# Towards Explainable and Safe Conversational Agents for Mental Health: A Survey

**Surjodeep Sarkar**[1] , **Manas Gaur**[1] , **Lujie Karen Chen**[1] ,
**Muskan Garg**[2] , **Biplav Srivastava**[3] , **Bhaktee Dongaonkar**[4]
[1]UMBC, MD, USA ; [2]Mayo Clinic, MN, USA; [3]AI Institute, USC, SC, USA: [4]IIIT Hyderabad, India
{ssarkar1, manas, lujiec}@umbc.edu, biplav.s@sc.edu,
garg.muskan@mayo.edu, bhaktee.dongaonkar@iiit.ac.in

## Abstract

Virtual Mental Health Assistants (VMHA) are seeing continual advancements to support the overburdened global healthcare system that gets 60 million primary care visits, and 6 million Emergency Room (ER) visits annually. These systems are built by clinical psychologists, psychiatrists, and Artificial Intelligence (AI) researchers for Cognitive Behavioral Therapy (CBT). At present, the role of VMHAs is to provide emotional support through information, focusing less on developing a reflective conversation with the patient. A more *comprehensive, safe* and *explainable* approach is required to build *responsible* VMHAs to ask follow-up questions or provide a well-informed response. This survey offers a systematic critical review of the existing conversational agents in mental health, followed by new insights into the improvements of VMHAs with contextual knowledge, datasets, and their emerging role in clinical decision support. We also provide new directions toward enriching the user experience of VMHAs with explainability, safety, and wholesome trustworthiness. Finally, we provide evaluation metrics and practical considerations for VMHAs beyond the current literature to build trust between VMHAs and patients in active communications.

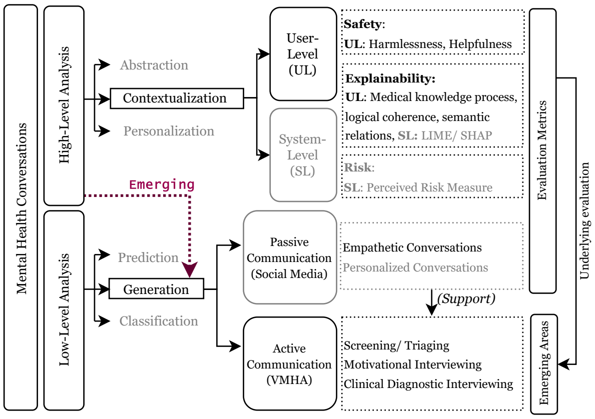

Mental Health Conversations
High-Level Analysis
Abstraction
Contextualization
Personalization
User-Level (UL)
System-Level (SL)
Safety:
UL: Harmlessness, Helpfulness
Explainability:
UL: Medical knowledge process, logical coherence, semantic relations, SL: LIME/ SHAP
Risk:
SL: Perceived Risk Measure
Evaluation Metrics
Underlying evaluation
Emerging
Low-Level Analysis
Prediction
Generation
Classification
Passive Communication (Social Media)
Empathetic Conversations
Personalized Conversations
(Support)
Active Communication (VMHA)
Screening/ Triaging
Motivational Interviewing
Clinical Diagnostic Interviewing
Emerging Areas

Figure 1: Taxonomy of Mental Health Conversations: While connecting dots in our investigation from NLP-centered low-level analysis (lexical, morphological, syntactic, semantic) over Mental health conversations to the higher-level analysis (discourses, pragmatics), we determine the evaluation metrics to support VMHAs for better user-level experience in terms of safety and explainability. We further support the emerging areas with AI model development and evaluation in passive conversations. The categories in black color defines the scope of our survey from the view point of user-level explainability and safety; dotted red colour highlights the emerging scope of question/response generation in mental health conversations between VHMAs and patients.

## 1 Introduction

Mental illness is highly prevalent nowadays, constituting a major cause of distress in people's lives with an impact on society's health and well-being, thereby projecting serious challenges for mental health professionals (MHPs) [Zhang et. al., 2022]. According to the National Survey on Drug Use and Health, nearly one in five U.S. adults live with a mental illness (52.9 million in 2020) [SAM, 2020]. Reports released in August 2021[1] indicate that *1.6 million people* in England were on waiting lists to seek professional help with mental health care. Such an overwhelming rise in the number of patients as compared to MHPs necessitated the use of (i) public health forums (e.g., dialogue4health), (ii) online communities (e.g., r/depression subreddit on Reddit), (iii) Talklife, and (iv) Virtual Mental Health Assistants (VMHAs), for informative healthcare. The anonymous functioning of (i), (ii), (iii) removed the psychological stigma in patients, which even refrained them from seeing an MHP [Hyman et al., 2008].

In addition, the unavailability of interpersonal interactions from other pure information agents resulted in the need to develop Virtual Mental Health Assistants (VMHAs).

**VMHAs**: VMHAs are artificial intelligence (AI)-based agents designed to provide emotional support through structured conversational sequences targeted to screen patients for mental health conditions and alert mental health professionals (MHPs) through *informed triaging*[2]. Despite the proliferation of research at the intersection of clinical psychology, artifi-

[1] https://www.theguardian.com/society/2021/aug/29/strain-on-mental-health-care-leaves-8m-people-without-help-say-nhs-leaders

