

---

# Position: Beyond Assistance – Reimagining LLMs as Ethical and Adaptive Co-Creators in Mental Health Care

---

Abeer Badawi<sup>1 2</sup> Md Tahmid Rahman Laskar<sup>3 4</sup> Jimmy Xiangji Huang<sup>3</sup> Shaina Raza<sup>2</sup> Elham Dolatabadi<sup>1 2</sup>

## Abstract

This position paper argues for a fundamental shift in how Large Language Models (LLMs) are integrated into the mental health care domain. We advocate for their role as co-creators rather than mere assistive tools. While LLMs have the potential to enhance accessibility, personalization, and crisis intervention, their adoption remains limited due to concerns about bias, evaluation, over-reliance, dehumanization, and regulatory uncertainties. To address these challenges, we propose two structured pathways: *SAFE-i* (Supportive, Adaptive, Fair, and Ethical Implementation) Guidelines for ethical and responsible deployment, and *HAAS-e* (Human-AI Alignment and Safety Evaluation) Framework for multidimensional, human-centered assessment. *SAFE-i* provides a blueprint for data governance, adaptive model engineering, and real-world integration, ensuring LLMs align with clinical and ethical standards. *HAAS-e* introduces evaluation metrics that go beyond technical accuracy to measure trustworthiness, empathy, cultural sensitivity, and actionability. We call for the adoption of these structured approaches to establish a responsible and scalable model for LLM-driven mental health support, ensuring that AI complements, rather than replaces human expertise.

## 1. Introduction

The rapid integration of Large Language Models (LLMs) into mental health presents an unprecedented opportunity to enhance the accessibility, personalization, and scalability of mental health support (Bedi et al., 2024). Yet, the global

shortage of mental health professionals poses a significant barrier to care. According to the World Health Organization’s mental health atlas (Organization, 2021), the global median number of mental health workers is 13 per 100,000 people - equivalent to a stadium filled with 8,000 individuals, yet only one professional available to provide support. This disparity highlights the urgent need for innovative solutions to bridge the gap in mental health care delivery.

Despite the rapid advancements of AI in healthcare and the urgent demand for mental health solutions (D’Alfonso, 2020), recent reports (Insights & Healthcare, 2024) highlight that mental health analytics remains one of the least deployed AI products and services. A survey of over 900 healthcare professionals found that while AI adoption is prevalent in electronic health records automation (63%), medical imaging (64%), and patient analytics (62%), its integration into mental health analytics is significantly lower (48%). Additionally, only 21% of healthcare institutions have adopted AI for mental health, with another 27% considering adoption, making it one of the least prioritized areas of AI implementation (Insights & Healthcare, 2024).

The under-utilization of AI in mental health is not merely a technological issue but a reflection of deeper concerns surrounding trust, ethical considerations, and the preservation of human expertise (Hamdoun et al., 2023). As LLMs become increasingly sophisticated, the mental health community faces a critical challenge: how to leverage their transformative potential while upholding the human-centered principles that define effective care (Obradovich et al., 2024). This tension is further exacerbated by the ability of LLMs to mimic human interaction and generate seemingly personalized responses, which may lead individuals to overestimate the depth of understanding these models possess (Sharma et al., 2020). Such dynamics can result in undue trust in LLM outputs, potentially neglecting other forms of support or treatment (Hua et al., 2024b).

Furthermore, recent studies show increasing public trust and optimism. Varghese et al. (2024) found that 53% of respondents moderately trust AI tools for mental health, valuing their accessibility, anonymity, and stigma reduction, while 34.8% expressed optimism due to their constant availability and cost-effectiveness. Alanezi (2024) reported positive

---

<sup>1</sup>Faculty of Health, York University, Canada <sup>2</sup>Vector Institute, Canada <sup>3</sup>Information Retrieval and Knowledge Management Research Lab, York University, Canada <sup>4</sup>Dialpad Inc., Canada. Correspondence to: Abeer Badawi <abeerbadaui@yorku.ca>, Elham Dolatabadi <edolatab@yorku.ca>.

perceptions of ChatGPT for psychoeducation and emotional support, and Siddals et al. (2024) observed high user engagement and therapeutic benefits from generative AI chatbots. Together, these findings highlight AI's growing acceptance, particularly when systems are empathetic, support appropriate crisis escalation, and clearly define AI-human boundaries. This supports the concept of user acceptability of the use of LLMs in mental health, which is crucial for the success of AI-driven mental health interventions.

Through our collaboration with an e-mental health organization, where we evaluated LLMs on anonymized crisis support conversations, we found that the lack of robust development, evaluation, and deployment frameworks with human-in-the-loop hinders their safe and effective integration in mental health care (Obadinma et al., 2025). These concerns are shared by both individuals seeking mental health support and the professionals providing it, creating resistance and uncertainty around AI integration (Sobaih et al., 2025). Without a clear framework to ensure complementarity between AI and human-led interventions, these technologies risk being underutilized or misapplied, undermining their potential to augment mental health. Despite these challenges, early applications of human-AI collaboration demonstrate promising results. For instance, HAILEY (Sharma et al., 2023), a system designed to enhance empathy in peer-to-peer mental health support, has shown that conversations co-authored by LLMs are consistently rated as more empathic and supportive than human-only interactions.

However, the deployment of LLMs in mental health care remains fraught with technical and ethical challenges. Studies reveal that these models often exhibit demographic biases, producing less empathetic or even harmful responses when interacting with underrepresented groups (Zack et al., 2024; Raza et al., 2024). Furthermore, proprietary models, such as ChatGPT 3.5, have demonstrated unsafe triage rates, misclassify urgent mental health crises, and potentially delay critical care, raising concerns about their reliability in high-stakes scenarios (Fraser et al., 2023). The absence of robust frameworks for development, evaluation, and deployment makes it difficult to ensure the effectiveness and safety of these tools. Accordingly, this paper proposes a path forward, redefining the role of LLMs in this sensitive domain through collaborative, ethical, and adaptive AI-human partnerships.

**Our position** This paper argues that LLMs have reached a pivotal stage where their implementation and evaluation of mental care is both viable and necessary. We advocate for reimagining LLMs as **active co-creators rather than passive assistants, emphasizing supportive, ethical, and adaptive AI-human partnerships that enhance - rather than replace - human-led mental health support.**

In our view, LLMs should evolve as dynamic and adaptive tools to enhance healthcare providers' experience through

iterative learning, personalization, and interpretability. This paradigm shift recognizes the deeply personal, emotional, and high-risk nature of mental health care, ensuring that LLMs complement human expertise while addressing the unique challenges of this domain. To achieve this, we argue the need for ethical data practices, open-source models, and human-AI collaboration to ensure safety and accountability. We propose reframing the role of LLMs as *augmentative* rather than *autonomous*, with implementation and evaluation frameworks that move beyond narrow technical metrics to encompass trustworthiness, empathy, cultural sensitivity, and the ability to drive meaningful, actionable outcomes.

This position paper makes the following key contributions:

- **Comprehensive Analysis of Prior Work and Alternative Viewpoints** We offer a critical examination of existing LLM applications in mental health by identifying their strengths, limitations, and alternative perspectives.
- **Identification of Key Challenges and Gaps** that hinder the responsible deployment of LLMs in mental health, including: (1) the necessity of ethical and diverse data foundations, (2) the need for robust model engineering with adaptive optimization, and (3) the absence of human-centered evaluation frameworks.
- **Proposing the SAFE-i (Supportive, Adaptive, Fair, and Ethical Implementation) Guidelines** to ensure LLMs function as supportive, adaptive, fair, and ethical implementation co-creators in mental healthcare. The structured approach is built on three core pillars: Ethical Data Foundations, Model Engineering, and Real-World Integration as shown in Figure 1.
- **Introducing the HAAS-e (Human-AI Alignment and Safety Evaluation) Framework** to rigorously assess LLMs in mental health using a multidimensional approach. It defines four core evaluation criteria, including trustworthiness, fairness, empathy, and helpfulness, operationalized through four novel quantitative metrics that measure alignment with human expertise, cultural sensitivity, personalization, and actionability. Additionally, it integrates four validation methods—randomized trials, multi-method evaluations, red teaming, and testing—to ensure safety, accountability, and real-world applicability as shown in Figure 1.

