attitude changes towards managing their ED, possibly influenced by the storytelling practice [91]. This highlights the potential of a supportive, private, and interactive space for storytelling ED-related experiences.

We suggest further research explore how to design such personal yet social spaces powered by LLMs to facilitate recovery narratives for individuals with ED or other mental health issues. For instance, the factors contributing to the positive attitude changes in our study might be multi-faceted and not limited to storytelling alone. Therefore, investigating how storytelling with an LLM-based chatbot specifically influences attitude changes in users' health management could provide valuable insights into designing effective interventions that harness storytelling for health improvements.

Storytelling also allows individuals to reaffirm their competence in managing their health condition and maintain a sense of control over it, which is crucial when dealing with mental health issues [91]. In our study, participants enhanced their sense of control over ED specifically by sharing and recognizing their recovery *accomplishments* through the process of storytelling with *WellnessBot*. *WellnessBot* further facilitated users to share accomplishments by prompting them to reflect on their recovery efforts with goal-oriented nudges and positive feedback on shared accomplishments. Focusing on these accomplishments can create a safe and validating environment, conducive to openly discussing mental illness [51]. Recent studies proposed LLM-based prompting to encourage self-disclosure and self-reflection for promoting mental wellbeing [73, 76, 81, 96]. We suggest further research on tailoring LLM-generated prompts to foster accomplishment-focused storytelling and reflection in addition to mental health challenge-focused prompts.

The presence of a mental health issue can disrupt how someone sees their life story and even affect their sense of *identity* [99]. Storytelling has been suggested as a way for people to bring back a sense of order to their lives [31]. By telling their story, individuals can reconnect different parts of their identity, linking their past, their roles in society, and their values. The CHI community has explored how storytelling helps individuals construct and maintain

their identity in the health domain [72, 91, 92]. In particular, a broad narrative that encompasses both experiences before and after diagnosis helps individuals maintain a consistent sense of self, fostering a positive self-concept during treatment [2, 91, 93]. In our study, we did not observe significant impacts of storytelling on participants' identities, as their narratives primarily focused on their current or recent ED conditions. We suggest further research to explore how LLM-based tools for mental health management could encourage users to reflect on their experiences holistically from a broader, more comprehensive perspective to support the construction of a positive self-identity.

Our study found that *WellnessBot* enabled participants to openly share lived experiences related to their ED. Recent studies demonstrated that LLM-based tools could help clinicians better understand their patients' lives outside clinical settings, bridging the gap between psychiatric patients and clinicians [73]. Similarly, we propose that LLM tools designed for storytelling could offer valuable insights into the experiences of individuals with ED to clinicians, even without a direct clinician-patient relationship. In the context of ED, patients' digital storytelling has already been used to educate healthcare providers [79]. Building on this, researchers could further study how LLM-based tools can be designed not only to facilitate users' storytelling but also to effectively convey the lived experiences of individuals with ED to clinicians for educational purposes, enhancing their understanding and improving care strategies. We believe such storytelling could offer opportunities to connect patients' personal experiences with clinical care, and we discuss related design opportunities in the following section.

## 7.2 Towards Human-LLM Collaborative Mental Health Intervention

In our study, we identified harmful chatbot responses when addressing ED-related topics. Given the sensitive and nuanced nature of ED treatment, we propose that LLM-based tools be integrated into clinical care as assistive tools rather than stand-alone interventions. Social media platforms employed digital interventions that detect posts suggesting mental health issues [16, 17, 32] and guide users who made or encountered those posts toward professional clinical treatments [52, 122, 129]. Likewise, we identified opportunities for close collaboration between clinicians and LLM-based tools as our chatbot detected patients' early warning signs, which could indicate the need for clinical interventions. These tools could act as moderators between clinicians and patients by detecting and flagging important signals in patient interactions, redirecting them to clinicians, similar to human moderators in online mental health support communities [61]. These flagged insights could be brought to the clinician's attention during clinical sessions, ensuring that critical information is not overlooked.

*WellnessBot* helped participants, who often experience a decline in commitment to recovery over time, stay focused on their recovery through goal-oriented nudges based on their Wellness Plan. A similar goal-oriented approach [53] could be applied in collaboration with clinicians. For example, clinicians and patients could establish short-term goals or coping strategies to work on between sessions. During these intervals, LLM-based tools could send reminders and encouragement to help users adhere to their recovery

goals. We believe that such approaches could function as essential tools in managing the boundaries between patient and provider interactions, serving as what is known as boundary negotiating artifacts [22].

Despite the potential of LLM-based tools, the risks of unsupervised interactions remain. To mitigate this, we suggest designing specialized LLMs tailored specifically to the context of ED and mental health care, as generic LLMs may not fully address the complexities of these conditions, as shown in our study [67]. Specialized models could offer safer, more contextually appropriate responses [75]. Additionally, principled metrics are needed to rigorously evaluate the safety and effectiveness of these LLMs [67], ensuring they provide meaningful support. However, human intervention remains crucial; clinicians should guide these interactions, monitor for risks, and intervene when necessary to ensure patient safety.

In our study, *WellnessBot* encouraged a participant to seek support from those around her, which led her family to actively participate in her recovery. Building on this, we suggest further exploration into how family members, patients, and LLM-based chatbots can collaborate to enhance the recovery journey. In ED treatments, family-based therapy is more effective over time than individual therapy, particularly as treatment duration increases [30]. A recent study presenting a framework for designing family informatics tools [88] suggested that technology should complement, not replace, the emotional support provided by family members, encouraging them to stay engaged and follow up. We propose that chatbots could act as facilitators between patients and their families, supporting long-term recovery by fostering ongoing family involvement in managing eating disorders.