[2] https://code4health.org/chat-bot/

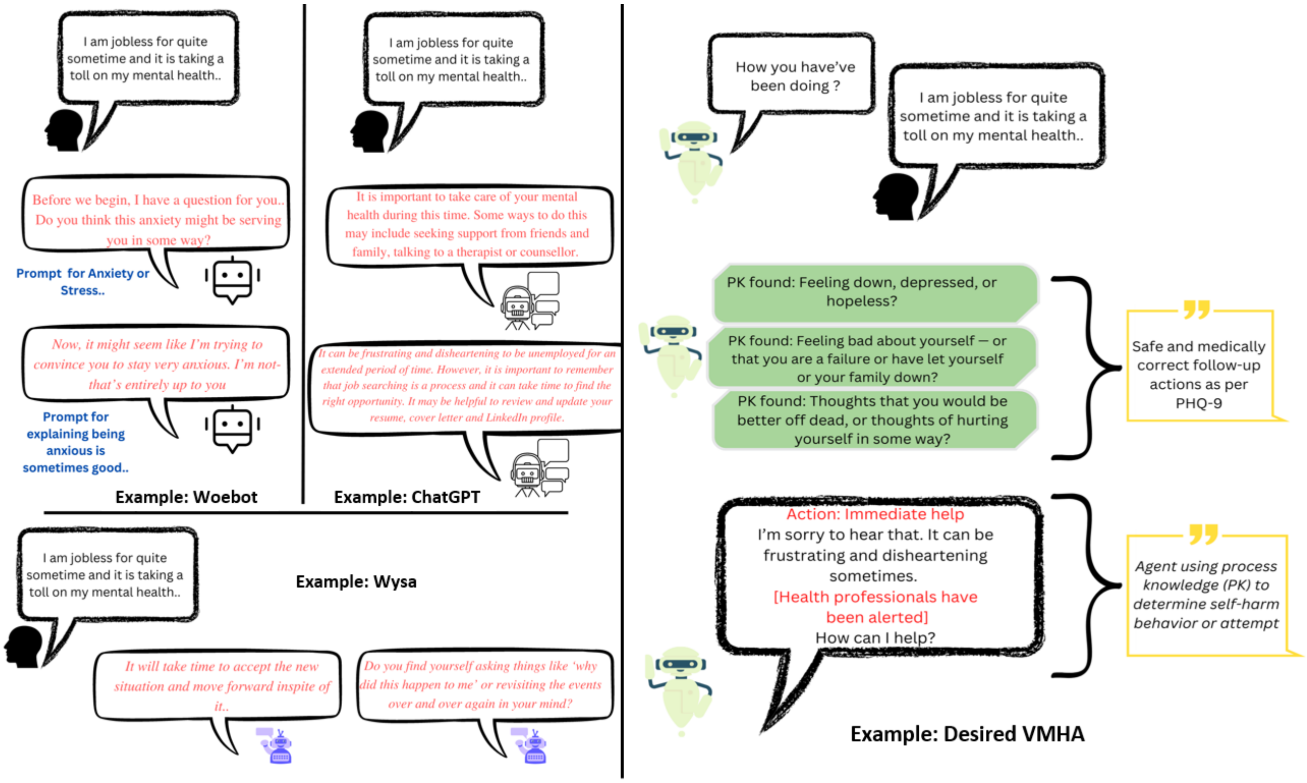

I am jobless for quite sometime and it is taking a toll on my mental health..
Before we begin, I have a question for you.. Do you think this anxiety might be serving you in some way?
Prompt for Anxiety or Stress..
Now, it might seem like I'm trying to convince you to stay very anxious. I'm not- that's entirely up to you
Prompt for explaining being anxious is sometimes good..
Example: Woebot
I am jobless for quite sometime and it is taking a toll on my mental health..
It is important to take care of your mental health during this time. Some ways to do this may include seeking support from friends and family, talking to a therapist or counsellor.
It can be frustrating and disheartening to be unemployed for an extended period of time. However, it is important to remember that job searching is a process and it can take time to find the right opportunity. It may be helpful to review and update your resume, cover letter and LinkedIn profile.
Example: ChatGPT
I am jobless for quite sometime and it is taking a toll on my mental health..
Example: Wysa
It will take time to accept the new situation and move forward inspite of it..
Do you find yourself asking things like 'why did this happen to me' or revisiting the events over and over again in your mind?
How you have've been doing ?
I am jobless for quite sometime and it is taking a toll on my mental health..
PK found: Feeling down, depressed, or hopeless?
PK found: Feeling bad about yourself — or that you are a failure or have let yourself or your family down?
PK found: Thoughts that you would be better off dead, or thoughts of hurting yourself in some way?
Safe and medically correct follow-up actions as per PHQ-9
Action: Immediate help
I'm sorry to hear that. It can be frustrating and disheartening sometimes.
[Health professionals have been alerted]
How can I help?
Agent using process knowledge (PK) to determine self-harm behavior or attempt
Example: Desired VMHA

Figure 2: (Left) The outcome from existing VMHAs (e.g., WoeBot, Wysa) and ChatGPT (general purpose chatbot). (Right) Illustration of a knowledge-driven conversational agent in mental health (desired VMHA). The use of questions in PHQ-9 to induce conceptual flow in mental health conversational agents. With clinical knowledge, the agent can detect the user's mental disturbance and alert MHPs accordingly.

cial intelligence (AI), and Natural Language Understanding (NLU), VMHAs missed an opportunity to serve as life-saving contextualized, personalized, and reliable decision support during COVID-19 under the *apollo* moment [Srivastava et al., 2021; Czeisler et al., 2020]. VMHAs' ability to function as simple information agents (e.g., suggest meditation, relaxation exercises, or give positive affirmations) *did not* bridge the gap between *monitoring the health condition* and *necessitating an MHP visit* for the patient.

To the best of our knowledge, this is the first critical evaluation that examines contextualization and question/response generation for VMHAs from the viewpoint of user-level explainability and trust (see Figure 1). *This survey facilitates the clinical psychologists, psychiatrists, and AI practitioners of VMHAs to support people at risk of chronic mental disease.*

**User-level Explainability**: The sensitive nature of VMHAs raises *safety* as a major concern of conversational systems, resulting in a negative outcome. For instance, figure 2 presents a real-world query from a user, which was common during the times of the COVID-19 recession. In response to the query-Woebot, Wysa and ChatGPT initiated a responsive conversation without focusing on the context (e.g., connecting mental health with its symptoms). We found assumptive questions (e.g., anxiety) and responses from Wysa, Woebot and ChatGPT with no association to clinical reference or clinical support. On the other hand, the desired VMHA (a) should capture the relationship between the user query and expert questionnaires and (b) tailor the response to reflect on the user's concerns (e.g., *frustrating* and *disheartening*) about the *long-term unemployment*, which is linked to *mental health* and *user immediate help*.

**Resources to support VMHA**: Prior research demonstrate extensive body of efforts in developing *mental health datasets* using social media to identify mental health conditions [Uban et. al., 2021]. These datasets represent real-world conversations and are annotated by experts leveraging clinically-grounded knowledge (e.g., MedChatbot [Kazi *et al.*, 2012]) or guidelines (e.g., PHQ-9). Augmenting such datasets with VMHAs can improve the quality of conversations with the user. Semantic enhancements with clinical knowledge and associated guidelines, if remain underexplored, may miss the hidden mental states in a given narrative which is an essential component of question generation.

**Trustworthiness**: By definition, *Trust* is a multi-faceted quality that is studied in the context of humans in humanities and now increasingly gaining importance in AI as systems and humans collaborate closely. Growing concern about (misplaced) *trust* on *VMHA* for *Social Media* (tackling mental health) hampers the adoption of AI techniques during emergency situations like COVID-19 [Srivastava et al., 2021]. A recent surge in the use of ChatGPT, in particular for mental health, is emergent for providing crucial personalized advice without clinical explanation, which might hurt user's *safety*, and thus, *trust* [3]. In [Varshney et al., 2022], the author identifies the support for human interaction and explainable alignment with human values as important for trust in AI systems.

To holistically contribute towards *trustworthy* behavior in a conversational system in mental health, there is a need to critically examine *user-level explainability*, *safety*, the use of clinical knowledge for contextualization, along with testing.

**Our Contributions**: This survey spans 5 major research dimensions: (i) What are explainability and safety in VMHAs? (ii) What are the current capabilities and limitations of VMHA?, (iii) What is the current state of AI and the hurdles in supporting VMHAs? (iv) What functionalities can be imagined in VMHA for which patients seek alternative solutions? and (v) What changes in evaluation is required with respect to explainability, safety, and trust? Figure 1 illustrates the survey coverage, exemplified in Figure 2.