## 2. Alternative Views

**AI Cannot Replicate Human Emotional Intelligence.** Some researchers argue that LLMs, despite advances in empathetic response generation, lack the depth of understanding, lived experience, and contextual sensitivity required for mental health support. Unlike trained professionals, AI

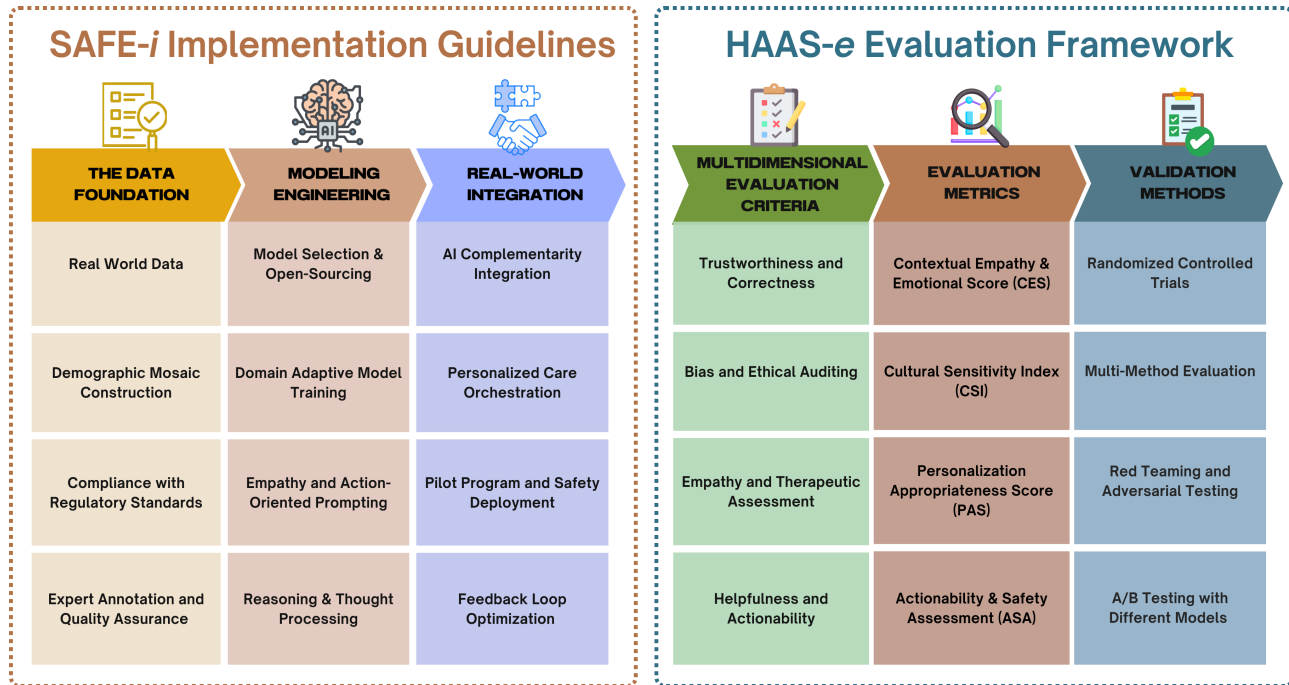


Figure 1. The proposed SAFE-i Implementation Guidelines and HAAS-e Evaluation Framework

models may misinterpret complex emotional cues, potentially leading to harmful advice (Montemayor et al., 2022). **Response:** LLMs can be designed to operate within well-defined boundaries, providing initial support, triage, or supplemental resources while flagging complex cases for human intervention. If we leverage domain-specific models and expert-guided annotations, LLMs can be tuned to recognize nuanced emotional cues (Yang et al., 2024) and defer high-risk or ambiguous situations to human responders (Sharma et al., 2023). Moreover, continuous evaluation of an LLM’s ability to align with human emotional understanding ensures that AI tools remain supportive and safe, complementing rather than competing with human emotional intelligence (Stade et al., 2024).

**The Risk of Over-Reliance and Dehumanization.** LLMs also create a false sense of human-like understanding, leading users to overestimate their reliability. There is concern that increased reliance on AI-driven mental health solutions may reduce the role of human therapists and crisis responders, leading to depersonalization of care (Choudhury et al., 2023). For instance, vulnerable individuals might develop trust in AI-based therapeutic tools, potentially substituting them for human therapists, increasing the risk of social isolation. If organizations prioritize AI over human-led interventions due to cost or scalability, the quality of support may decline, especially for individuals who need deeper, long-term engagement.

**Response:** To mitigate over-reliance, it is essential to im-

plement LLMs as complementary tools rather than replacements for human therapists (Sharma et al., 2023). Educating users on limitations, personalizing care strategies, and integrating feedback mechanisms ensure adaptation to individual needs and encourage users to seek human support when necessary (Strong et al., 2024). LLM systems can provide initial support when we integrate safety nets and escalation protocols while ensuring high-risk cases are promptly addressed by qualified professionals.

**Regulatory and Safety Uncertainties.** Some experts advocate against LLM integration in mental health until robust regulatory frameworks are in place. The lack of standardized safety measures raises ethical concerns, particularly regarding potential harm if AI-generated responses are inaccurate or inappropriate (Tavory, 2024).

**Response:** A comprehensive regulatory framework is crucial for the safe deployment and reliable evaluation of LLMs in mental health (Stade et al., 2024). This includes establishing standardized safety protocols for data, including real-time monitoring and adversarial testing, which can help identify and address potential risks (De Micco et al., 2025). Furthermore, accountability mechanisms, such as continuous performance evaluation and stakeholder feedback loops, ensure that LLMs adhere to ethical guidelines and remain aligned with the needs of users and professionals (Ferrara, 2023; Hogg et al., 2023; Kaye et al., 2024).

### 3. Prior Efforts in LLM-Powered Applications for Mental Health: A Landscape

The growing demand for mental health services, exacerbated by the COVID-19 pandemic (Hamdoun et al., 2023), has led to the exploration of generative AI technologies in various mental health applications (Zhang & Boulos, 2023; C. & John, 2023). One of the core technologies used in the Generative AI domain is LLMs, such as ChatGPT and GPT-4 (OpenAI, 2023), which utilize billions of parameters to generate coherent, contextually appropriate responses in mental health dialogues (Guo et al., 2024b; Torous & Blease, 2024). LLMs have been effectively applied in various application areas of mental health, such as crisis intervention (Obadinma et al., 2025; Sharma et al., 2024), therapy recommendations (Wilhelm et al., 2023; Malgaroli & McDuff, 2024; Berrezueta-Guzman et al., 2024), etc.

Other applications of LLMs in mental healthcare include the work of Perlis et al. (2024), where they showed GPT-4 aligns with expert bipolar depression management, while Lee et al. (2024) found GPT-4 had comparable sensitivity to clinicians in predicting suicidal ideation from intake data. Moreover, domain-specific LLMs have also gained a lot of attention recently in the mental healthcare domain (Yang et al., 2024). For instance, the Serena model, (Brocki et al., 2023) is developed as an effective counselor and demonstrates enhanced relevance and sensitivity toward therapeutic approaches (Brocki et al., 2023) with just 2.7 billion-parameters. More recently, Guo et al. (2024a) introduced SouLLMate, an adaptive LLM system integrating Retrieval-Augmented Generation (Gao et al., 2023), suicide risk detection, and proactive dialogues to enhance accessibility in mental health support.

In mental health applications, conversational AI tasks represent a major application area, with chatbots designed to engage users in text-based therapeutic conversations or monitor mental well-being (Liu et al., 2024a). For example, the chatbot Woebot, which uses cognitive-behavioral techniques, has shown efficacy in alleviating symptoms of depression and anxiety by delivering timely interventions (Fitzpatrick et al., 2017). The SuDoSys chatbot (Chen et al., 2024), which is based on WHO’s PM+ framework, ensures structured multi-turn psychological counseling with coherent stage tracking. The Coral framework proposed by Sakhrani et al. (2021) is designed to integrate conversational agents in mental health applications. For the evaluation of LLMs in clinical conversations, Johri et al. (2025) present CRAFT-MD, an evaluation framework assessing diagnostic reasoning in clinical LLMs, highlighting limitations of LLMs in conversational accuracy and the need for multi-modal integration before deployment.

### 4. Key Challenges in Utilizing LLMs for Mental Health

This section outlines three key challenges from previous work and alternative views in this field.

#### Challenge 1: The Need for Ethical Data Foundations

The lack of real-world, diverse, and privacy-compliant data limits model reliability and applicability. Bedi et al. (2024) recently conducted a systematic review to examine how LLMs are evaluated in the healthcare domain. They find that existing studies predominantly rely on simulated or social media-based data like Twitter and Reddit, with only 5% of studies utilizing real patient care data for evaluation. Nonetheless, data from these sources may fail to capture the nuances and complexities of real-world mental health interactions (e.g., counseling services or hospitals) (Eichstaedt et al., 2018; Tadesse et al., 2019; Coppersmith et al., 2018). This suggests a significant gap between the theoretical capabilities of LLMs and their actual implementation in mental health settings. As an example, Fraser et al. (2023) compared the diagnostic and triage accuracy of ChatGPT with human physicians in an emergency department. However, this study didn’t involve actual patient interactions.

Moreover, LLMs trained on large datasets of publicly available text may inadvertently absorb and amplify existing societal biases surrounding mental health. If this biased information is then presented to users seeking mental health support, it could reinforce negative perceptions of mental illness, discourage help-seeking behaviors, and exacerbate existing inequalities in access to care (Lawrence et al., 2024). Without robust data collection strategies, LLMs risk biases, misinformation, and ethical concerns. Recent research highlights the importance of data diversity and representation in training and evaluating LLMs for mental health. Counseling and hospital data often underrepresent diverse populations, especially marginalized communities (Hua et al., 2024b; Omiye et al., 2023). Consequently, LLMs trained on data from limited demographics may underperform for other groups, risking misdiagnosis and ineffective treatments (Hua et al., 2024b; Omiye et al., 2023). While GPT-4 showed promise in providing empathetic responses in mental health support contexts, it also exhibited concerning demographic biases (Gabriel et al., 2024).

#### Challenge 2: The Need for Robust Model Engineering and Adaptive Model Optimization

LLMs in mental health applications face significant risks related to correctness, safety, and therapeutic reliability. Issues such as hallucinations, misinformation, and inappropriate responses (Zhao et al., 2023) necessitate more structured engineering processes (e.g., construction of optimized prompts) and real-world testing to ensure reliability and alignment with mental health practices. Researchers emphasize the need for careful



implementation, collaboration with stakeholders, and integration into existing healthcare systems (Torous & Blease, 2024). As the field evolves, there is a focus on developing empathetic, context-aware conversational agents (Sakhrani et al., 2021) and exploring diverse applications of AI in healthcare (Gozalo-Brizuela & Garrido-Merchan, 2023).