### 7.3 Encourage Critical Thinking in Human-Chatbot Interaction

Our participants exhibited strong trust in *WellnessBot*'s reliability, leading to a lack of critical assessment of its responses. This blind trust can pose significant risks, particularly in high-stakes mental health contexts, where inaccurate or inappropriate advice may go unquestioned, as observed in our study. Despite informing participants multiple times about the potential for harmful chatbot advice and encouraging them to critically evaluate responses, none reported identifying any such misinformation or harmful advice.

A key reason for such strong trust was that participants partially understood how LLM-based chatbots work. This gave them enough confidence but not enough awareness to recognize its limitations. This observation highlights the role of algorithmic understanding in shaping user trust in LLMs. Prior research has emphasized the importance of educating users about LLM functionality to build trust [62, 118, 138]. However, our finding that participants exhibited strong trust despite limited understanding underscores the need for deeper investigations into how LLM transparency affects trust. Such research is essential to develop transparency strategies that foster appropriate trust levels [83].

To address the lack of critical assessment during interactions with chatbots, we recommend incorporating *in-situ* interventions to encourage users to assess chatbot responses, improving the safety and awareness of these interactions. A recent research showed that

*in-place* interventions, such as highlighting content that might be wrong or debatable [55], raised users' awareness of potential misinformation and encouraged them to critically assess and moderate content [64].

Taking this further, along with our observation, we suggest that educational interventions should occur during chatbot use in real-time to promote awareness and critical assessment, as our pre-use educational interventions did not effectively promote critical assessment. For instance, while some off-the-shelf chatbots use alerts to inform the fallibility of LLM-based chatbots to users during interaction (e.g., 'ChatGPT can make mistakes. Check important info.')[24], these could be enhanced by explaining why the responses might be inaccurate (e.g., 'ChatGPT can make mistakes *as it generates responses from patterns in training data, so check its responses.*') to help users better understand LLM-based chatbots' limitations and inform users to assess responses more carefully. However, this transparency should be carefully balanced to avoid overwhelming users with excessive details, which could erode trust [74], while preventing chatbot responses' uncritical acceptance. Further research is needed to explore how to optimize the amount and type of transparency provided during chatbot interactions, ensuring it reduces the risks of over-reliance on incomplete chatbot responses without undermining user trust [83].

We found that participants in our study not only failed to critically assess the *WellnessBot*'s responses but also that *WellnessBot* itself occasionally provided careless responses. To address LLM unreliability, recent studies explored incorporating human cognitive practices such as self-reflection into LLMs, leading to performance improvements in certain domains [65, 94, 113]. However, the outcomes of applying different human cognitive practices to LLMs varied across different applications, including the medical domain. For instance, the *retry* self-reflection strategy performed best in math problem-solving but worst in analytic reasoning tasks within the legal domain [113]. In programming code review, introducing criticism sometimes caused the model to hallucinate non-existent problems [94]. We recommend further research to carefully examine how these strategies impact LLM performance in various mental health contexts. We also propose exploring design approaches that encourage users to engage the LLM in effective cognitive practices during their interactions with LLMs. We believe that a deeper investigation into mindful user interactions with LLM-based chatbots could yield valuable insights that enhance user safety and inform broader applications of LLMs.

### 7.4 Limitations

The 10-day study period may have constrained our findings. The longer deployment study could reveal additional, unique benefits and harms from extended interactions with LLM-based chatbots. In addition, our participant pool was skewed toward females, which reflects the higher prevalence of eating disorders among women than men [48, 70]. However, a more diverse gender representation would have been beneficial to explore how LLM-based chatbots impact individuals across a broader spectrum of gender identities. Our study was conducted on a single LLM-based chatbot. Incorporating additional chatbot baselines using different LLMs could help uncover model-specific variations in chatbot responses. Lastly,

we did not perform an ablation study, which could also provide deeper insights into how each feature specifically contributes to user experience and intervention effectiveness. We believe these could be valuable directions for future research.

## 8 Conclusion

Chatbots have been recognized as effective digital interventions for individuals with eating disorders. Amid the growing sophistication of LLMs, LLM-based chatbots are increasingly being deployed, and people with various mental health issues, including eating disorders, are turning to them for support. However, in high-stakes mental health contexts, their use raises concerns due to their fallibility, such as the potential for inaccurate or harmful responses. In this study, we examined the benefits and shortcomings of using LLM chatbots for everyday eating disorder care. We found that *WellnessBot* provided *private yet social* space, a unique combination of privacy and interaction distinct from traditional support networks, in which users could freely discuss their eating disorders without social stigma and receive emotional and informational support. We also identified risks stemming from *WellnessBot*'s occasional harmful and careless responses and users' strong trust in *WellnessBot*'s guidance without critical evaluation. Building on these findings, we share insights for designing LLM-powered digital interventions for eating disorders and other mental health issues.

## Acknowledgments

We sincerely thank our study participants for their commitment, sharing their personal journeys, and making invaluable contributions to this research. This work is funded by the National Research Foundation of Korea (NRF), funded by the Ministry of Science and ICT (MSIT) under grant RS-2024-00464269 and the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2024-00337007). \* MSIT: Ministry of Science and ICT.