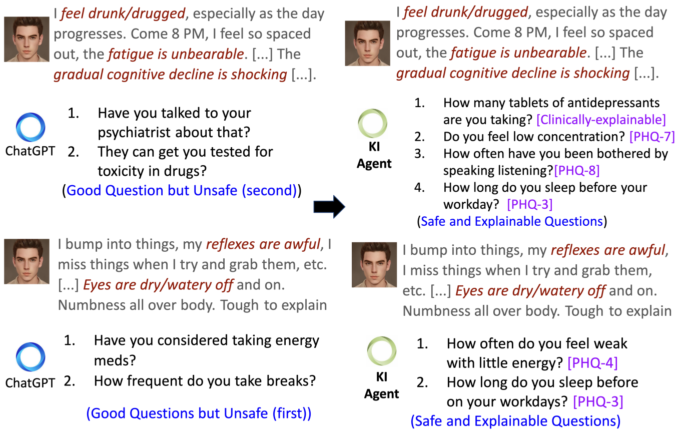
## 2 Scope of Survey

In this section, we explore the state of research in explainability and safety in conversational systems to ensure trust [Hoffman et al., 2018].

### 2.1 Explanation

Conversations in AI happen through large and complex language models (e.g., GPT-3, ChatGPT), which are established as state-of-the-art models for developing intelligent agents to chat with the users by generating human-like questions or responses. The reasons behind the output generated by the Large Language Models (LLM) are unclear and hard to interpret, also known as the "*black box*" effect. The consequences of the black box effect are more concerning than their utility, particularly in mental health. Figure 3 presents a scenario where ChatGPT advises the user about *toxicity in drugs*, which may have a negative consequence. To this end, [Bommasani et al., 2021] reports hallucination and harmful question generations as unexpected behaviors shown by such black box models. The study characterizes *hallucination* as a generated content that *deviates* significantly from the subject matter or is unreasonable. Recently, Replika, a VMHA, augmented with a GPT-3, provides meditative suggestions to a user expressing self-harm tendencies[4]. The analysis above supports the critical need for a comprehensive and explainable approach toward the decision-making of VMHAs. According to [Weick et. al., 1995], the explanations are human-centered sentences that signify the reason or justification behind an action and are comprehensible to a human expert. There are many types of explanations [Longo *et al.*, 2020] and surveys of deployed systems [Bhatt *et al.*, 2020] has revealed that most are targeted towards model developers and not the end-users. The users interacting with the VMHAs may need more systematic information than just the decision-making. Thus, this survey is more focused towards "*User-level Explainability*".

**User-level Explainability (UsEx)** *is defined as the capability of an AI methodology to provide a post-hoc explanation upon the need of a user and in the form of traceable links to real-world entities and definitions [Gaur et al., 2022b].*


I *feel drunk/drugged*, especially as the day progresses. Come 8 PM, I feel so spaced out, the *fatigue is unbearable*. [...] The *gradual cognitive decline is shocking* [...].

1. Have you talked to your psychiatrist about that?
2. They can get you tested for toxicity in drugs?

(Good Question but Unsafe (second))

I bump into things, my *reflexes are awful*, I miss things when I try and grab them, etc. [...] *Eyes are dry/watery off* and on. Numbness all over body. Tough to explain

1. Have you considered taking energy meds?
2. How frequent do you take breaks?

(Good Questions but Unsafe (first))

I *feel drunk/drugged*, especially as the day progresses. Come 8 PM, I feel so spaced out, the *fatigue is unbearable*. [...] The *gradual cognitive decline is shocking* [...].

1. How many tablets of antidepressants are you taking? [Clinically-explainable]
2. Do you feel low concentration? [PHQ-7]
3. How often have you been bothered by speaking listening?[PHQ-8]
4. How long do you sleep before your workday? [PHQ-3]

(Safe and Explainable Questions)

I bump into things, my *reflexes are awful*, I miss things when I try and grab them, etc. [...] *Eyes are dry/watery off* and on. Numbness all over body. Tough to explain

1. How often do you feel weak with little energy? [PHQ-4]
2. How long do you sleep before on your workdays? [PHQ-3]

(Safe and Explainable Questions)

Figure 3: A conversational scenario in which a user asks a query with multiple symptoms. Left is a set of generated questions obtained by repetitive prompting ChatGPT. Right is a generation from ALLEVIATE, a knowledge-infused (KI) conversational agent with access to PHQ-9 and clinical knowledge from Mayo Clinic.

Figure 3 illustrates the UsEx wherein the generated follow-up questions from a safe and user-level explainable agent establish semantic connections with clinical guidelines (e.g., PHQ-9). Though UsEx sees promise over foundational general-purpose NLP tasks, its applicability in the mental health context is yet to be examined [Gunaratna et al., 2022].

### 2.2 Safety

VMHAs are required to be predominantly safe while at the same time being explainable to prevent undesirable behaviors. One such method is aligning the functioning of VMHA to MHP-defined specifications [Koulouri et al., 2022]. Such specifications allow VMHAs absolve the control of generating fabricated content and render it unsafe.

[Dinan et al., 2021] identifies three major effects on safety in general-purpose conversational systems: (a) Generating Offensive Content, also known as the *Instigator (Tay) Effect*. It describes the tendencies of a conversational agent to display behaviors like the Microsoft Tay chatbot, which went racial after learning from the internet. (b) *YEA-SAYER (ELIZA)* effect is defined as the response from a conversational agent to an offensive input from the user. People have been proven to be particularly forthcoming about their mental health problems in interactions with conversational agents, which may increase the danger of "*agreeing with user utterances implying self-harm*". (c) *Imposter* effect applies to VMHAs that tend to respond *inappropriately* in sensitive scenarios. To overcome the imposter effect, Deepmind designed *Sparrow*, a conversational agent. It responsibly leverages the live google search to talk with users [Gupta et. al., 2022a]. The agent generates answers by following the *23 rules* determined by researchers, such as *not offering financial advice*, *making threatening statements*, or *claiming to be a person* [Heikkilä et. al., 2022].

In the context of mental health, such rules can be replaced

[3] https://tinyurl.com/4sr2hw9b

[4] https://ineqe.com/2022/01/20/replika-ai-friend/

by clinical specifications to validate the functioning of AI model within the *safe limits*. Source for such specifications are: Systematized Nomenclature of Medicine-Clinical Terms (SNOMED-CT) [Donnelly et al., 2006], International Classification of Diseases (ICD-10) [Quan et al., 2005], Diagnostic Statistical Manual for Mental Health Disorder (DSM-5) [Regier et. al., 2013], Structured Clinical Interviews for DSM-5 (SCID) [First, 2014], and clinical questionnaire-guided lexicons. [Hennemann et al., 2022] performs a comparative study on psychotherapy of outpatients in mental health where an AI model within VMHA aligns to clinical guidelines for easy understanding of domain experts through UsEx.