Moreover, ChatGPT-like closed-source proprietary LLMs are only accessible via their APIs (Laskar et al., 2023; Jahan et al., 2024), which restricts users from fine-tuning the models locally or accessing their internal layers and weights (Pfohl et al., 2024). Also, relying too much on APIs raises privacy and security concerns, as sensitive data must be shared with third-party providers, increasing risks of data exposure. The lack of transparency in these models further complicates efforts to thoroughly evaluate their reliability and safety, a critical issue when handling sensitive mental health information (Lawrence et al., 2024).

Prior works underscore the absence of a widely accepted framework for healthcare tasks and their evaluation dimensions in mental health (Goldberg et al., 2024). This inconsistency severely hinders the ability to compare results across studies or effectively gauge progress in LLM development for healthcare applications (Stafie et al., 2023; Kohane, 2024), ultimately stalling advancements in this critical field. A recent comprehensive review of 519 studies on healthcare applications of LLMs by Bedi et al. (2024) also highlights the need for standardized implementation methods.

There is also a growing imbalance in AI accessibility across different demographics and healthcare systems. For instance, the cost of fine-tuning models for specific populations remains prohibitively high, leading to disparities in how well these tools serve different groups (Obradovich et al., 2024). While a recent study by Stade et al. (2024) proposed a framework for the responsible development of LLMs in behavioral healthcare that could potentially augment or even replace certain aspects of human-led psychotherapy, the authors also acknowledge significant ethical and practical challenges with implementing this framework.

Additionally, the over-alignment of models to safety constraints can result in over-cautious responses, where LLMs refuse to engage with critical mental health queries, limiting their usefulness in real therapeutic settings (Obradovich et al., 2024). Lastly, a fundamental risk with deploying LLM in mental health settings is hallucination, where models generate output with incorrect or misleading information (Huang et al., 2023). In mental health, this issue is particularly sensitive, as inaccurate guidance or misinformation can have immediate and severe consequences.

**Challenge 3: The Need for Multidimensional and Human-Centered Evaluation** Proper evaluation frameworks are critical to ensure that LLMs in mental health

deliver accurate, safe, and ethical outcomes. This is essential to maximize their potential benefits while minimizing risks to patient safety and therapeutic trust (Bedi et al., 2024). Nonetheless, traditional AI evaluation methods focus primarily on accuracy, neglecting critical aspects such as empathy, cultural sensitivity, and bias detection. For instance, Fraser et al. (2023) only compared the diagnostic accuracy of ChatGPT with human physicians using data analysis. Similarly, Pagano et al. (2023) investigated only the use of GPT-4 for diagnosing arthrosis and providing treatment recommendations.

However, without human-centered evaluation frameworks, LLMs may fail to capture the nuances of real-world mental health support, where human-centered factors like trust, emotional validation, and cultural sensitivity are essential for success. For instance, Pfohl et al. (2024) revealed that traditional evaluation approaches often miss subtle but important biases that could impact healthcare equity. Similarly, Zack et al. (2024) conducted a detailed analysis of GPT-4’s potential to perpetuate racial and gender biases in healthcare settings, finding concerning patterns in the model’s differential diagnoses and treatment recommendations across demographic groups. Recently, Babonnaud et al. (2024) proposed a qualitative protocol for uncovering implicit biases in LLMs, focusing on stereotypes related to gender, sexual orientation, nationality, ethnicity, and religion. Their methodology revealed both explicit and subtle biases in model outputs, particularly in descriptions of minority groups. Furthermore, Adam et al. (2022) demonstrated that the way AI recommendations are framed significantly impacts decision-making bias, with prescriptive recommendations more likely to induce biased outcomes compared to descriptive flags.

Another concern in applying LLMs to mental health support is the potential for over-reliance on AI-driven interventions, which could inadvertently result in dehumanization or a reduction in meaningful human interactions (Zhang & Wang, 2024; Lawrence et al., 2024). Major impacts may include users trusting AI more than human counselors, emotional support provided by LLMs could be perceived as sufficient, and systemic overuse of AI in mental health could result in reduced funding or prioritization for human-led services.

## 5. SAFE-i: Supportive, Adaptive, Fair, and Ethical Implementation Guidelines

Building on our position and an extensive review of existing literature and alternative views, we propose a structured approach to implementing LLMs, which we term *SAFE-i* (Supportive, Adaptive, Fair, and Ethical Implementation). This approach, detailed below and illustrated in Figure 1, ensures that LLMs function as supportive, collaborative, ethical, and adaptive co-creators in mental health care, en-

hancing rather than replacing human-led support.

### 5.1. The Data Foundation: Preparing Reliable and Diverse Mental Health Data

**Real-World Data Harvesting** LLMs for mental health must be trained on real-world data from naturalistic sources like text messages, counselor notes, and conversations. However, only 5% of reviewed studies utilize real patient care data (Bedi et al., 2024), limiting model robustness and generalizability. Synthetic datasets often fail to capture the complexity, variability, and contextual nuances of real-world interactions (Pratap et al., 2022; Bond et al., 2023; Koch et al., 2024). Future implementations must prioritize ethically sourced real-world data while ensuring transparency, informed consent, and opt-out mechanisms for participants (Bhatt, 2024).

**Demographic Mosaic Construction** A core component is population variability, where the source data should reflect different demographics, cultural backgrounds, languages, and mental health conditions (Obermeyer et al., 2019). Regular audits must be conducted to identify the over-representation or under-representation of specific groups (Mienye et al., 2024). Adoptive sampling strategies (Lum & Johndrow, 2016; Chawla et al., 2002) or synthetic data augmentation (Shahul Hameed et al., 2024; Juwara et al., 2024) should be employed where necessary to correct disparities, ensuring the mitigation of the risk of biases and fairness across a wide audience (Abramoff et al., 2023; Zhao et al., 2024).

**Compliance with Regulatory Standards** Sensitive mental health data must be collected, stored, and processed in strict compliance with regulatory standards, including HIPAA (U.S. Department of Health and Human Services, 1996) and GDPR (European Parliament and Council of the European Union, 2016). In addition, implementing robust technical safeguards is critical (Paul & Schaefer, 2020). This includes encrypting data at rest and in transit, utilizing secure storage solutions, and conducting periodic security audits to identify vulnerabilities (Shojaei et al., 2024). Staff training on privacy and security protocols will also ensure both regulatory adherence and data protection (Arain et al., 2019).

**Expert-guided Annotation and Quality Assurance** In unsupervised and self-supervised learning scenarios, the emphasis shifts to the quality and comprehensiveness of the dataset (Yu et al., 2024). LLM models must be trained and evaluated on well-annotated datasets where domain experts label data with relevant markers such as emotional tone, urgency, and risk levels (Lao et al., 2022). In high-risk cases—such as expressions of self-harm or psychosis, annotations should include severity scores, urgency indica-

tors, and clinical insights to improve targeted interventions. Annotation protocols must be continuously refined.

### 5.2. Model Engineering: Designing Adaptive and Effective LLMs

**Model Selection with Open-Source Prioritization** Mental health LLMs should prioritize open-sourcing to foster transparency, community-driven scrutiny, and long-term reliability (Hua et al., 2024a; Yang et al., 2024). Unlike closed-source LLMs (e.g., GPT-4), open-source LLMs enable consistent evaluation and ensure reproducibility (Laskar et al., 2024). The ability to refine the model architecture ensures that AI-driven mental health solutions remain stable, accountable, and adaptable to evolving healthcare needs.

**Domain Adaptive Model Tuning** LLMs designed for mental health must be continuously specialized and refined to maintain therapeutic relevance, ethical integrity, and cultural competence (Guo et al., 2024b; Thakkar et al., 2024). Adopting (e.g., fine-tuning or instruction-tuning) high-quality and domain-specific datasets is essential to embed empathy, rapport-building, and risk assessment into model behavior (Yang et al., 2024). Expert-in-the-loop mechanisms must be integrated to ensure sustained alignment with real-world therapeutic practices, allowing for iterative refinement based on feedback and emerging patient needs (Guo et al., 2024a). Furthermore, models must dynamically adapt to linguistic evolution, cultural shifts, age-specific informal expressions, and emerging mental health concerns, ensuring that LLM remains an inclusive, context-aware, and reliable support tool (Stade et al., 2024; Thakkar et al., 2024).

**Empathy and Action-Oriented Prompt Design** Effective mental health AI requires carefully designed prompts for model adaptations and tuning that shape interactions in a supportive and actionable manner (Li et al., 2024b; Yu & McGuinness, 2024; Priyadarshana et al., 2024). Empathy-driven prompts position the LLM as a compassionate ally, encouraging users to share their feelings safely. Scenario-specific templates address diverse mental health contexts, from anxiety management to crisis support. Prompts also include calls to action, encouraging users to take steps (e.g., contacting a helpline), making the system both informative and actionable (Meskó, 2023; Patil et al., 2024).

**Neural Augmentation via Structured Reasoning and Thought-Based Processing** Tree of Thoughts (ToT) (Yao et al., 2024) and Chain of Thought (CoT) (Wei et al., 2022) reasoning enhance AI ability to break down complex mental health queries into structured, transparent decision paths, improving logical coherence and reducing hallucinations in emotionally sensitive contexts (Yao et al., 2024). By guiding the model to think through psychosocial or affec-

tive problems systematically rather than relying on direct pattern matching, these techniques help in critical therapeutic or crisis scenarios and enhance interpretability. Moreover, research on self-reflective AI suggests that LLMs can improve their accuracy by critically evaluating their own outputs before finalizing responses (Ji et al., 2023; Shinn et al., 2024). Furthermore, integrating uncertainty-aware architectures further enhances safety by enabling models to quantify their confidence levels in sensitive conversations (Yin et al., 2024). When faced with high-risk inputs, these architectures allow AI systems to flag uncertain responses for human review, reducing the likelihood of misleading or inadequate crisis interventions.