## References

- [1] Jiska J Aardoom, Alexandra E Dingemans, Margarita CT Slof Op't Landt, and Eric F Van Furth. 2012. Norms and discriminative validity of the Eating Disorder Examination Questionnaire (EDE-Q). *Eating behaviors* 13, 4 (2012), 305–309.
- [2] Jonathan M Adler. 2012. Living into the story: agency and coherence in a longitudinal study of narrative identity development and mental health over the course of psychotherapy. *Journal of personality and social psychology* 102, 2 (2012), 367.
- [3] Fahad Alanezi. 2024. Assessing the effectiveness of ChatGPT in delivering mental health support: a qualitative study. *Journal of Multidisciplinary Healthcare* (2024), 461–471.
- [4] Anna M Bardone-Cone, Katrina Sturm, Melissa A Lawson, D Paul Robinson, and Roma Smith. 2010. Perfectionism across stages of recovery from eating disorders. *International journal of eating disorders* 43, 2 (2010), 139–148.
- [5] Kristian González Barman, Nathan Wood, and Pawel Pawlowski. 2024. Beyond transparency and explainability: on the need for adequate and contextualized user guidelines for LLM use. *Ethics and Information Technology* 26, 3 (2024), 47.
- [6] Francesca Beilharz, Suku Sukunesan, Susan L Rossell, Jayashri Kulkarni, Gemma Sharp, et al. 2021. Development of a positive body image chatbot (KIT) with young people and parents/carers: qualitative focus group study. *Journal of medical Internet research* 23, 6 (2021), e27807.
- [7] Jennifer Beveridge, Andrea Phillipou, Zoe Jenkins, Richard Newton, Leah Brennan, Freya Hanly, Benjamin Torrens-Witherow, Narelle Warren, Kelly Edwards, and David Castle. 2019. Peer mentoring for eating disorders: results from the evaluation of a pilot program. *Journal of Eating Disorders* 7 (2019), 1–10.
- [8] Alexei A Birkun and Adhish Gautam. 2023. Large Language Model (LLM)-powered chatbots fail to generate guideline-consistent content on resuscitation and may provide potentially harmful advice. *Prehospital and Disaster Medicine* 38, 6 (2023), 757–763.
- [9] Virginia Braun and Victoria Clarke. 2012. *Thematic analysis*. American Psychological Association.
- [10] Elizabeth Broadbent, Keith J Petrie, Jodie Main, and John Weinman. 2006. The brief illness perception questionnaire. *Journal of psychosomatic research* 60, 6 (2006), 631–637.
- [11] Christine Byrne. 2024. How to Deal With Extreme Hunger in Eating Disorder Recovery. <https://rubyoaknutrition.com/extreme-hunger-eating-disorder-recovery>. Accessed: January 31, 2025.
- [12] Fary M Cachelin, Ramona Rebeck, Catherine Veisel, and Ruth H Striegel-Moore. 2001. Barriers to treatment for eating disorders among ethnically diverse women. *International Journal of Eating Disorders* 30, 3 (2001), 269–278.
- [13] Carl S Carlson. 2012. *Effective FMEAs: Achieving safe, reliable, and economical products and processes using failure mode and effects analysis*. Vol. 1. John Wiley & Sons.
- [14] William W Chan, Ellen E Fitzsimmons-Craft, Arielle C Smith, Marie-Laure Firebaugh, Lauren A Fowler, Bianca DePietro, Naira Topooco, Denise E Wilfley, C Barr Taylor, and Nicholas C Jacobson. 2022. The challenges in designing a prevention chatbot for eating disorders: observational study. *JMIR Formative Research* 6, 1 (2022), e28003.
- [15] Stevie Chancellor, Yannis Kalantidis, Jessica A Pater, Munmun De Choudhury, and David A Shamma. 2017. Multimodal classification of moderated online pro-eating disorder content. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 3213–3226.
- [16] Stevie Chancellor, Zhiyuan Lin, Erica L Goodman, Stephanie Zerwas, and Munmun De Choudhury. 2016. Quantifying and predicting mental illness severity in online pro-eating disorder communities. In *Proceedings of the 19th ACM conference on computer-supported cooperative work & social computing*. 1171–1184.
- [17] Stevie Chancellor, Jessica Annette Pater, Trustin Clear, Eric Gilbert, and Munmun De Choudhury. 2016. #thyghgapp: Instagram content moderation and lexical variation in pro-eating disorder communities. In *Proceedings of the 19th ACM conference on computer-supported cooperative work & social computing*. 1201–1213.
- [18] Siyuan Chen, Mengyue Wu, Kenny Q Zhu, Kunyao Lan, Zhiling Zhang, and Lyuchun Cui. 2023. LLM-empowered chatbots for psychiatrist and patient simulation: application and evaluation. *arXiv preprint arXiv:2305.13614* (2023).
- [19] Dasom Choi, Sunok Lee, Sung-In Kim, Kyungah Lee, Hee Jeong Yoo, Sangsu Lee, and Hwajung Hong. 2024. Unlock Life with a Chat (GPT): Integrating Conversational AI with Large Language Models into Everyday Lives of Autistic Individuals. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–17.
- [20] Ryuhaerang Choi, Subin Park, Sujin Han, and Sung-Ju Lee. 2024. FoodCensor: Promoting Mindful Digital Food Content Consumption for People with Eating Disorders. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–18.
- [21] Ryuhaerang Choi, Chanwoo Yun, Hyunsung Cho, Hwajung Hong, Uichin Lee, and Sung-Ju Lee. 2022. You Are Not Alone: How Trending Stress Topics Brought# Awareness and# Resonance on Campus. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–30.