## 3 Knowledge Infused (KI) Learning for Mental Health Conversations

Machine-readable knowledge can be categorized into five forms: (a) lexical and linguistic, (b) general-purpose (e.g., Wikipedia, Wikidata), (c) commonsense (e.g., ConceptNet), (d) domain-specific (Unified Medical Language System), and (e) procedural or process-oriented (**PK**) [Sheth et. al., 2021]. Knowledge-infused Learning (KIL), a paradigm within AI, defines a set of methodologies incorporating these broader forms of knowledge to address the limitations of current black-box AI. In addition, KiL benefits from data and knowledge to enable safe and explainable operations in mental health [Gaur et al., 2022b]. We categorize the KIL-driven efforts at the intersection of conversational AI and mental health into two categories:

- **Knowledge Graph-guided Conversations:** Question answering using knowledge graph (KG) is seeing tremendous interest from AI and NLU community through various technological improvements in query understanding, query rewriting, knowledge retrieval, question generation, response shaping, and others. The methods proposed can improve the high-level functionalities of VMHA. For instance, [Welivita et. al., 2022b]'s HEAL KG can generate a better empathetic response by capturing empathy, expectations, affect, stressors, and feedback types from distress conversations. With HEAL, the model picks an appropriate phrase in the user's query to tailor its response. EmoKG is another KG that connects BioPortal, SNOMEDCT, RxNORM, MedDRA, and emotion ontologies to have a conversation with a user to boost their mental health with food recommendation [Gyrard et al., 2022]. Likewise, [Cao et al., 2020] created suicide KG to train conversational agents that can sense whether the user interacting has suicidal indication (e.g., relationship issues, family issues) or suicide risk tendencies before yielding a response or asking follow-up questions. [Sheth et al., 2019] explained the evolution of KG in VMHA during a conversation for adaptive communications. Augmentation of KG demands improvement in metrics to examine the safety and user-level explainability through proxy measures such as logical coherence, semantic relations, and others (covered in section 6.1 and [Gaur et al., 2022a]).
- **Lexicon or Process-guided Conversations:** Lexicons in mental health were created to resolve ambiguities in human language. For instance, the following two sentences: "I am feeling on the edge." and "I am feeling anxious," are similar, provided there is a lexicon with "Anxiety" as a category and "feeling on the edge" as its concept. [Yazdavar et al., 2017] created a PHQ-9 lexicon to study realistic mental health conversations on social media clinically. [Roy et al, 2022b] leveraged PHQ-9 and SNOMED-CT lexicons to train a question-generating agent for paraphrasing questions in PHQ-9 to introduce *Diversity in Generation* (**DiG**) [Limsopatham et al., 2016]. With DiG, a VMHA can paraphrase its question to acquire a meaningful response from a user while still keeping engagement. *Clinical specifications*[5](PK) include questionnaires such as PHQ-9 (depression), Columbia Suicide Severity Rating Scale (C-SSRS; suicide), Generalized Anxiety Disorder (GAD-7) [Gaur et al., 2022b]. It provides a sequence of questions clinicians follow to interview patients. Such questions are safe and medically validated. [Noble et al., 2022] developed MIRA, a VMHA with knowledge of clinical specification to meaningfully respond to queries on mental health issues and interpersonal needs during COVID-19. [Miner et al., 2016] leverage Relational Frame Theory (RFT), a procedural knowledge in clinical psychology to capture events between conversations and labels as positive and negative. [Chung et al., 2021] develops KakaoTalk, a chatbot with prenatal and postnatal care knowledge database of Korean clinical assessment questionnaires and responses that enable the VMHA to carry out thoughtful and contextual conversations with users.

Using KGs through mechanisms of KIL can propel context understanding in VMHA for a safe and explainable conversation. Datasets or VMHAs which use mental health-related knowledge (**MK**) as either KG or lexical are marked as ✓ in Tables 1 and 2.

## 4 Safe and Explainable Language Models in Mental Health

Language models (e.g., Blenderbot, DialoGPT) and in-use conversational agents (e.g., Xiaoice, Tay, Siri) were questioned in the context of safety during the *first workshop on safety in conversational AI*. 70% participants in the workshop were unsure of whether present-day conversational systems or language models within them are capable of safe generation. Following it, [Xu et al., 2020] introduced *Bot-Adversarial Dialogue* and *Bot Baked In* methods to introduce *safety* in conversational systems. The study was performed on *Blenderbot*, which had mixed opinions on safety, and *DialoGPT*, to enable AI models to detect unsafe/safe utterances, avoid sensitive topics, and provide responses that are gender-neutral. The study utilizes knowledge from Wikipedia (for offensive words) and knowledge-powered methods to train conversational agents [Dinan et al., 2018]. Alternatively, safety

[5] also called clinical guidelines and clinical process knowledge

| Datasets | | Safety | UsEx | KI | | DiG | FAIR Principle | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | PK | MK | | F | A | I | R |
| [CounselChat, 2015] | CounselChat | ✓ | ✗ | ✗ | ✗ | ✗ | ✓ | ✓ | ✗ | † |
| [Huang et. al., 2015] | CC | ✗ | ✓ | ✗ | ✓ | ✗ | ✓ | ✓ | ✗ | † |
| [Althoff et al., 2016] | SNAP Counseling | ✓ | ✗ | ✗ | ✗ | ✓ | ✗ | ✗ | ✗ | ✗ |
| [Raskin et al., 2019] | Empathetic Dialogues | ✓ | ✗ | ✗ | ✗ | ✓ | ✓ | ✓ | ✓ | ✓ |
| [Demasi et al., 2019] | Roleplay | ✓ | ✓ | ✓ | ✗ | ✓ | ✓ | ✓ | ✗ | ✓ |
| [Liang et al., 2021] | CC-44 | ✗ | ✗ | ✗ | ✗ | ✗ | ✓ | † | ✗ | † |
| [Gupta et al., 2022b] | PRIMATE | ✓ | ✓ | ✓ | ✗ | ✗ | ✓ | ✓ | ✓ | ✓ |
| [Roy et al, 2022b] | ProKnow-data | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| [Welivita et al., 2022a] | MITI | ✓ | ✓ | ✗ | ✗ | ✗ | ✓ | ✓ | ✓ | ✓ |

Table 1: Lists of conversational datasets created with support from MHPs, crisis counselors, nurse practitioners, or trained annotators. We have not included datasets created using crowdsource workers without proper annotation guidelines. KI: Knowledge infusion; PK: Process Knowledge; MK: Medical Knowledge; DiG: Diversity in Generation; UsEx: User-level Explainability. Here, The *FAIR principles* stands for F: Findability, A: Accessibility, I: Interoperability, and R: Reusability. †: partial fulfillment of the corresponding principle.

in conversational systems can be introduced through clinical guidelines. [Roy et al, 2022b] develop safety lexicons from PHQ-9 and GAD-7, for safe and explainable functioning of language models. The study showed an 85% improvement in safety across Sequence to Sequence and Attention-based language models. In addition, explainability saw an uptake of 23% across the same language models. Similar results were observed when PHQ-9 was used in explainable training of language models [Zirikly et. al., 2022]. VMHA can align with clinical guidelines through reinforcement learning. For example, the *policy gradient-based learning* can assist conversational systems in accounting for safe generation, either through specialized datasets on response rewriting [Sharma et al., 2021] or tree-based rewards guided by process knowledge in mental health [Roy et. al., 2022a].