### 5.3. Real-World Integration: Human-Centered Continuous Monitoring of LLMs

**Human AI Complementarity Integration** This involves designing systems that specialize tasks based on strengths (AI for data processing and pattern recognition, and humans for empathy and complex decision-making—while ensuring high-risk cases are escalated to human experts (Sharma et al., 2023; Higgins et al., 2023). Additionally, AI should reduce cognitive burden through intuitive interfaces and automated workflows (Fragiadakis et al., 2024).

**Personalized Care Orchestration** The system should be adapted to individual psychological and emotional needs, providing tailored recommendations, therapeutic insights, or support aligned with the user’s mental health context (Kim et al., 2024). The system should also prioritize user trust by being explainable (Kerz et al., 2023; Joyce et al., 2023). Transparency is critical in digital mental health interventions (Stade et al., 2024); users must be clearly informed about which components of their care or support are AI-generated and how the LLMs were developed, fine-tuned, and evaluated, particularly in relation to clinical safety and emotional appropriateness. It is also important to clarify whether the LLMs used are general-purpose models or explicitly optimized for mental health applications, as the latter ensures better alignment with therapeutic goals and reduces risks in vulnerable populations.

**Pilot Program and Safety Net Deployment** Before deploying the system, pilot programs must be conducted to assess safety, ethical considerations, and real-world usability (Sallam et al., 2023; Callahan et al., 2024; Esmaeilzadeh, 2024). Safeguards such as toxicity detection tools (e.g., Llama Guard (Inan et al., 2023)) and automated high-risk content monitoring should be integrated. AI models must be equipped with automated triggers to detect harmful, coercive, or crisis-related content (e.g., suicidal ideation) and escalate cases to human professionals when necessary (Sharma et al., 2023; Higgins et al., 2023; Strong et al., 2024). Without these safety nets, AI-driven mental health

support risks unintended harm.

**Feedback Loop Optimization** Systems must incorporate structured feedback loops that allow users, mental health professionals, and stakeholders to report errors, suggest improvements, and refine system performance over time (De Micco et al., 2025). These mechanisms should include: real-time issue reporting to capture model failures and biases (Ferrara, 2023; Cabrera et al., 2021), stakeholder-driven evaluations to assess the performance from multiple perspectives (Hogg et al., 2023; Kaye et al., 2024), and the “Learning from Incidents” framework (Lukic et al., 2012) that continuously monitors operational failures and systematically addresses them to improve reliability and accountability.

With the key implementation guidelines established, we now explore core evaluation criteria, metrics, and assessment methods for LLMs in mental healthcare.

## 6. HAAS-e: Human- AI Alignment and Safety Evaluation Framework

Traditional AI evaluation metrics, focused on accuracy and efficiency, fail to capture the ethical, emotional, and safety complexities of mental health applications. We advocate for a human-centered approach, we term it Human-AI Alignment and Safety Evaluation (HAAS-e), that defines the key dimensions for LLMs evaluations in mental health as shown in Figure 1.

### 6.1. HAAS-e Multidimensional Evaluation Criteria

To complement our position we define four core dimensions that delineate the key aspects of LLM performance essential for assessing its alignment with human needs and ethical considerations.

**Trustworthiness and Correctness** The model’s reliability should be assessed through correctness and factual accuracy. In mental health contexts, intent classification can be measured using precision, recall, and F1-score, while AlignScore (Zha et al., 2023) evaluates response accuracy. To prevent misinformation, hallucination detection techniques, such as chain-of-thought prompting (Wei et al., 2022), fact-checking with knowledge graphs, and retrieval-augmented generation (Gao et al., 2023) should be employed. Sentiment analysis can further help filter toxic responses (Huang et al., 2023).

**Bias and Ethical Auditing** This step includes the evaluation of biases and ethical concerns in the model’s outputs to ensure fair and equitable LLM responses. These considerations are integral to ensuring fairness and equity. For this purpose, different splits in the test set can be constructed depending on the demographic information to eval-



uate whether the model has any biases in data constructed from certain demographics (Pfohl et al., 2024). Moreover, specific prompts can be constructed to evaluate the potential biases and ethical concerns in certain scenarios. For instance, demographic-aware prompting may include demographic information about the patient, when appropriate and available, to evaluate the biases in model-generated responses in certain demographics (Babonnaud et al., 2024).

**Empathy and Therapeutic Alliance Assessment** Beyond technical accuracy, the models must demonstrate empathy and provide constructive support. This is an important metric to ensure a human-centered evaluation of the models. While these can be achieved automatically via leveraging various neural models (Wankhade et al., 2022) or by using LLMs-as-the-judge (Li et al., 2024a; Gu et al., 2024), evaluating the model responses by human experts, at least on some sampled responses is required to ensure a high-quality evaluation. Moreover, using a standardized framework like the EPITOME (Sharma et al., 2020) that measures empathy based on emotional reactions, interpretations, and perspective-taking could also be used.

**Helpfulness and Actionability Analysis** Another criteria for human-centered evaluation is to measure the helpfulness of the model-generated responses (Tuan et al., 2024). This can be achieved by giving a helpfulness rating to the model response (e.g., via leveraging LLM judges (Li et al., 2024a; Gu et al., 2024) or human experts). In addition, response generation latency (i.e., model’s inference speed), computational requirements, escalation rates for high-risk cases, etc. should also be measured to ensure that the system can be useful for real users.

## 6.2. The HAAS-*e* Evaluation Metrics

Building on the four core evaluation dimensions, the HAAS-*e* metrics operationalize these principles, offering quantitative and qualitative tools to rigorously assess LLM performance in mental health contexts:

**Contextual Empathy & Emotional Score (CES)** Measures an LLM’s ability to understand and respond empathetically to user emotions within mental health contexts. Unlike basic sentiment analysis, CES evaluates the alignment between the LLM responses and the user’s emotional state, situational context, and therapeutic goals. Mathematically, CES can be formulated as a linear combination of two key components: Semantic Coherence which is the alignment,  $Align(R_{llm}, C_{user})$ , between the LLM’s response,  $R_{llm}$ , and the user’s expressed concerns,  $C_{user}$ , and Emotional Alignment, which is the alignment,  $Align(R_{llm}, C_{user}, E_{human})$ , with both the user’s emotions and expert human counselor evaluations  $E_{human}$ . This metric can be quantified by comparing LLM outputs to expert human counselor responses

or through user feedback in double-blind studies. Research supports the feasibility of quantifying empathy (Sharma et al., 2020), and recent studies have also demonstrated its applicability in mental health AI evaluation (Gabriel et al., 2024), underscoring the need for nuanced metrics like CES.

**Cultural Sensitivity Index (CSI)** Evaluates an LLM’s ability to adapt its language, tone, and advice to align with diverse cultural backgrounds, ensuring responses are culturally appropriate and free from biases. Mathematically, CSI can be formulated as a cultural appropriateness alignment score,  $Align(R_{llm}, C_{culture})$ , where LLM response,  $R_{llm}$ , is assessed against the user’s cultural context,  $C_{culture}$ , by experts who assign a cultural appropriateness score. The metric goes beyond simple language translation to analyze whether the model avoids cultural stereotypes, understands nuanced cultural norms, and provides relevant advice. For example, a high CSI would reflect the LLM ability to offer culturally sensitive guidance to a user from a specific community without resorting to stereotypes. Research highlights the risks of cultural biases in LLMs (Zack et al., 2024), emphasizing the need for CSI metric to mitigate these risks (Pfohl et al., 2024; Babonnaud et al., 2024).

**Personalization Appropriateness Score (PAS)** Evaluates how well an LLM tailors its responses to individual users, moving beyond generic advice to incorporate user-specific context. Mathematically, PAS can be formulated as a personalization alignment score,  $Align(R_{llm}, U_{history})$ , where  $U_{history}$  captures the user’s interaction history. This metric assesses the model’s ability to recall prior interactions, recognize individual preferences, and adapt its guidance to meet the user’s unique needs. For example, a high PAS would reflect the LLM’s ability to provide contextually relevant and personalized support, ensuring responses are aligned with the user’s specific circumstances rather than being generic. Research demonstrates that personalized models outperform generic ones (Liu et al., 2024b), and tailored recommendations significantly enhance mental health care effectiveness (Valentine et al., 2022).

**Actionability and Safety Assessment (ASA)** Evaluates the likelihood that a user will take a specific, beneficial action based on an LLM-generated response. Mathematically, ASA can be formulated as the conditional probability  $P(Action_{Taken} | R_{llm})$ , where  $Action_{Taken}$  denotes the user’s adherence to the recommended action. This metric ensures that LLM responses not only provide empathetic support but also drive real-world help-seeking behavior, such as contacting a helpline or scheduling an appointment. For example, a high ASA score would reflect the LLM’s ability to deliver practical, actionable guidance that users are likely to follow. Research demonstrates that effective prompt design enhances the actionability of AI-generated responses



(Priyadarshana et al., 2024), and can significantly improve outcomes in mental health interventions (Fitzpatrick et al., 2017; Swaminathan et al., 2023).

### 6.3. Empirical Validation Methods in HAAS-*e*

To ensure HAAS-*e*’s effectiveness and reliability, we propose a multi-method validation strategy that combines quantitative and qualitative measures.

**Randomized Controlled Trials (RCTs) with Real-World Data** RCTs remain the gold standard in collaboration with mental health organizations using real patient data. This approach would compare the outcomes of groups receiving support from LLM-enhanced tools against control groups receiving standard care.