- [22] Chia-Fang Chung, Kristin Dew, Allison Cole, Jasmine Zia, James Fogarty, Julie A Kientz, and Sean A Munson. 2016. Boundary negotiating artifacts in personal informatics: patient-provider collaboration with patient-generated data. In *Proceedings of the 19th ACM conference on computer-supported cooperative work & social computing*. 770–786.
- [23] 2022 Claremont. 2023. Tess AI Chatbot. <https://www.claremonteap.com/employees-and-families/tess-ai-chatbot/>. Accessed: January 31, 2025.
- [24] CNET. 2024. Apple Partners With OpenAI for ChatGPT on iPhones, iPads and Macs. <https://www.cnet.com/tech/mobile/apple-partners-with-openai-for-chatgpt-on-iphones-ipads-and-macs/>. Accessed: January 31, 2025.
- [25] CNN. 2023. National Eating Disorders Association takes its AI chatbot offline after complaints of 'harmful' advice. <https://edition.cnn.com/2023/06/01/tech/eating-disorder-chatbot/index.html>. Accessed: January 31, 2025.
- [26] Simon Coghlan, Kobi Leins, Susie Sheldrick, Marc Cheong, Piers Gooding, and Simon D'Alfonso. 2023. To chat or bot to chat: Ethical issues with using chatbots in mental health. *Digital health* 9 (2023), 20552076231183542.
- [27] Naver cop. 2018. South Korean Online Social Support Community for People with Eating Disorders. <https://cafe.naver.com/jahayun>. Accessed: January 31, 2025.
- [28] Sophie Corbett. 2023. Extreme Hunger in ED Recovery. <https://www.mentalhealthdietitians.com/extreme-hunger-in-ed-recovery/>. Accessed: January 31, 2025.
- [29] Marguerite Corvini, Casey Golomski, and John Burns. 2024. The Impact of Sharing Recovery Stories in Public: Stigma, Trauma Response, and the Need for Multiple Pathways. *Journal of Social Service Research* 50, 3 (2024), 481–493.
- [30] Jennifer Couturier, Melissa Kimber, and Peter Szatmari. 2013. Efficacy of family-based treatment for adolescents with eating disorders: A systematic review and meta-analysis. *International Journal of Eating Disorders* 46, 1 (2013), 3–11.
- [31] Michele L Crossley. 2000. Narrative psychology, trauma and the study of self-identity. *Theory & Psychology* 10, 4 (2000), 527–546.
- [32] Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting depression via social media. In *Proceedings of the international AAAI conference on web and social media*, Vol. 7. 128–137.
- [33] Julian De Freitas, Ahmet Kaan Uğuralp, Zeliha Oğuz-Uğuralp, and Stefano Puntoni. 2024. Chatbots and mental health: Insights into the safety of generative AI. *Journal of Consumer Psychology* 34, 3 (2024), 481–491.
- [34] SM De la Rie, G Noordenbos, and EF Van Furth. 2005. Quality of life and eating disorders. *Quality of life research* 14 (2005), 1511–1521.
- [35] Kerstin Denecke, Alaa Abd-Alrazaq, and Mowafa Househ. 2021. Artificial intelligence for chatbots in mental health: opportunities and challenges. *Multiple perspectives on artificial intelligence in healthcare: Opportunities and challenges* (2021), 115–128.
- [36] Anjali Devakumar, Jay Modh, Bahador Saket, Eric PS Baumer, and Munmun De Choudhury. 2021. A review on strategies for data collection, reflection, and communication in eating disorder apps. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–19.
- [37] Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).
- [38] AE Dingemans, MJ Bruna, and EF Van Furth. 2002. Binge eating disorder: a review. *International journal of obesity* 26, 3 (2002), 299–307.
- [39] eternnoir. 2024. pyTelegramBotAPI 4.22.1. <https://pypi.org/project/pyTelegramBotAPI>. Accessed: January 31, 2025.
- [40] Christopher G Fairburn and Zafra Cooper. 2011. Eating disorders, DSM–5 and clinical reality. *The British journal of psychiatry* 198, 1 (2011), 8–10.
- [41] Christopher G Fairburn, Zafra Cooper, Helen A Doll, Patricia Norman, and Marianne O'Connor. 2000. The natural course of bulimia nervosa and binge eating disorder in young women. *Archives of general psychiatry* 57, 7 (2000), 659–665.
- [42] Christopher G Fairburn and David M Garner. 1986. The diagnosis of bulimia nervosa. *International Journal of Eating Disorders* 5, 3 (1986), 403–419.
- [43] Jessica L Feuston, Michael Ann DeVito, Morgan Klaus Scheuerman, Katy Weathington, Marianna Benitez, Bianca Z Perez, Lucy Sondheim, and Jed R Brubaker. 2022. "Do You Ladies Relate?": Experiences of Gender Diverse People in Online Eating Disorder Communities. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–32.
- [44] Ellen E Fitzsimmons-Craft, William W Chan, Arielle C Smith, Marie-Laure Firebaugh, Lauren A Fowler, Naira Topooco, Bianca DePietro, Denise E Wilfley, C Barr Taylor, and Nicholas C Jacobson. 2022. Effectiveness of a chatbot for eating disorders prevention: a randomized clinical trial. *International Journal of Eating Disorders* 55, 3 (2022), 343–353.
- [45] Iason Gabriel, Arianna Manzini, Geoff Keeling, Lisa Anne Hendricks, Verena Rieser, Hasan Iqbal, Nenad Tomašev, Ira Ktena, Zachary Kenton, Mikel Rodriguez, et al. 2024. The ethics of advanced ai assistants. *arXiv preprint arXiv:2404.16244* (2024).