Though there is an initiative to attain safety in conversations from AI-powered agents, an effort is needed to achieve UsEx. In mental health, the indicators of signs and symptoms, causes, disorders, medications, and other comorbid conditions possess probabilistic relationships with one another. Hence, the augmentation of the knowledge base or infusion of knowledge to improve AI's decision-making is crucial to human understandability [Joyce et al., 2023].

## 5 Virtual Mental Health Assistants

Despite the positive potentials of the language models, our observations indicate the in-capabilities of VMHAs to comprehend the behavioral and emotional instability, self-harm tendencies, and user's latent psychological mindset. VMHAs (e.g., as exemplified in Figure 3 and 2) generate incoherent and unsafe responses when a user tries to seek a response for clinically relevant questions. In this section, we outline the capabilities of well-established VMHAs and inspect limitations in the context of UsEx and safety following taxonomy in figure 1.

- WoeBot is introduced as a part of the growing industry of digital mental health space as an "*Automated Coach*" that can deliver a coach-like or sponsor-like experience without the human intervention to facilitate the "*good thinking hygiene*" [6]. WoeBot deploys featuring lessons (via texts and "stories"), interactive exercises, and videos that were tuned around Cognitive Behavioral Therapy (CBT) [Fitzpatrick et al., 2017].
- Wysa, a mental health application, uses CBT conversational agent to have empathetic/ therapeutic conversations and activities thereby helping its users with several mental health problems [Inkster et al., 2018]. Based on a series of question-answering mechanisms, Wysa suggests a set of relaxing activities for elevating mental well-being.

With the historical evolution of VMHAs (see Table 2) from behavioral health coaching [Ginger, 2011] to KG-based intellectual VMHAs such as ALLEVIATE [Roy et. al., 2023], we examine the possibilities of new research directions to facilitate the expression of empathy in passive communications [Sharma et. al., 2023]. The existing studies suggest the risk of oversimplification of mental conditions and therapeutic approaches without considering latent or external contextual knowledge [Cirille et al., 2020]. Thinking beyond the low-level analysis of classification and prediction, the high-level analysis of VMHAs would enrich the User-Level (UL) experience and informedness of MHPs [Roy et. al., 2023].

Limiting our discovery to context-based high-level analysis, the System-Level (SL) observations for WoeBot and Wysa suggest the UL tracking of human behavior, such as gratitude/ mindfulness and frequent mood changes (an emotional spectrum) during the day. Contributions toward this endeavor have emerged through exclusive studies with *trustworthiness* of WoeBot and Wysa through ethical research protocols, as it is mandatory to concur ethical dimensions due to the sensitive nature of VMHAs. The lack of *ethical dimensions* in WoeBot and Wysa is exemplified through non-clinical grounding and lack of contextual awareness in responses to the emergencies such as disclosure of immediate harm or suicidal ideation [Koutsouleris et al., 2022]. To this end, the development of *safe and explainable* VMHAs shall enhance their capabilities of reading between the lines result-

[6] https://woebothealth.com/why-we-need-mental-health-chatbots/

| VMHA | | Objective | KI | | DiG | Safety | UsEx | QM |
|---|---|---|---|---|---|---|---|---|
| | | | PK | MK | | | | |
| [Ginger, 2011] | Ginger | Behavioral Health Coaching | ✗ | ✓ | ✗ | ✗ | ✗ | H |
| [CompanionMX, 2011] | CompanionMX | PTSD | ✗ | ✗ | ✗ | ✗ | ✗ | H |
| [Quartet, 2014] | Quartet | Therapy & Counseling | ✗ | ✗ | ✗ | ✗ | ✗ | H |
| [Fitzpatrick et al., 2017] | Woebot | CBT | ✓ | ✓ | ✗ | ✗ | ✗ | A |
| [Limbic, 2017] | Limbic | CBT | ✗ | ✗ | ✗ | ✓ | ✗ | H |
| [Inkster et al., 2018] | Wysa | CBT | ✗ | ✗ | ✗ | ✗ | ✗ | A |
| [Fulmer et al., 2018] | Tess | Anxiety & Depression | ✗ | ✗ | ✗ | ✗ | ✗ | - |
| [Ghandeharioun *et al.*, 2019] | EMMA | CBT | ✗ | ✗ | ✗ | ✗ | ✗ | H |
| [Denecke et al., 2020] | SERMO | CBT | ✗ | ✗ | ✗ | ✗ | ✗ | H |
| [Possati, 2022] | Replika | Empathetic & Supportive | ✗ | ✗ | ✗ | ✗ | ✗ | A |
| [Roy et. al., 2023] | ALLEVIATE | Depression | ✓ | ✓ | ✓ | ✓ | ✗ | H |
| Our Survey Paper | Desired System | Screening, Triaging, & MI | ✓ | ✓ | ✓ | ✓ | ✓ | H,A,T |

Table 2: Prominent and in-use VMHAs with different objectives for supporting patients with mental disturbance. We performed a high-level analysis of all the VMHAs based on publicly-available user reviews on forums (e.g., WebMD, AskaPatient, MedicineNet), and Reddit. For Woebot, Wysa, and Alleviate, a survey of 40 participants was carried out at Prisma Health. Here we define QM: Qualitative Metrics as H: Harmlessness, A: Adherence, T: Transparency.

ing in accountable and fair conversational agents. For a well-aware (about user's depression) dialogue agent, it is perhaps *safer* to avoid mentioning or inquiring about the topics that can worsen the users' mental health condition [Henderson et al., 2018].

Although WoeBot employs medical and process knowledge, to **explain** the decision-making, we investigate the relevant datasets for FAIR principles[7] (see Table 1) and evaluation metrics for quantitative and qualitative performance analysis of VMHAs' question-response generation module in active communication [Brocki et al., 2023]. We further investigate existing evaluation metrics from *passive communication* to support the VMHAs for *active communication*.

## 6 Discussion

The field of AI-powered automated VMHAs is still in its nascent phase and continuously evolving to provide accessible health care to an increasing number of patients with mental illnesses. However, repetitive question/answer functionality within the models fails to sustain the user's engagement. Irrespective of deploying state-of-the-art VMHAs to mitigate the problems of the overburdened healthcare systems, the gap still remains between user's clinical needs and VMHAs that is yet to be connected. Despite the significant amount of studies in realizing the requirement of *safety, harmlessness, explainability, curation of process and medical knowledge-based datasets and knowledge-infused learning methods*, they have never been incorporated or evaluated to enhance the contextualized conversations within a VMHA and their role in the emerging areas of mental healthcare. Hence, there is an urgent need to incorporate high-level contextual analysis and infuse new technical abilities of AI for VMHA. We outline two sub-sections to discuss: (i) the need of revamping the *evaluation metrics*, and (ii) *emerging* areas for developing safe and explainable VMHAs.