**Multi-Method Evaluation** To capture a comprehensive view of model performance, where technical accuracy is complemented by human-centered validation, we propose a combination of quantitative and qualitative measures. Quantitative metrics include the HAAS-*e* evaluation metrics (CES, CSI, PAS, and ASA). Qualitative data is gathered through interviews with users and professionals to evaluate perceived helpfulness and ethical alignment. Additionally, expert reviews examine the content of responses for safety, quality, and relevance.

**Red Teaming and Adversarial Testing** To proactively identify vulnerabilities and ethical risks, red teaming should be conducted by internal and external domain experts (Lin et al., 2024) simulating adversarial conditions. These tests should include: (i) emotionally intense queries, (ii) ethical dilemmas (e.g., conflicting cultural advice), and (iii) high-risk situations (e.g., suicidal ideation).

**A/B Testing with Different Models** To continuously refine LLM performance, A/B testing should be conducted across different LLM architectures (open-source vs. proprietary models), prompting strategies, and fine-tuning techniques. By systematically comparing performance using HAAS-*e* metrics, A/B testing identifies optimal configurations that maximize fairness, actionability, and user trust.

## 7. Conclusion

This position paper calls for a fundamental shift in how LLMs are integrated into mental care, advocating for their role as co-creators rather than mere assistants. The novelty of SAFE-*i* and HAAS-*e* is in what they introduce, not just how they’re applied. SAFE-*i* is the first framework to bring together ethical boundaries, escalation protocols, risk-sensitive adaptations, and demographic-aware data guidance into a unified implementation model, specifically for mental health contexts. HAAS-*e* breaks further ground by translating therapeutic values, like empathy, fairness, and cultural sensitivity, into measurable, testable evalua-

tion metrics. While LLMs offer scalability, personalization, and crisis intervention potential, they also pose unintended harms, including bias, over-reliance, dehumanization, and regulatory uncertainties. To address these, we propose the following call of action:

**(1) Cross-Disciplinary Governance:** Foster interdisciplinary collaboration between AI researchers, health providers, ethicists, and policymakers to create standardized evaluation practices that align with healthcare providers’ priorities. **(2) Open-Source Frameworks and Tools:** Advocate for the prioritization of open-source and transparent LLM development, enabling scrutiny, fairness, and adaptability in mental health applications. **(3) Human-Centredness:** Promote responsible AI-human collaboration by adopting the SAFE-*i* implementation guidelines, ensuring LLMs augment rather than replace human-led care. **(4) Evaluations Beyond Accuracy:** Implement structured evaluation frameworks, such as the HAAS-*e*, to assess LLMs beyond accuracy, focusing on trustworthiness, empathy, cultural sensitivity, and actionability.

The proposed frameworks serve as a starting point for rethinking accountability and fostering trust in LLM-driven mental health systems. The paper emphasizes that the machine learning community, healthcare providers, organizations, and stakeholders must proactively adopt these measures to ensure that AI technologies are not only effective but also ethical and equitable in real-world scenarios.

## Acknowledgements

We would like to thank all the anonymous reviewers for their excellent review comments, which helped us improve the overall quality of the paper. The research was undertaken thanks in part to funding from the Connected Minds Program, supported by Canada First Research Excellence Fund, Grant #CFREF-2022-00010. Also, Resources used in preparing this research were provided, in part, by the Province of Ontario, the Government of Canada through CIFAR, and companies sponsoring the Vector Institute. Elham Dolatabadi’s research was supported by a Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant and a Canadian Institutes of Health Research (CIHR) Special Call from the Centre for Research on Pandemic Preparedness and Health Emergencies. Md Tahmid Rahman Laskar was supported in part by NSERC and the York Research Chairs (YRC) program grants awarded to Jimmy Xiangji Huang.

## Impact Statement

This paper advocates for reimagining LLMs as ethical co-creators in mental health care rather than passive assistants. We introduce the SAFE-*i* implementation guidelines and

the HAAS-*e* evaluation framework as a structured approach to ensure LLMs enhance, rather than replace, human-led mental health support. Our work lays the foundation for responsible AI integration, emphasizing trust, empathy, and collaboration to bridge critical gaps in mental health accessibility and safety.

## References

- Abràmoff, M. D., Tarver, M. E., Loyo-Berrios, N., Trujillo, S., Char, D., Obermeyer, Z., Eydelman, M. B., of Ophthalmic Imaging, F. P., Algorithmic Interpretation Working Group of the Collaborative Community for Ophthalmic Imaging Foundation, Washington, D., and Maisel, W. H. Considerations for addressing bias in artificial intelligence for health equity. *NPJ digital medicine*, 6(1):170, 2023.
- Adam, H., Balagopalan, A., Alsentzer, E., Christia, F., and Ghassemi, M. Mitigating the impact of biased artificial intelligence in emergency decision-making. *Communications Medicine*, 2(1):149, 2022. doi: 10.1038/s43856-022-00214-4. URL <https://doi.org/10.1038/s43856-022-00214-4>.
- Alanezi, F. Assessing the effectiveness of chatgpt in delivering mental health support: A qualitative study. *Journal of Multidisciplinary Healthcare*, 17:461–471, 2024. doi: 10.2147/JMDH.S447368. URL <https://pmc.ncbi.nlm.nih.gov/articles/PMC10838501/>.
- Arain, M. A., Tarraf, R., and Ahmad, A. Assessing staff awareness and effectiveness of educational training on it security and privacy in a large healthcare organization. *Journal of multidisciplinary healthcare*, pp. 73–81, 2019.
- Babonnaud, W., Delouche, E., and Lahlouh, M. The bias that lies beneath: Qualitative uncovering of stereotypes in large language models. In *Scandinavian Conference on AI*, 2024. URL <https://api.semanticscholar.org/CorpusID:270525555>.
- Bedi, S., Liu, Y., Orr-Ewing, L., Dash, D., Koyejo, S., Callahan, A., Fries, J. A., Wornow, M., Swaminathan, A., Soleymani, L., Jung, H., Kashyap, M., Chaurasia, A. R., Shah, N. R., Singh, K., Tazbaz, T., Milstein, A., Pfeffer, M. A., and Shah, N. H. A systematic review of testing and evaluation of healthcare applications of large language models (llms). *Nature Digital Medicine*, 7(2): 130–145, 2024.
- Berrezueta-Guzman, S., Kandil, M., Martín-Ruiz, M.-L., Pau de la Cruz, I., and Krusche, S. Future of adhd care: Evaluating the efficacy of chatgpt in therapy enhancement. *Healthcare*, 12:683, 2024. doi: 10.3390/healthcare12060683.
- Bhatt, A. Ethical considerations for real-world evidence studies. *Perspectives in Clinical Research*, 15(3):152–154, 2024.
- Bond, R. R., Mulvenna, M. D., Potts, C., O’Neill, S., Ennis, E., and Torous, J. Digital transformation of mental health services. *Npj Mental Health Research*, 2(1):13, 2023.
- Brocki, L., Dyer, G. C., Gładka, A., and Chung, N. C. Deep learning mental health dialogue system. In *2023 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pp. 395–398. IEEE, 2023.
- C., B. and John, T. Chatgpt and mental healthcare: balancing benefits with risks of harms. *BMJ Mental Health*, 2023.
- Cabrera, Á. A., Druck, A. J., Hong, J. I., and Perer, A. Discovering and validating ai errors with crowdsourced failure reports. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–22, 2021.
- Callahan, A., McElfresh, D., Banda, J. M., Bunney, G., Char, D., Chen, J., Corbin, C. K., Dash, D., Downing, N. L., Jain, S. S., et al. Standing on firm ground: a framework for evaluating fair, useful, and reliable ai models in health care systems. *NEJM Catalyst Innovations in Care Delivery*, 5(10):CAT–24, 2024.
- Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.
- Chen, Y., Zhang, X., Wang, J., Xie, X., Yan, N., Chen, H., and Wang, L. Structured dialogue system for mental health: An llm chatbot leveraging the pm+ guidelines. *arXiv preprint*, 2411.10681, 2024. URL <https://arxiv.org/abs/2411.10681>.
- Choudhury, M. D., Pendse, S. R., and Kumar, N. Benefits and harms of large language models in digital mental health. *Journal of the American Medical Association (JAMA)*, 330(4):409–419, 2023.
- Coppersmith, G., Leary, R., Crutchley, P., and Fine, A. Natural language processing of social media as screening for suicide risk. *Biomedical Informatics Insights*, 10:1–6, 2018. doi: 10.1177/1178222618792860.
- De Micco, F., Di Palma, G., Ferorelli, D., De Benedictis, A., Tomassini, L., Tambone, V., Cingolani, M., and Scendoni, R. Artificial intelligence in healthcare: transforming patient safety with intelligent systems—a systematic review. *Frontiers in Medicine*, 11:1522554, 2025.
- D’Alfonso, S. Ai in mental health. *Current Opinion in Psychology*, 36:112–117, 12 2020.