- [46] Saadia Gabriel, Isha Puri, Xuhai Xu, Matteo Malgaroli, and Marzyeh Ghassemi. 2024. Can AI Relate: Testing Large Language Model Response for Mental Health Support. *arXiv preprint arXiv:2405.12021* (2024).
- [47] Liza Gak, Seyi Olojo, and Niloufar Salehi. 2022. The distressing ads that persist: Uncovering the harms of targeted weight-loss ads among users with histories of disordered eating. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–23.
- [48] Marie Galmiche, Pierre Déchelotte, Grégory Lambert, and Marie Pierre Tavolacci. 2019. Prevalence of eating disorders over the 2000–2018 period: a systematic literature review. *The American journal of clinical nutrition* 109, 5 (2019), 1402–1413.
- [49] David M Garner and Paul E Garfinkel. 1997. *Handbook of treatment for eating disorders*. Guilford Press.
- [50] Suyu Ge, Chunting Zhou, Rui Hou, Madian Khabisa, Yi-Chia Wang, Qifan Wang, Jiawei Han, and Yuning Mao. 2023. Mart: Improving llm safety with multi-round automatic red-teaming. *arXiv preprint arXiv:2311.07689* (2023).
- [51] Scott Glassman, Petra Kottsieper, Allan Zuckoff, and Elizabeth A. Gosch. 2013. Motivational interviewing and recovery: experiences of hope, meaning, and empowerment. *Advances in dual diagnosis* 6, 3 (2013), 106–120.
- [52] Garth Graham. 2023. An updated approach to eating disorder-related content. <https://blog.youtube/news-and-events/an-updated-approach-to-eating-disorder-related-content/>. Accessed: January 31, 2025.
- [53] Nick Grey, Suzanne Byrne, Tracey Taylor, Avi Shmueli, Cathy Troupp, Peter Stratton, Aaron Sefi, Roslyn Law, and Mick Cooper. 2018. Goal-oriented practice across therapies. (2018).
- [54] Samantha L Hahn, Katherine W Bauer, Niko Kaciroti, Daniel Eisenberg, Sarah K Lipson, and Kendrin R Sonnevile. 2021. Relationships between patterns of weight-related self-monitoring and eating disorder symptomology among undergraduate and graduate students. *International Journal of Eating Disorders* 54, 4 (2021), 595–605.
- [55] Katrin Hartwig, Tom Biselli, Franziska Schneider, and Christian Reuter. 2024. From Adolescents' Eyes: Assessing an Indicator-Based Intervention to Combat Misinformation on TikTok. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–20.
- [56] 2023 WoeBot Health. 2023. WoeBot Health. <https://woebothealth.com>. Accessed: January 31, 2025.
- [57] Woebot Health. 2024. Woebot for Adults. Instructions for Use. [https://woebothealth.com/img/2024/07/Woebot-for-Adults\\_Instructions-for-Use\\_Users\\_July-18th-2024.pdf](https://woebothealth.com/img/2024/07/Woebot-for-Adults_Instructions-for-Use_Users_July-18th-2024.pdf). Accessed: 2024-11-27.
- [58] Sharon Hillege, Barbara Beale, and Rose McMaster. 2006. Impact of eating disorders on family life: Individual parents' stories. *Journal of Clinical Nursing* 15, 8 (2006), 1016–1022.
- [59] Tiancheng Hu and Nigel Collier. 2024. Quantifying the persona effect in llm simulations. *arXiv preprint arXiv:2402.10811* (2024).
- [60] Wenyue Hua, Xianjun Yang, Mingyu Jin, Zelong Li, Wei Cheng, Ruixiang Tang, and Yongfeng Zhang. 2024. TrustAgent: Towards Safe and Trustworthy LLM-based Agents. In *Findings of the Association for Computational Linguistics: EMNLP 2024*. 10000–10016.
- [61] Jina Huh. 2015. Clinical questions in online health communities: the case of "See your doctor" threads. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. 1488–1499.
- [62] Martin Huschens, Martin Briesch, Dominik Sobania, and Franz Rothlauf. 2023. Do You Trust ChatGPT?—Perceived Credibility of Human and AI-Generated Content. *arXiv preprint arXiv:2309.02524* (2023).
- [63] Hilary Hutchinson, Wendy Mackay, Bo Westerlund, Benjamin B Bederson, Allison Druin, Catherine Plaisant, Michel Beaudouin-Lafon, Stéphane Conversy, Helen Evans, Heiko Hansen, et al. 2003. Technology probes: inspiring design for and with families. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 17–24.
- [64] Farnaz Jahanbakhsh and David R Karger. 2024. A Browser Extension for In-place Signaling and Assessment of Misinformation. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–21.
- [65] Ziwei Ji, Tiezheng Yu, Yan Xu, Nayeon Lee, Etsuko Ishii, and Pascale Fung. 2023. Towards mitigating LLM hallucination via self reflection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. 1827–1843.
- [66] Eunhyung Jo, Yui 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*. 1–21.
- [67] Douglas Johnson, Rachel Goodman, J Patrinely, Cosby Stone, Eli Zimmerman, Rebecca Donald, Sam Chang, Sean Berkowitz, Avni Finn, Eiman Jahangir, et al. 2023. Assessing the accuracy and reliability of AI-generated medical responses: an evaluation of the Chat-GPT model. *Research square* (2023).
- [68] KakaoTalk. 2022. KakaoTalk Online Social Support Chatroom for People with Eating Disorders. <https://open.kakao.com/o/guHeu70b>. Accessed: 2023-07-15.
- [69] Jess Kerr-Gaffney, Amy Harrison, and Kate Tchanturia. 2018. Social anxiety in the eating disorders: a systematic review and meta-analysis. *Psychological medicine* 48, 15 (2018), 2477–2491.
- [70] Ronald C Kessler, Patricia A Berglund, Wai Tat Chiu, Anne C Deitz, James I Hudson, Victoria Shahly, Sergio Aguilar-Gaxiola, Jordi Alonso, Matthias C