### 6.1 Evaluation Method

All the notable earlier work [Walker *et al.*, 1997] included subjective measures involving human-in-the-loop to evaluate a conversational system for its utility in the general purpose domain. Due to the expensive nature of human-based evaluation procedures, researchers have started using machine learning-based automatic quantitative metrics (e.g., BLEURT, BERTScore [Clinciu *et al.*, 2021], BLEU [Papineni *et al.*, 2002], ROUGE [Lin, 2004]) to evaluate the semantic similarity of the machine-translated text. [Liu et al., 2017] highlights the disagreement of users with existing metrics thereby lowering their expectations. Also, most of these traditional quantitative metrics are reference-based which is limited in availability and very difficult to ensure the quality of the human-written references [Bao et al., 2022]. To address these issues and holistically evaluate a desired VMHA with respect to *explainability*, *safety*, and *knowledge process inclusion*, we need to revamp the metrics to bring VMHA systems closer to real-time applications.

**Qualitative Metric** We define mutlimetric evaluation strategy by instilling metrics that correlate well with human judgement and can provide more granular analysis towards more realistic VMHAs.

- **Adherence:** Adherence, a long-standing discussion in the healthcare sector [Fadhil et. al., 2018], is defined as a commitment towards the goal (e.g., long-term therapy, physical activity, or medicine). Despite the AI community showing a significant interest in evaluating the adherence of users [Davis et al., 2020] towards health assistants, the lack of *safe* response, in terms of *DiG* and *UsEx* in VMHAs, add the criticism with loss of adherence. This situation necessitates the requirement of adherence as a qualitative metric towards realizing more *realistic* and *contextual* VMHAs while treating patients with serious mental illness.
- **Harmlessness:** The conversational agents tend to generate harmful, unsafe, and sometimes incoherent infor-

[7] https://www.go-fair.org/fair-principles/

mation [Welbl et. al., 2021]. Although researchers have made many efforts to curb the toxicity in the proliferation of hateful speech and biases in social media, much need to be realized when VMHAs are trained using the same datsets.

- **Transparency:** The transparency and interpretability for understandable models (TIFU) framework emphasize the "explainability" of VMHAs by focusing on *UsEx* and *DiG*, thereby processing the knowledge to obtain clinically-verified responses [Joyce et al., 2023].

**KI Metric:** In this section, we provide metrics that describe *DiG*, *safety*, *MK* and *PK* in table 2. ✓ and ✗ tell whether VHMA has been tested for these KI metrics.

- **Safety:** Even though the datasets have been verified to be safe [Sezgin et al., 2022], it is quite difficult to evaluate the models based on acceptable standards of safety because of their black-box nature. To include safety as a metric of evaluating conversational models, [Roy et al, 2022b] introduces a safety lexicon as a glossary of clinical terms that the MHP would understand in their dataset. [Henderson et al., 2018] emphasizes on the underlying bias of a data-driven model and the need of an idea for contextual safety in the dialogue systems.
- **Logical Coherence (LC):** LC is a qualitative check of the logical relationship between a user's input and the follow-up questions measuring *PK* and *MK*. [Kane et al., 2020] used LC to ensure the reliable output from the RoBERTa model trained on the MNLI challenge and natural language inference GLUE benchmark, hence, opening new research directions towards safer models for MedNLI dataset [Romanov et al., 2018].
- **Semantic Relations (SR):** SR measures the extent of similarity between the response generation and the user's query [Kane et al., 2020]. [Gaur et al., 2022a] highlights the use of SR for logical ordering of question-generations and hence, preventing language models from hallucinations. It further enhances the use of VMHAs for profiling the user's illness and generating appropriate responses through *DiG*.

### 6.2 Emerging Areas of VMHAs

**Mental Health Triage** Mental Health Triage[8] is a risk assessment that categorizes the severity of the mental disturbance before suggesting psychiatric help to the users and categorizes them on the basis of urgency.The screening and triage system could fulfill more complex requirements to achieve automated triage empowered by AI. A recent surge in the use of screening mechanisms by Babylon[9] and Limbic[10] has given new research directions towards a *trustworthy* and *safe* models in near future.

**Motivational Interviewing** Motivational Interviewing (MI) is a directive, user-centered counseling style for eliciting behavior change by helping clients to explore and resolve ambivalence. In contrast to the assessment of severity in mental health triaging, MI enables more interpersonal relationships for cure with a possible extension of MI for mental illness domain [Westra *et al.*, 2011]. [Wu et. al., 2020] suggest human-like empathetic response generation in MI with support for *UsEx* and *contextualization* with clinical knowledge. Recent works in identifying the interpersonal risk factors [Ghosh et al., 2022] from offline text documents further support MI for active communications.

**Clinical Diagnostic Interviewing (CDI)** : CDI is a direct client-centered interview between a clinician and patient without any intervention. With multiple modalities of the CDI data (e.g., video, text, audio), the applications are developed in accordance with Diagnostic and Statistical Manual of Mental Disorders (DSM-V) to facilitate a quick gathering of detailed information about the patient. In contrast to the in-person sessions (leveraged on both verbal and non-verbal communication), the conversational agents miss the *personalized* and *contextual* information from non-verbal communication hindering the efficacy of VMHAs.

### 6.3 Practical Considerations

We now consider two practical considerations with VMHAs.
**Difference in human v/s machine assistance:** For the VMHAs to be accepted by people in need, it is important that they feel the output of the system is valuable and useful. If the user had sought the help of a human mental health professional in the past, she would expect a similarly realistic conversational experience from the VMHA. However, getting training data from real conversations is expensive and fraught with data privacy and annotation challenges. To maintain the confidentiality of user data, approaches akin to popular methods used in recommendation literature for creating training data from user data could be used: (a) anonymize real data, (b) abstract from real data to create representative (but inaccurate) samples, and (c) generate synthetic conversations based on characteristics of real data. In recommendations, user data is used to create personas while in the case of VMHA, real conversations can be used to create conversation templates and assign user profiles [Qiao et al., 2018]. But high-quality annotations on (conversation) data are a more significant problem and widespread in learning-based AI tasks.
**Perception of quality with assistance offered:** A well-understood result in marketing is that people perceive the quality of a service based on the price paid for it as well as the word of mouth buzz around it [Liu et al., 2016]. In the case of VMHAs, it is an open question whether help offered by VMHAs will be considered inferior to that offered by professionals. More crucially, if a user perceives it negatively, will this further aggravate the user's mental condition?

## 7 Conclusion

From 297 studies on mental health (active and passive communications), we present a systematic survey of $\sim 80$ intelligent technologies for improving the user experience through VMHA or potential VMHA. We first propose a taxonomy of the mental healthcare domain for social NLP research, thrusting on benchmarking evaluation metrics for active communi-

[8] https://en.wikipedia.org/wiki/Mental_health_triage

[9] https://tinyurl.com/2p8be7d4

[10] https://tinyurl.com/2s44uxnk

cations. We then discussed the efforts in knowledge-driven AI for mental health, its connection with *UsEx* and *safety*, and provided methods of improving and enhancing VMHA to support triaging, motivational interviewing, and diagnostic interviews. Finally, the survey sees its extension to "personalization" in VMHA, which is needed to perform tasks like screening, triaging and MI. Recently, Anthropic's Claude(a competitor of ChatGPT) is another effort to induce better safety with UsEx in the conversational system [Bai et al., 2022].