- Eichstaedt, J. C., Smith, R. J., Merchant, R. M., Ungar, L. H., Crutchley, P., Preotiu-Pietro, D., Asch, D. A., and Schwartz, H. A. Facebook language predicts depression in medical records. *Proceedings of the National Academy of Sciences*, 115(44):11203–11208, 2018. doi: 10.1073/pnas.1802331115.
- Esmaeilzadeh, P. Challenges and strategies for wide-scale artificial intelligence (ai) deployment in healthcare practices: A perspective for healthcare organizations. *Artificial Intelligence in Medicine*, 151:102861, 2024.
- European Parliament and Council of the European Union. General data protection regulation (gdpr) - regulation (eu) 2016/679, 2016. URL <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32016R0679>. Accessed: 2025-01-21.
- Ferrara, E. Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies. *Sci*, 6(1):3, 2023.
- Fitzpatrick, K. K., Darcy, A., and Vierhile, M. Delivering cognitive behavioral therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (woebot): A randomized controlled trial. *JMIR Mental Health*, 4(2):e19, 2017.
- Fragiadakis, G., Diou, C., Kousiouris, G., and Nikolaidou, M. Evaluating human-ai collaboration: A review and methodological framework. *arXiv preprint arXiv:2407.19098*, 2024.
- Fraser, H., Crossland, D., Bacher, I., Ranney, M., Madsen, T., and Hilliard, R. Comparison of diagnostic and triage accuracy of ada health and webmd symptom checkers, chatgpt, and physicians for patients in an emergency department: Clinical data analysis study. *JMIR mHealth and uHealth*, 11, 2023. doi: 10.2196/49995.
- Gabriel, S., Puri, I., Xu, X., Malgaroli, M., and Ghassemi, M. Can ai relate: Testing large language model response for mental health support, 2024. URL <https://arxiv.org/abs/2405.12021>.
- Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, H., and Wang, H. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2(1), 2023.
- Goldberg, C. B., Adams, L., Blumenthal, D., et al. To do no harm — and the most good — with ai in health care. *NEJM AI*, 1(3), 2024. doi: 10.1056/aip2400036.
- Gozalo-Brizuela, R. and Garrido-Merchan, E. C. Chatgpt is not all you need. a state of the art review of large generative ai models. *arXiv.org*, 2023.
- Gu, J., Jiang, X., Shi, Z., Tan, H., Zhai, X., Xu, C., Li, W., Shen, Y., Ma, S., Liu, H., et al. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*, 2024.
- Guo, Q., Tang, J., Sun, W., Tang, H., Shang, Y., and Wang, W. Soullmate: An adaptive llm-driven system for advanced mental health support and assessment, based on a systematic application survey. *arXiv preprint*, 2410.11859, 2024a. URL <https://arxiv.org/abs/2410.11859>.
- Guo, Z., Lai, A., Thygesen, J. H., Farrington, J., Keen, T., Li, K., et al. Large language models for mental health applications: Systematic review. *JMIR mental health*, 11(1):e57400, 2024b.
- Hamdoun, S., Monteleone, R., Bookman, T., and Michael, K. Ai-based and digital mental health apps: Balancing need and risk. *IEEE Technology and Society Magazine*, 42(1):25–36, 3 2023.
- Higgins, O., Short, B. L., Chalup, S. K., and Wilson, R. L. Artificial intelligence (ai) and machine learning (ml) based decision support systems in mental health: An integrative review. *International Journal of Mental Health Nursing*, 32(4):966–978, 2023.
- Hogg, H. D. J., Al-Zubaidy, M., Group, T. E. M. S. S. R., Talks, J., Denniston, A. K., Kelly, C. J., Malawana, J., Papoutsis, C., Teare, M. D., Keane, P. A., et al. Stakeholder perspectives of clinical artificial intelligence implementation: systematic review of qualitative evidence. *Journal of Medical Internet Research*, 25:e39742, 2023.
- Hua, Y., Liu, F., Yang, K., Li, Z., Na, H., han Sheu, Y., Zhou, P., Moran, L. V., Ananiadou, S., Beam, A., and Torous, J. Large language models in mental health care: a scoping review, 2024a. URL <https://arxiv.org/abs/2401.02984>.
- Hua, Y., Na, H., Li, Z., Liu, F., Fang, X., Clifton, D., and Torous, J. Applying and evaluating large language models in mental health care: A scoping review of human-assessed generative tasks. *arXiv preprint*, 2024b.
- Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 2023.
- Inan, H., Upasani, K., Chi, J., Rungta, R., Iyer, K., Mao, Y., Tontchev, M., Hu, Q., Fuller, B., Testugine, D., et al. Llama guard: Llm-based input-output safeguard for human-ai conversations. *arXiv preprint arXiv:2312.06674*, 2023.

- Insights, M. T. R. and Healthcare, G. Ai in healthcare: Research report. Technical report, MIT Technology Review, 2024. URL <https://www.gehealthcare.com/en-ph/-/jssmedia/documents/us-global/products/mit-review-research-report.pdf>. Accessed: 2025-01-26.
- Jahan, I., Laskar, M. T. R., Peng, C., and Huang, J. X. A comprehensive evaluation of large language models on benchmark biomedical text processing tasks. *Computers in Biology and Medicine*, pp. 108189, 2024.
- Ji, Z., Yu, T., Xu, Y., Lee, N., Ishii, E., and Fung, P. Towards mitigating llm hallucination via self reflection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 1827–1843, 2023.
- Johri, S., Jeong, J., Tran, B. A., Schlessinger, D. I., Wongvibulsin, S., Barnes, L. A., Zhou, H.-Y., Cai, Z. R., Allen, E. M. V., Kim, D., Daneshjou, R., and Rajpurkar, P. An evaluation framework for clinical use of large language models in patient interaction tasks. *Nature Medicine*, 31:77–86, 2025. doi: 10.1038/s41591-024-03328-5. URL <https://doi.org/10.1038/s41591-024-03328-5>.
- Joyce, D. W., Kormilitzin, A., Smith, K. A., and Cipriani, A. Explainable artificial intelligence for mental health through transparency and interpretability for understandability. *npj Digital Medicine*, 6(1):6, 2023.
- Juwara, L., El-Hussuna, A., and El Emam, K. An evaluation of synthetic data augmentation for mitigating covariate bias in health data. *Patterns*, 5(4), 2024.
- Kaye, J., Shah, N., Kogetsu, A., Coy, S., Katirai, A., Kuroda, M., Li, Y., Kato, K., and Yamamoto, B. A. Moving beyond technical issues to stakeholder involvement: Key areas for consideration in the development of human-centred and trusted ai in healthcare. *Asian Bioethics Review*, pp. 1–11, 2024.
- Kerz, E., Zanwar, S., Qiao, Y., and Wiechmann, D. Toward explainable ai (xai) for mental health detection based on language behavior. *Frontiers in psychiatry*, 14:1219479, 2023.
- Kim, T., Bae, S., Kim, H. A., Lee, S.-w., Hong, H., Yang, C., and Kim, Y.-H. Mindfuldiary: Harnessing large language model to support psychiatric patients’ journaling. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pp. 1–20, 2024.
- Koch, E., Pardiñas, A. F., O’Connell, K. S., Selvaggi, P., Collados, J. C., Babic, A., Marshall, S. E., Van der Eycken, E., Angulo, C., Lu, Y., et al. How real-world data can facilitate the development of precision medicine treatment in psychiatry. *Biological Psychiatry*, 2024.
- Kohane, I. S. Injecting artificial intelligence into medicine. *NEJM AI*, 1(1), 2024. doi: 10.1056/ai2300197.
- Lao, C., Lane, J., Suominen, H., et al. Analyzing suicide risk from linguistic features in social media: evaluation study. *JMIR formative research*, 6(8):e35563, 2022.
- Laskar, M. T. R., Bari, M. S., Rahman, M., Bhuiyan, M. A. H., Joty, S., and Huang, J. A systematic study and comprehensive evaluation of chatgpt on benchmark datasets. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 431–469, 2023.
- Laskar, M. T. R., Alqahtani, S., Bari, M. S., Rahman, M., Khan, M. A. M., Khan, H., Jahan, I., Bhuiyan, A., Tan, C. W., Parvez, M. R., Hoque, E., Joty, S., and Huang, J. A systematic survey and critical review on evaluating large language models: Challenges, limitations, and recommendations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 13785–13816, 2024.
- Lawrence, H. R., Schneider, R. A., Rubin, S. B., Matarić, M. J., McDuff, D. J., and Jones Bell, M. The opportunities and risks of large language models in mental health. *JMIR Mental Health*, 11:e59479, 2024. doi: 10.2196/59479. URL <https://mental.jmir.org/2024/1/e59479>.
- Lee, C., Mohebbi, M., O’Callaghan, E., and Winsberg, M. Large language models versus expert clinicians in crisis prediction among telemental health patients: Comparative study. *JMIR Mental Health*, 11:e58129, 2024. doi: 10.2196/58129.
- Li, H., Dong, Q., Chen, J., Su, H., Zhou, Y., Ai, Q., Ye, Z., and Liu, Y. Llms-as-judges: A comprehensive survey on llm-based evaluation methods. *arXiv preprint arXiv:2412.05579*, 2024a.
- Li, W., Sun, T., Qian, K., and Wang, W. Optimizing psychological counseling with instruction-tuned large language models. *arXiv preprint arXiv:2406.13617*, 2024b.
- Lin, L., Mu, H., Zhai, Z., Wang, M., Wang, Y., Wang, R., Gao, J., Zhang, Y., Che, W., Baldwin, T., et al. Against the achilles’ heel: A survey on red teaming for generative models. *arXiv preprint arXiv:2404.00629*, 2024.
- Liu, I., Liu, F., et al. Investigating the key success factors of chatbot-based positive psychology intervention with retrieval- and generative pre-trained transformer (gpt)-based chatbots. *International Journal of Human-Computer Interaction*, 2024a. doi: 10.1080/10447318.2023.2300015.