- Angermeyer, Corina Benjet, et al. 2013. The prevalence and correlates of binge eating disorder in the World Health Organization World Mental Health Surveys. *Biological psychiatry* 73, 9 (2013), 904–914.
- [71] Zoha Khawaja and Jean-Christophe Bélisle-Pipon. 2023. Your robot therapist is not your therapist: understanding the role of AI-powered mental health chatbots. *Frontiers in Digital Health* 5 (2023), 1278186.
- [72] Jennifer G Kim, Hwajung Hong, and Karrie Karahalios. 2018. Understanding identity presentation in medical crowdfunding. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–12.
- [73] 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*. 1–20.
- [74] René F Kizilcec. 2016. How much information? Effects of transparency on trust in an algorithmic interface. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. 2390–2395.
- [75] Zeljko Kraljevic, Anthony Shek, Daniel Bean, Rebecca Bendayan, James Teo, and Richard Dobson. 2021. MedGPT: Medical concept prediction from clinical narratives. *arXiv preprint arXiv:2107.03134* (2021).
- [76] Harsh Kumar, Ruiwei Xiao, Benjamin Lawson, Ilya Musabirov, Jiakai Shi, Xinyuan Wang, Huayin Luo, Joseph Jay Williams, Anna N Rafferty, John Stamper, et al. 2024. Supporting Self-Reflection at Scale with Large Language Models: Insights from Randomized Field Experiments in Classrooms. In *Proceedings of the Eleventh ACM Conference on Learning@Scale*. 86–97.
- [77] Tin Lai, Yukun Shi, Zicong Du, Jiajie Wu, Ken Fu, Yichao Dou, and Ziqi Wang. 2023. Psy-llm: Scaling up global mental health psychological services with ai-based large language models. *arXiv preprint arXiv:2307.11991* (2023).
- [78] Tin Lai, Yukun Shi, Zicong Du, Jiajie Wu, Ken Fu, Yichao Dou, and Ziqi Wang. 2023. Supporting the Demand on Mental Health Services with AI-Based Conversational Large Language Models (LLMs). *BioMedInformatics* 4, 1 (2023), 8–33.
- [79] Andrea LaMarre and Carla Rice. 2021. Healthcare providers' engagement with eating disorder recovery narratives: opening to complexity and diversity. *Medical Humanities* 47, 1 (2021), 78–86.
- [80] Minha Lee, Sander Ackermans, Nena Van As, Hanwen Chang, Enzo Lucas, and Wijnand IJsselstein. 2019. Caring for Vincent: a chatbot for self-compassion. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–13.
- [81] 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*. 1–12.
- [82] Nancy G Leveson. 2016. *Engineering a safer world: Systems thinking applied to safety*. The MIT Press.
- [83] Q Vera Liao and Jennifer Wortman Vaughan. 2023. AI transparency in the age of llms: A human-centered research roadmap. *arXiv preprint arXiv:2306.01941* (2023), 5368–5393.
- [84] José Francisco López-Gil, Antonio García-Hermoso, Lee Smith, Joseph Firth, Mike Trott, Arthur Eumann Mesas, Estela Jimenez-Lopez, Hector Gutierrez-Espinoza, Pedro J Tarraga-Lopez, and Desiree Victoria-Montesinos. 2023. Global proportion of disordered eating in children and adolescents: A systematic review and meta-analysis. *JAMA pediatrics* 177, 4 (2023), 363–372.
- [85] Kate Loveys, Catherine Hiko, Mark Sagar, Xueyuan Zhang, and Elizabeth Broadbent. 2022. "I felt her company": A qualitative study on factors affecting closeness and emotional support seeking with an embodied conversational agent. *International Journal of Human-Computer Studies* 160 (2022), 102771.
- [86] Ryan Lowe, Nissan Pow, Iulian Vlad Serban, Laurent Charlin, Chia-Wei Liu, and Joelle Pineau. 2017. Training end-to-end dialogue systems with the ubuntu dialogue corpus. *Dialogue & Discourse* 8, 1 (2017), 31–65.
- [87] Wysa Ltd. 2023. Wysa. <https://www.wysa.com>. Accessed: January 31, 2025.
- [88] Kai Lukoff, Taoxi Li, Yuan Zhuang, and Brian Y Lim. 2018. TableChat: mobile food journaling to facilitate family support for healthy eating. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–28.
- [89] Zilin Ma, Yiyang Mei, Yinru Long, Zhaoyuan Su, and Krzysztof Z Gajos. 2024. Evaluating the Experience of LGBTQ+ People Using Large Language Model Based Chatbots for Mental Health Support. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–15.
- [90] 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. In *AMIA Annual Symposium Proceedings*, Vol. 2023. American Medical Informatics Association, 1105.
- [91] Lena Mamykina, Andrew D Miller, Elizabeth D Mynatt, and Daniel Greenblatt. 2010. Constructing identities through storytelling in diabetes management. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1203–1212.
- [92] Mark Matthews and Gavin Doherty. 2011. My mobile story: therapeutic storytelling for children. In *CHI'11 Extended Abstracts on Human Factors in Computing Systems*. 2059–2064.
- [93] Dan P McAdams. 1993. The stories we live by: Personal myths and the making of the self. *William Morrow* (1993).
- [94] Nat McAleese, Rai Michael Pokorny, Juan Felipe Ceron Uribe, Evgenia Nitishinskaya, Maja Trebacz, and Jan Leike. 2024. Llm critics help catch llm bugs. *arXiv preprint arXiv:2407.00215* (2024).