## Ethical Statement

We adhere to anonymity, data privacy, intended use, and practical implication of the VMHAs. The questionnaires and rating scales described as clinical process knowledge do not contain personally identifiable information. The datasets covered in the survey are publicly available and can be obtained from user-author agreement forms. The text conversation in the figures are abstract and has no relevance with the real-time data source or any person.

## References

[Althoff et al., 2016] Tim Althoff et al. Large-scale analysis of counseling conversations: An application of natural language processing to mental health. *In TACL*, 2016.

[Bai et al., 2022] Yuntao Bai et al. Constitutional ai: Harmlessness from ai feedback. *ArXiv*, 2022.

[Bao et al., 2022] Forrest Sheng Bao et al. Docasref: A pilot empirical study on repurposing reference-based summary quality metrics reference-freely. *ArXiv*, 2022.

[Bhatt *et al.*, 2020] Umang Bhatt, Alice Xiang, Shubham Sharma, Adrian Weller, Ankur Taly, Yunhan Jia, Joydeep Ghosh, Ruchir Puri, José M. F. Moura, and Peter Eckersley. Explainable machine learning in deployment. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, 2020.

[Bommasani et al., 2021] Rishi Bommasani et al. On the opportunities and risks of foundation models. *ArXiv*, 2021.

[Brocki et al., 2023] Lennart Brocki et al. Deep learning mental health dialogue system. *IEEE BDSC*, 2023.

[Cao et al., 2020] Lei Cao et al. Building and using personal knowledge graph to improve suicidal ideation detection on social media. *IEEE ToM*, 2020.

[Chung et al., 2021] Kyungmi Chung et al. A chatbot for perinatal women's and partners' obstetric and mental health care: Development and usability evaluation study. *JMIR*, 2021.

[Cirille et al., 2020] Davide Cirille et al. Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare. *NPJ digital medicine*, 2020.

[Clinciu *et al.*, 2021] Miruna Clinciu, Arash Eshghi, and Helen Hastie. A study of automatic metrics for the evaluation of natural language explanations. *arXiv preprint arXiv:2103.08545*, 2021.

[CompanionMX, 2011] CompanionMX, 2011.

[CounselChat, 2015] CounselChat. Mental health answers from counselors. *CounselChat*, 2015.

[Czeisler et al., 2020] Mark É Czeisler et al. Mental health, substance use, and suicidal ideation during the covid-19 pandemic—united states, june 24–30, 2020. *Morbidity and Mortality Weekly Report*, 2020.

[Davis et al., 2020] Courtney R Davis et al. A process evaluation examining the performance, adherence, and acceptability of a physical activity and diet artificial intelligence virtual health assistant. *IJERPH*, 2020.

[Demasi et al., 2019] Orianna Demasi et al. Towards augmenting crisis counselor training by improving message retrieval. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, 2019.

[Denecke et al., 2020] Kerstin Denecke et al. A mental health chatbot for regulating emotions (sermo)-concept and usability test. *IEEE TETC*, 2020.

[Dinan et al., 2018] Emily Dinan et al. Wizard of wikipedia: Knowledge-powered conversational agents. *ArXiv*, 2018.

[Dinan et al., 2021] Emily Dinan et al. Anticipating safety issues in e2e conversational ai: Framework and tooling. *ArXiv*, 2021.

[Donnelly et al., 2006] Kevin Donnelly et al. Snomed-ct: The advanced terminology and coding system for ehealth. *Studies in HT&I*, 2006.

[Fadhil et. al., 2018] Ahmed Fadhil et. al. A conversational interface to improve medication adherence: towards ai support in patient's treatment. *ArXiv*, 2018.

[First, 2014] Michael B First. Structured clinical interview for the dsm (scid). *Clinical Psychology*, 2014.

[Fitzpatrick et al., 2017] Kathleen Kara Fitzpatrick et al. Woebot: a randomized controlled trial. *JMIR*, 2017.

[Fulmer et al., 2018] Russell Fulmer et al. Tess: randomized controlled trial. *JMIR*, 2018.

[Gaur et al., 2022a] Manas Gaur et al. Iseeq: Information seeking question generation using dynamic meta-information retrieval and knowledge graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022.

[Gaur et al., 2022b] Manas Gaur et al. Knowledge-infused learning: A sweet spot in neuro-symbolic ai. *IEEE IC*, 2022.

[Ghandeharioun *et al.*, 2019] Asma Ghandeharioun, Daniel McDuff, Mary Czerwinski, and Kael Rowan. Emma: An emotion-aware wellbeing chatbot. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, 2019.

[Ghosh et al., 2022] Soumitra Ghosh et al. Am i no good? towards detecting perceived burdensomeness and thwarted belongingness from suicide notes. *IJCAI*, 2022.

[Ginger, 2011] Ginger. In-the-moment care for every emotion., 2011.

[Gunaratna et al., 2022] Kalpa Gunaratna et al. Explainable slot type attentions to improve joint intent detection and slot filling. *ArXiv*, 2022.

[Gupta et. al., 2022a] Khushboo Gupta et. al. Deepmind introduces 'sparrow,' an artificial intelligence-powered chatbot developed to build safer machine learning systems. *MarkTechPost*, Sep 2022.

[Gupta et al., 2022b] Shrey Gupta et al. Learning to automate follow-up question generation using process knowledge for depression triage on reddit posts. In *CLPsych*, 2022.

[Gyrard et al., 2022] Amelie Gyrard et al. Interdisciplinary iot and emotion knowledge graph-based recommendation system to boost mental health. *Applied Sciences*, 2022.

[Heikkilä et. al., 2022] Melissa Heikkilä et. al. Deepmind's new chatbot uses google searches plus humans to give better answers. *MIT Technology Review*, 2022.

[Henderson et al., 2018] Peter Henderson et al. Ethical challenges in data-driven dialogue systems. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 2018.

[Hennemann et al., 2022] Severin Hennemann et al. Diagnostic performance of an app-based symptom checker in mental disorders: comparative study in psychotherapy outpatients. *JMIR*, 2022.

[Hoffman et al., 2018] Robert R. Hoffman et al. Metrics for explainable ai: Challenges and prospects. In *arXiv preprint arXiv:1812.04608*, 2018.

[Huang et. al., 2015] Rongyao Huang et. al. *Language use in teenage crisis intervention and the immediate outcome: A machine automated analysis of large scale text data*. PhD thesis, Master's thesis, Columbia University, 2015.

[Hyman et al., 2008] I Hyman et al. Self-disclosure and its impact on individuals who receive mental health services. *SAMHSA*, 2008.

[Inkster et al., 2018] Becky Inkster et al. Wysa: real-world data evaluation mixed-methods study. *JMIR*, 2018.

[Joyce et al., 2023] Dan W Joyce et al. Explainable artificial intelligence for mental health through transparency and interpretability for understandability. *Nature*, 2023.