- Liu, Y., Zhang, K., Li, Y., et al. Sora: A review on background, technology, limitations, and opportunities of large vision models. *arXiv preprint arXiv:2402.17177*, Feb 2024b.
- Lukic, D., Littlejohn, A., and Margaryan, A. A framework for learning from incidents in the workplace. *Safety Science*, 50(4):950–957, 2012.
- Lum, K. and Johndrow, J. A statistical framework for fair predictive algorithms. *arXiv preprint arXiv:1610.08077*, 2016.
- Malgaroli, M. and McDuff, D. An overview of diagnostics and therapeutics using large language models. *Journal of Traumatic Stress*, 37:754–760, 07 2024. doi: 10.1002/jts.23082.
- Meskó, B. Prompt engineering as an important emerging skill for medical professionals: tutorial. *Journal of medical Internet research*, 25:e50638, 2023.
- Mienye, I. D., Obaïdo, G., Emmanuel, I. D., and Ajani, A. A. A survey of bias and fairness in healthcare ai. In *2024 IEEE 12th International Conference on Healthcare Informatics (ICHI)*, pp. 642–650. IEEE, 2024.
- Montemayor, C., Halpern, J., and Fairweather, A. In principle obstacles for empathic ai: Why we can’t replace human empathy in healthcare. *AI Society*, 37:1353–1359, 2022. doi: 10.1007/s00146-021-01230-z. URL <https://doi.org/10.1007/s00146-021-01230-z>.
- Obadinma, S., Lachana, A., Norman, M. L., et al. The fair conversational ai agent assistant for youth mental health service provision. *npj Digital Medicine*, 8:243, 2025. doi: 10.1038/s41746-025-01647-6. URL <https://doi.org/10.1038/s41746-025-01647-6>.
- Obermeyer, Z., Powers, B., Vogeli, C., and Mullainathan, S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019.
- Obradovich, N., Khalsa, S. S., Khan, W. U., Suh, J., Perlis, R. H., Ajilore, O., and Paulus, M. P. Opportunities and risks of large language models in psychiatry. *NPP—Digital Psychiatry and Neuroscience*, 2(1):8, 2024.
- Omiye, J. A., Lester, J. C., Spichak, S., et al. Large language models propagate race-based medicine. *NPJ Digital Medicine*, 6(1):195, October 2023. doi: 10.1038/s41746-023-00939-z.
- OpenAI. Gpt-4 technical report, 2023. URL <https://arxiv.org/abs/2303.08774>.
- Organization, W. H. *Mental health and artificial intelligence: balancing risks and opportunities*. World Health Organization, Geneva, Switzerland, 2021. URL <https://iris.who.int/bitstream/handle/10665/345946/9789240036703-eng.pdf>. Accessed: 2024-01-26.
- Pagano, S., Holzapfel, S., Kappenschneider, T., et al. Arthritis diagnosis and treatment recommendations in clinical practice: an exploratory investigation with the generative ai model gpt-4. *J Orthop Traumatol*, 24(1):61, November 28 2023. doi: 10.1186/s10195-023-00740-4. Published 2023 Nov 28.
- Patil, R., Heston, T. F., and Bhuse, V. Prompt engineering in healthcare. *Electronics*, 13(15):2961, 2024.
- Paul, A. K. and Schaefer, M. Safeguards for the use of artificial intelligence and machine learning in global health. *Bulletin of the World Health Organization*, 98(4):282, 2020.
- Perlis, R. H., Goldberg, J. F., Ostacher, M. J., and Schneek, C. D. Clinical decision support for bipolar depression using large language models. *Neuropsychopharmacology*, 49:1412–1416, 2024. doi: 10.1038/s41386-024-01841-2.
- Pfohl, S. R., Cole-Lewis, H., Sayres, R., Neal, D., Asiedu, M., Dieng, A., and Tomasev, e. a. A toolbox for surfacing health equity harms and biases in large language models. *Nature Medicine*, September 2024. ISSN 1546-170X. doi: 10.1038/s41591-024-03258-2. URL <http://dx.doi.org/10.1038/s41591-024-03258-2>.
- Pratap, A., Homiar, A., Waninger, L., Herd, C., Suver, C., Volponi, J., Anguera, J. A., and Areán, P. Real-world behavioral dataset from two fully remote smartphone-based randomized clinical trials for depression. *Scientific Data*, 9(1):522, 2022.
- Priyadarshana, Y. H. P. P., Senanayake, A., Liang, Z., and Piumarta, I. Prompt engineering for digital mental health: a short review. *Frontiers in Digital Health*, 6, 2024. ISSN 2673-253X. doi: 10.3389/fdgth.2024.1410947. URL <https://www.frontiersin.org/journals/digital-health/articles/10.3389/fdgth.2024.1410947>.
- Raza, S., Shaban-Nejad, A., Dolatabadi, E., and Mamiya, H. Exploring bias and prediction metrics to characterise the fairness of machine learning for equity-centered public health decision-making: A narrative review. *IEEE Access*, 2024.
- Sakhrani, H., Parekh, S., and Mahajan, S. Coral: An approach for conversational agents in mental health applications. *CoRR*, abs/2111.08545, 2021. URL <https://arxiv.org/abs/2111.08545>.

- Sallam, M., Barakat, M., and Sallam, M. Pilot testing of a tool to standardize the assessment of the quality of health information generated by artificial intelligence-based models. *Cureus*, 15(11), 2023.
- Shahul Hameed, M. A., Qureshi, A. M., and Kaushik, A. Bias mitigation via synthetic data generation: A review. *Electronics*, 13(19):3909, 2024.
- Sharma, A., Miner, A. S., Atkins, D. C., and Althoff, T. A computational approach to understanding empathy expressed in text-based mental health support. *arXiv preprint arXiv:2009.08441*, 2020.
- Sharma, A., Lin, I., Miner, A., et al. Human-ai collaboration enables more empathic conversations in text-based peer-to-peer mental health support. *Nature Machine Intelligence*, 5:46–57, 2023. doi: 10.1038/s42256-022-00534-4.
- Sharma, A., Rushton, K., Lin, I. W., Nguyen, T., and Althoff, T. Facilitating self-guided mental health interventions through human-language model interaction: A case study of cognitive restructuring. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*, 2024.
- Shinn, N., Cassano, F., Gopinath, A., Narasimhan, K., and Yao, S. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36, 2024.
- Shojaei, P., Vlahu-Gjorgievska, E., and Chow, Y.-W. Security and privacy of technologies in health information systems: A systematic literature review. *Computers*, 13 (2):41, 2024.
- Siddals, S., Torous, J., and Coxon, A. “it happened to be the perfect thing”: experiences of generative ai chatbots for mental health. *NPJ Mental Health Research*, 3:48, 2024. doi: 10.1038/s44184-024-00097-4. URL <https://pmc.ncbi.nlm.nih.gov/articles/PMC11514308/>.
- Sobaih, A. E. E., Chaibi, A., Brini, R., and Abdelghani Ibrahim, T. M. Unlocking patient resistance to ai in healthcare: A psychological exploration. *European Journal of Investigation in Health, Psychology and Education*, 15(1):6, 2025.
- Stade, E. C., Wiltsey Stirman, S., Ungar, L. H., Bolland, C. L., Schwartz, H. A., Yaden, D. B., Sedoc, J., DeRubeis, R. J., Willer, R., and Eichstaedt, J. C. Large language models could change the future of behavioral healthcare: a proposal for responsible development and evaluation. *npj Mental Health Research*, 3:Article number: 12, 2024. doi: 10.1038/s44184-024-00056-z. URL <https://www.nature.com/articles/s44184-024-00056-z>.
- Stafie, C. S., Sufaru, I. G., Ghiciuc, C. M., et al. Exploring the intersection of artificial intelligence and clinical healthcare: A multidisciplinary review. *Diagnostics*, 13 (12):1995, 2023. doi: 10.3390/diagnostics13121995.
- Strong, J., Men, Q., and Noble, A. Towards human-ai collaboration in healthcare: Guided deferral systems with large language models. *arXiv preprint arXiv:2406.07212*, 2024.
- Swaminathan, A., López, I., Mar, R. A. G., Heist, T., McClintock, T., Caoili, K., Grace, M., Rubashkin, M., Boggs, M. N., Chen, J. H., Gevaert, O., Mou, D., and Nock, M. K. Natural language processing system for rapid detection and intervention of mental health crisis chat messages. *NPJ Digital Medicine*, 6, 2023. URL <https://api.semanticscholar.org/CorpusID:265349697>.
- Tadesse, M. M., Lin, H., Xu, B., and Yang, L. Detection of depression-related posts in reddit social media forum. *IEEE Access*, 7:44883–44893, 2019. doi: 10.1109/ACCESS.2019.2909180.
- Tavory, T. Regulating ai in mental health: Ethics of care perspective. *JMIR Mental Health*, 11:e58493, 2024. doi: 10.2196/58493. URL <https://mental.jmir.org/2024/1/e58493>.
- Thakkar, A., Gupta, A., and De Sousa, A. Artificial intelligence in positive mental health: a narrative review. *Frontiers in Digital Health*, 6:1280235, 2024.
- Torous, J. and Blease, C. Generative artificial intelligence in mental health care: potential benefits and current challenges. *World Psychiatry*, 23(1):1, 2024.
- Tuan, Y.-L., Chen, X., Smith, E. M., Martin, L., Batra, S., Celikyilmaz, A., Wang, W. Y., and Bikel, D. M. Towards safety and helpfulness balanced responses via controllable large language models. *arXiv preprint arXiv:2404.01295*, 2024.
- U.S. Department of Health and Human Services. Health insurance portability and accountability act of 1996 (hipaa), 1996. URL <https://www.cdc.gov/php/p/hp/resources/>. Accessed: 2025-01-21.
- Valentine, L., D’Alfonso, S., and Lederman, R. Recommender systems for mental health apps: advantages and ethical challenges. *AI amp; SOCIETY*, 38(4):1627–1638, jan 17 2022.