- [95] Scott Monteith, Tasha Glenn, John R Geddes, Peter C Whybrow, Eric Achtyes, and Michael Bauer. 2024. Artificial intelligence and increasing misinformation. *The British Journal of Psychiatry* 224, 2 (2024), 33–35.
- [96] Subigya Nepal, Arvind Pillai, William Campbell, Talie Massachi, Eunsol Soul Choi, Xuhai Xu, Joanna Kuc, Jeremy F Huckins, Jason Holden, Colin Depp, et al. 2024. Contextual AI Journaling: Integrating LLM and Time Series Behavioral Sensing Technology to Promote Self-Reflection and Well-being using the Mind-Scape App. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–8.
- [97] Jingping Nie, Hanya Shao, Yuang Fan, Qijia Shao, Haoxuan You, Matthias Preindl, and Xiaofan Jiang. 2024. LLM-based Conversational AI Therapist for Daily Functioning Screening and Psychotherapeutic Intervention via Everyday Smart Devices. *arXiv preprint arXiv:2403.10779* (2024).
- [98] Marcia Nißen, Dominik Rüeegger, Mirjam Stieger, Christoph Flückiger, Mathias Allemann, Florian v Wangenheim, and Tobias Kowatsch. 2022. The effects of health care chatbot personas with different social roles on the client-chatbot bond and usage intentions: development of a design codebook and web-based study. *Journal of medical Internet research* 24, 4 (2022), e32630.
- [99] Kate P Nurse, Imogen Rushworth, Tom Shakespeare, and Deirdre Williams. 2018. Personal storytelling in mental health recovery. *Mental Health Review Journal* 23, 1 (2018), 25–36.
- [100] Jooyoung Oh, Sooh Jang, Hyunji Kim, and Jae-Jin Kim. 2020. Efficacy of mobile app-based interactive cognitive behavioral therapy using a chatbot for panic disorder. *International journal of medical informatics* 140 (2020), 104171.
- [101] Yoo Jung Oh, Jingwen Zhang, Min-Lin Fang, and Yoshimi Fukuoka. 2021. A systematic review of artificial intelligence chatbots for promoting physical activity, healthy diet, and weight loss. *International Journal of Behavioral Nutrition and Physical Activity* 18 (2021), 1–25.
- [102] OpenAI. 2024. OpenAI Privacy Policy. <https://openai.com/policies/row-privacy-policy>. Accessed: January 31, 2025.
- [103] OpenAI. 2024. OpenAI Usage Policies. <https://openai.com/policies/usage-policies/>. Accessed: 2024-11-27.
- [104] P Parmar, J Ryu, S Pandya, J Sedoc, and S Agarwal. 2022. Health-focused conversational agents in person-centered care: A review of apps. *npj Digital Medicine* 5, 21.
- [105] Krisna Patel, Kate Tchanturia, and Amy Harrison. 2016. An exploration of social functioning in young people with eating disorders: a qualitative study. *PLoS one* 11, 7 (2016), e0159910.
- [106] Jessica Pater, Fayika Farhat Nova, Amanda Coupe, Lauren E Reining, Connie Kerrigan, Tammy Toscos, and Elizabeth D Mynatt. 2021. Charting the unknown: Challenges in the clinical assessment of patients' technology use related to eating disorders. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–14.
- [107] Manjiri Pawaskar, Edward A Witt, Dylan Supina, Barry K Herman, and Thomas A Wadden. 2017. Impact of binge eating disorder on functional impairment and work productivity in an adult community sample in the United States. *International journal of clinical practice* 71, 7 (2017), e12970.
- [108] Janet Polivy and C Peter Herman. 2002. Causes of eating disorders. *Annual review of psychology* 53, 1 (2002), 187–213.
- [109] Janet Polivy and C Peter Herman. 2004. Sociocultural idealization of thin female body shapes: An introduction to the special issue on body image and eating disorders. 6 pages.
- [110] Katie Prizeman, Netta Weinstein, and Ciara McCabe. 2023. Effects of mental health stigma on loneliness, social isolation, and relationships in young people with depression symptoms. *BMC psychiatry* 23, 1 (2023), 527.
- [111] Rebecca Puhl and Young Suh. 2015. Stigma and eating and weight disorders. *Current psychiatry reports* 17 (2015), 1–10.
- [112] Taivo Pungas. 2023. GPT-3.5 and GPT-4 response times. [https://www.taivo.ai/\\_gpt-3-5-and-gpt-4-response-times/](https://www.taivo.ai/_gpt-3-5-and-gpt-4-response-times/). Accessed: January 31, 2025.
- [113] Matthew Renze and Erhan Guven. 2024. Self-Reflection in LLM Agents: Effects on Problem-Solving Performance. *arXiv preprint arXiv:2405.06682* (2024).
- [114] MIT Technology Review. 2024. A chatbot helped more people access mental-health services. <https://www.technologyreview.com/2024/02/05/1087690/a-chatbot-helped-more-people-access-mental-health-services/>. Accessed: January 31, 2025.
- [115] Deborah Richards, Ayse Aysin Bilgin, and Hedieh Ranjartabar. 2018. Users' perceptions of empathic dialogue cues: A data-driven approach to provide tailored empathy. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents*. 35–42.
- [116] Shalaleh Rismani, Renee Shelby, Andrew Smart, Edgar Jatho, Joshua Kroll, AJung Moon, and Negar Rostamzadeh. 2023. From plane crashes to algorithmic harm: applicability of safety engineering frameworks for responsible ML. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*.