[Kane et al., 2020] Hassan Kane et al. Nubia: Neural based interchangeability assessor for text generation. *ArXiv*, 2020.

[Kazi *et al.*, 2012] Hameedullah Kazi, Bhavani Shankar Chowdhry, and Zeesha Memon. Medchatbot: an umls based chatbot for medical students. 2012.

[Koulouri et al., 2022] Theodora Koulouri et al. Chatbots to support young adults' mental health: An exploratory study of acceptability. *ACM TiiS*, 2022.

[Koutsouleris et al., 2022] Nikolaos Koutsouleris et al. From promise to practice: towards the realisation of ai-informed mental health care. *Lancet*, 2022.

[Liang et al., 2021] Kai-Hui Liang et al. Evaluation of in-person counseling strategies to develop physical activity chatbot for women. *ArXiv*, 2021.

[Limbic, 2017] Limbic. Enabling the best psychological therapy, 2017.

[Limsopatham et al., 2016] Nut Limsopatham et al. Normalising medical concepts in social media texts by learning semantic representation. In *Proceedings of ACL*, 2016.

[Lin, 2004] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, 2004.

[Liu et al., 2016] Chih-Hsing Sam Liu et al. Service quality and price perception of service: Influence on word-of-mouth and revisit intention. *JATM*, 2016.

[Liu et al., 2017] Siqi Liu et al. Improved image captioning via policy gradient optimization of spider. In *IEEE ICCV*, 2017.

[Longo *et al.*, 2020] Luca Longo, Randy Goebel, Freddy Lecue, Peter Kieseberg, and Andreas Holzinger. Explainable artificial intelligence: Concepts, applications, research challenges and visions. In Andreas Holzinger, Peter Kieseberg, A Min Tjoa, and Edgar Weippl, editors, *Machine Learning and Knowledge Extraction*. Springer International Publishing, 2020.

[Miner et al., 2016] Adam Miner et al. Conversational agents and mental health: Theory-informed assessment of language and affect. In *Proceedings of the fourth international conference on human agent interaction*, 2016.

[Noble et al., 2022] Jasmine M Noble et al. Mira covid 19 chatbot. *JMIR*, 2022.

[Papineni *et al.*, 2002] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002.

[Possati, 2022] Luca M Possati. Psychoanalyzing artificial intelligence: The case of replika. *AI & SOCIETY*, 2022.

[Qiao et al., 2018] Qian Qiao et al. Assigning personality/profile to a chatting machine for coherent conversation generation. In *IJCAI*, 2018.

[Quan et al., 2005] Hude Quan et al. Coding algorithms for defining comorbidities in icd-9-cm and icd-10 administrative data. *Medical care*, 2005.

[Quartet, 2014] Quartet. Mental health care, made easier, 2014.

[Raskin et al., 2019] Hannah Raskin et al. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proc. ACL*, 2019.

[Regier et. al., 2013] Darrel A Regier et. al. The dsm-5: Classification and criteria changes. *World psychiatry*, 2013.

[Romanov et al., 2018] Alexey Romanov et al. Lessons from natural language inference in the clinical domain. *ArXiv*, 2018.

[Roy et. al., 2022a] Kaushik Roy et. al. Process knowledge-infused learning for suicidality assessment on social media. *arXiv preprint arXiv:2204.12560*, 2022.

[Roy et al, 2022b] Kaushik Roy et al. Proknow: Process knowledge for safety constrained and explainable question generation for mental health diagnostic assistance. *In Frontiers BD*, 2022.

[Roy et. al., 2023] Kaushik Roy et. al. Alleviate chatbot. *UMBC Faculty Collection*, 2023.

[SAM, 2020] 2020 national survey of drug use and health (nsduh) releases. *SAMHSA*, 2020.

[Sezgin et al., 2022] Emre Sezgin et al. Gpt3 healthcare language model. *JMIR*, 2022.

[Sharma et al., 2021] Ashish Sharma et al. Towards facilitating empathic conversations in online mental health support: A reinforcement learning approach. In *Proceedings of the Web Conference 2021*, 2021.

[Sharma et. al., 2023] Ashish Sharma et. al. Human–ai collaboration enables more empathic conversations in text-based peer-to-peer mental health support. *Nature Machine Intelligence*, 2023.

[Sheth et al., 2019] Amit Sheth et al. Extending patient-chatbot experience with internet-of-things and background knowledge: case studies with healthcare applications. *IEEE IS*, 2019.

[Sheth et. al., 2021] Thirunarayan Krishnaprasad Sheth et. al., Amit. The duality of data and knowledge across the three waves of ai. *IT professional*, 2021.

[Srivastava et al., 2021] Biplav Srivastava et al. Did chatbots miss their "apollo moment"? potential, gaps, and lessons from using collaboration assistants during covid-19. In *Patterns*, 2021.

[Uban et. al., 2021] Ana-Sabina Uban et. al. An emotion and cognitive based analysis of mental health disorders from social media data. *FGCS*, 2021.

[Varshney et al., 2022] Kush R Varshney et al. Trustworthy machine learning. *ISBNL 979-8411903959*, 2022.

[Walker *et al.*, 1997] Marilyn A Walker, Diane J Litman, Candace A Kamm, and Alicia Abella. Paradise: A framework for evaluating spoken dialogue agents. *arXiv preprint cmp-lg/9704004*, 1997.

[Weick et. al., 1995] Karl E Weick et. al. *Sensemaking in organizations*. Sage, 1995.

[Welbl et. al., 2021] Johannes Welbl et. al. Challenges in detoxifying language models. In *Proceedings of EMNLP*, 2021.

[Welivita et al., 2022a] Anuradha Welivita et al. Motivational interviewing dataset. In *COLING*, 2022.

[Welivita et. al., 2022b] Pu Pearl Welivita et. al., Anuradha. Heal: A knowledge graph for distress management conversations. In *Proceedings of the AAAI*, 2022.

[Westra *et al.*, 2011] Henny A Westra, Adi Aviram, and Faye K Doell. Extending motivational interviewing to the treatment of major mental health problems: current directions and evidence. *The Canadian Journal of Psychiatry*, 2011.

[Wu et. al., 2020] Zixiu Wu et. al. Towards detecting need for empathetic response in motivational interviewing. In *Companion Publication of the 2020 ICMI*, pages 497–502, 2020.

[Xu et al., 2020] Jing Xu et al. Recipes for safety in open-domain chatbots. *arXiv preprint arXiv:2010.07079*, 2020.

[Yazdavar et al., 2017] Amir Hossein Yazdavar et al. Semi-supervised approach to monitoring clinical depressive symptoms in social media. In *Proceedings of the 2017 IEEE/ACM international conference on advances in social networks analysis and mining 2017*, 2017.

[Zhang et. al., 2022] Tianlin Zhang et. al. Natural language processing applied to mental illness detection: a narrative review. *NPJ digital medicine*, 2022.

[Zirikly et. al., 2022] Dredze Zirikly et. al., Ayah. Explaining models of mental health via clinically grounded auxiliary tasks. In *CLPsych*, 2022.