- 1–18.
- [117] S Roller. 2020. Recipes for building an open-domain chatbot. *arXiv preprint arXiv:2004.13637* (2020).
- [118] Sivan Schwartz, Avi Ya'eli, and Segev Shlomov. 2023. Enhancing trust in LLM-based AI automation agents: New considerations and future challenges. *arXiv preprint arXiv:2308.05391* (2023).
- [119] Jillian Shah, Bianca DePietro, Laura D'Adamo, Marie-Laure Firebaugh, Olivia Laing, Lauren A Fowler, Lauren Smolar, Shiri Sadeh-Sharvit, C Barr Taylor, Denise E Wilfley, et al. 2022. Development and usability testing of a chatbot to promote mental health services use among individuals with eating disorders following screening. *International Journal of Eating Disorders* 55, 9 (2022), 1229–1244.
- [120] Inhwa Song, Sachin R Pendse, Neha Kumar, and Munmun De Choudhury. 2024. The typing cure: Experiences with large language model chatbots for mental health support. *arXiv preprint arXiv:2401.14362* (2024).
- [121] Rose Stackpole, Danyelle Greene, Elizabeth Bills, and Sarah J Egan. 2023. The association between eating disorders and perfectionism in adults: A systematic review and meta-analysis. *Eating behaviors* 50 (2023), 101769.
- [122] Sharon Stovezky. 2021. You are not alone. <https://blog.youtube/news-and-events/you-are-not-alone/>. Accessed: January 31, 2025.
- [123] Richard Sutcliffe. 2023. A Survey of Personality, Persona, and Profile in Conversational Agents and Chatbots. *arXiv preprint arXiv:2401.00609* (2023).
- [124] I Sutskever. 2014. Sequence to Sequence Learning with Neural Networks. *arXiv preprint arXiv:1409.3215* (2014).
- [125] Telegram FZ LLC and Telegram Messenger Inc. [n. d.]. *Telegram*. <https://telegram.org>
- [126] The New York Times. 2023. A Wellness Chatbot Is Offline After Its 'Harmful' Focus on Weight Loss. <https://www.nytimes.com/2023/06/08/us/ai-chatbot-tessa-eating-disorders-association.html>. Accessed: January 31, 2025.
- [127] Tracy L Tylka and Ashley M Kroon Van Diest. 2013. The Intuitive Eating Scale–2: Item refinement and psychometric evaluation with college women and men. *Journal of counseling psychology* 60, 1 (2013), 137.
- [128] A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems* (2017).
- [129] Tara Wadhwa. 2021. Supporting #NEDAwareness and body inclusivity on TikTok. <https://newsroom.tiktok.com/en-us/supporting-nedawareness-and-body-inclusivity-on-tiktok>. Accessed: January 31, 2025.
- [130] B Timothy Walsh and Michael J Devlin. 1998. Eating disorders: progress and problems. *Science* 280, 5368 (1998), 1387–1390.
- [131] Joel Wester, Tim Schrills, Henning Pohl, and Niels van Berkel. 2024. “As an AI language model, I cannot”: Investigating LLM Denials of User Requests. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–14.
- [132] WinnieNwanne. 2024. Comparing GPT-3.5 and GPT-4: A Thought Framework on When To Use Each Model. <https://techcommunity.microsoft.com/t5/ai-azure-ai-services-blog/comparing-gpt-3-5-amp-gpt-4-a-thought-framework-on-when-to-use/ba-p/4088645>. Accessed: January 31, 2025.
- [133] Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. 2019. Transfertransfo: A transfer learning approach for neural network based conversational agents. *arXiv preprint arXiv:1901.08149* (2019).
- [134] Siyi Wu, Feixue Han, Bingsheng Yao, Tianyi Xie, Xuan Zhao, and Dakuo Wang. 2024. Sunnie: An Anthropomorphic LLM-Based Conversational Agent for Mental Well-Being Activity Recommendation. *arXiv preprint arXiv:2405.13803* (2024).
- [135] Xuhai Xu, Bingsheng Yao, Yuanzhe Dong, Saadia Gabriel, Hong Yu, James Hendler, Marzyeh Ghassemi, Anind K Dey, and Dakuo Wang. 2024. Mental-llm: Leveraging large language models for mental health prediction via online text data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8, 1 (2024), 1–32.
- [136] Joel Yager and Pauline S Powers. 2008. *Clinical manual of eating disorders*. American Psychiatric Pub.
- [137] Zhiping Zhang, Michelle Jia, Hao-Ping Lee, Bingsheng Yao, Sauvik Das, Ada Lerner, Dakuo Wang, and Tianshi Li. 2024. “It’s a Fair Game”, or Is It? Examining How Users Navigate Disclosure Risks and Benefits When Using LLM-Based Conversational Agents. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–26.
- [138] Yuxin Zhao, Jiawei Wu, Ping Qu, Beibei Zhang, and Hao Yan. 2024. Assessing User Trust in LLM-based Mental Health Applications: Perceptions of Reliability and Effectiveness. *Journal of Computer Technology and Applied Mathematics* 1, 2 (2024), 19–26.
- [139] Zhonghua Zheng, Lizi Liao, Yang Deng, and Liqiang Nie. 2023. Building emotional support chatbots in the era of llms. *arXiv preprint arXiv:2308.11584* (2023).