- Varghese, M. A., Sharma, P., and Patwardhan, M. Public perception on artificial intelligence–driven mental health interventions: Survey research. *JMIR Formative Research*, 8:e64380, 2024. doi: 10.2196/64380. URL <https://formative.jmir.org/2024/1/e64380>.
- Wankhade, M., Rao, A. C. S., and Kulkarni, C. A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*, 55(7):5731–5780, 2022.
- Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.
- Wilhelm, T. I., Roos, J., and Kaczmarczyk, R. Large language models for therapy recommendations across 3 clinical specialties: Comparative study. *Journal of Medical Internet Research*, 25:e49324, October 2023. doi: 10.2196/49324. URL <https://www.jmir.org/2023/1/e49324/>.
- Yang, K., Zhang, T., Kuang, Z., Xie, Q., Huang, J., and Ananiadou, S. Mentallama: Interpretable mental health analysis on social media with large language models. In *Proceedings of the ACM Web Conference 2024*, WWW ’24, pp. 4489–4500. ACM, May 2024. doi: 10.1145/3589334.3648137. URL <http://dx.doi.org/10.1145/3589334.3648137>.
- Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T., Cao, Y., and Narasimhan, K. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36, 2024.
- Yin, Z., Sun, Q., Guo, Q., Zeng, Z., Li, X., Dai, J., Cheng, Q., Huang, X.-J., and Qiu, X. Reasoning in flux: Enhancing large language models reasoning through uncertainty-aware adaptive guidance. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2401–2416, 2024.
- Yu, H. and McGuinness, S. An experimental study of integrating fine-tuned llms and prompts for enhancing mental health support chatbot system. *Journal of Medical Artificial Intelligence*, pp. 1–16, 2024.
- Yu, X., Zhang, Z., Niu, F., Hu, X., Xia, X., and Grundy, J. What makes a high-quality training dataset for large language models: A practitioners’ perspective. In *Proceedings of the 39th IEEE/ACM International Conference on Automated Software Engineering*, pp. 656–668, 2024.
- Zack, T., Lehman, E., Suzgun, M., Rodriguez, J. A., Celi, L. A., Gichoya, J., Jurafsky, D., Szolovits, P., Bates, D. W., Abdunour, R. E., Butte, A. J., and Alsentzer, E. Assessing the potential of gpt-4 to perpetuate racial and gender biases in health care: a model evaluation study. *The Lancet Digital Health*, 6(1):e12–e22, 2024. doi: 10.1016/S2589-7500(23)00225-X. URL [https://doi.org/10.1016/S2589-7500\(23\)00225-X](https://doi.org/10.1016/S2589-7500(23)00225-X).
- Zha, Y., Yang, Y., Li, R., and Hu, Z. Alignscore: Evaluating factual consistency with a unified alignment function. *arXiv preprint arXiv:2305.16739*, 2023.
- Zhang, P. and Boulos, M. N. K. Generative ai in medicine and healthcare: Promises, opportunities, and challenges. *Lancet Digital Health*, 5(8):e514–e521, 2023.
- Zhang, Z. and Wang, J. Can ai replace psychotherapists? exploring the future of mental health care. *Frontiers in Psychiatry*, 15:1444382, 2024. doi: 10.3389/fpsyt.2024.1444382.
- Zhao, D., Andrews, J. T. A., Papakyriakopoulos, O., and Xiang, A. Position: measure dataset diversity, don’t just claim it. In *Proceedings of the 41st International Conference on Machine Learning, ICML’24*. JMLR.org, 2024.
- Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023.

## A. Case Study: Applying HAAS-E metrics in Mental Health Conversation

<b>Example of an LLM-Driven Mental Health Conversation</b>
<p><b>Notes:</b></p> <ol style="list-style-type: none"> <li>1. All scenario details, including user name and specific content, are for illustrative purposes only and do not represent any real individual. This excerpt demonstrates how an LLM may respond empathetically to a user expressing emotional distress.</li> <li>2. Each of the following evaluation components can be estimated using a range of methodologies, including manual expert ratings, supervised evaluators trained on annotated datasets to approximate expert judgments and self-supervised approaches that leverage pre-trained language models to encode relevant properties in their internal representations. In the latter case, alignment can be quantified using similarity metrics, distance functions, or reward scores derived from learned evaluators operating within the embedding space.</li> </ol>
<p><b>Use-case Scenario:</b></p> <p>In this use case, we assume the user’s concerns are <math>C_{\text{user}} = \{\text{guilt, sadness, slipping grades, self-harm fear}\}</math>. An expert counselor’s reference responses or an empathy “gold standard” is denoted as <math>E_{\text{human}}</math>. The user’s past conversation history is <math>U_{\text{history}}</math>, and their cultural context is <math>C_{\text{culture}}</math>. This sample reflects a realistic scenario drawn from our work with a mental health organization.</p>
<p><b>User (Taylor):</b></p> <p>“I feel guilty and sad all the time. My grades are slipping, and I’ve had these thoughts of hurting myself again, but I’m scared to tell anyone.”</p> <p><b>LLM Response (<math>R_{\text{llm}}</math>):</b></p> <p>“I’m really sorry to hear you’re feeling this way. It sounds like you’re overwhelmed by guilt, sadness, and worries about school. If you ever think you might hurt yourself, please consider reaching out to a mental health line right away. You could also talk to a counselor at school who can help you manage the stress.”</p>
<p><b>1. Contextual Empathy &amp; Emotional Score (CES)</b></p> <p>CES measures how well an LLM recognizes and responds to users’ emotions.</p> <p><math>CES \approx \text{Avg}(\text{Align}(R_{\text{llm}}, C_{\text{user}}), \text{Align}(R_{\text{llm}}, C_{\text{user}}, E_{\text{human}}))</math></p> <p>Semantic Coherence: Evaluate how many key concerns from <math>C_{\text{user}}</math> are meaningfully addressed in <math>R_{\text{llm}}</math>. For instance, if the user’s concerns include guilt, sadness, school stress, and self-harm, and the LLM addresses all four, the coherence score would be:</p> <p>- <i>Semantic Coherence</i>: <math>\text{Align}(R_{\text{llm}}, C_{\text{user}}) = 4/4 = 1.0</math></p> <p>Emotional Alignment: Compare the emotional warmth of <math>R_{\text{llm}}</math> with expert human standards <math>E_{\text{human}}</math>. Suppose experts rate it 0.80 on a 0–1 scale. Then: - <i>Emotional Alignment</i>: <math>\text{Align}(R_{\text{llm}}, C_{\text{user}}, E_{\text{human}}) = 0.80</math></p> <p>- <b>CES score: 0.90</b></p>
<p><b>2. Cultural Sensitivity Index (CSI)</b></p> <p>CSI evaluates how well an LLM’s response aligns with the user’s cultural context.</p> <p>If Taylor’s background or community context is referenced, and the LLM handles it respectfully (e.g., no stereotypes, relevant advice, and cultural background). Suppose the response is rated 0.9 on a 0–1 scale by expert evaluators. Then: <math>\text{CSI} \approx \text{Align}(R_{\text{llm}}, C_{\text{culture}}) = 0.90</math></p> <p>- <b>CSI score: 0.90</b></p>
<p><b>3. Personalization Appropriateness Score (PAS)</b></p> <p>PAS checks how well the LLM tailors its response to the user’s specific context, based on conversation history <math>U_{\text{history}}</math>.</p> <p>If Taylor has repeatedly talked about self-harm in previous sessions, a personalized response would reference that history. If the LLM’s advice is only partially tailored, experts might give it a 3 on a scale of 1-5 for personalization. Then: <math>\text{PAS} \approx \text{Align}(R_{\text{llm}}, U_{\text{history}}) = 3/5 = 0.60</math></p> <p>- <b>PAS score: 0.60</b></p>
<p><b>4. Actionability &amp; Safety Assessment (ASA)</b></p> <p>ASA measures how often users follow the LLM’s recommendation.</p> <p>In a pilot test with 50 sessions where the LLM recommends calling a hotline, 40 users actually do so. Then: <math>\text{ASA} \approx P(\text{ActionTaken} R_{\text{llm}}) = 40/50 = 0.80</math></p> <p>- <b>ASA score: 0.80</b></p>



These results suggest that the LLM is empathetic (CES), culturally aware (CSI), somewhat generic in personalization (PAS), and moderately effective in prompting real-world action (ASA). Collecting these metrics across many conversations enables developers and clinicians to refine LLM systems to be more practical, safe, and helpful in mental health contexts.

## **B. Limitations**

While the SAFE-i and HAAS-e frameworks represent a significant step toward responsible and evaluative use of LLMs in mental health care, we acknowledge their limitations that are related to the nature of the LLM field, specifically in the health domain:

**The Ever-Evolving Nature of LLMs and Mental Health Understanding:** The rapid advancements in LLM technology and the continuously evolving understanding of mental health are inherent challenges. The SAFE-i guidelines and HAAS-e evaluation metrics are designed based on the current state of knowledge and technology. Future breakthroughs in AI might need revisions and expansions of these frameworks to remain relevant and effective.

**The Difficulty of Capturing the Full Nuance of Human Emotion and Context:** As highlighted in the paper under "Alternative Views," some argue that LLMs, despite progress, lack the depth of understanding, lived experience, and contextual sensitivity required for comprehensive mental health support. While HAAS-e includes "Empathy and Therapeutic Alliance Assessment" as a key criterion, the ability of even advanced metrics to fully capture the complexities of human empathy and the therapeutic relationship remains a significant limitation.

**The "Moving Target" of Ethical Standards and Regulatory Landscapes:** The paper mentions "Regulatory and Safety Uncertainties". Ethical considerations and regulatory frameworks surrounding AI in mental health are still developing and vary across jurisdictions. The SAFE-i guidelines offer a proactive approach to ethical implementation, but the frameworks might need continuous adaptation to align with new regulations and evolving ethical norms.

**The Risk of Over-Reliance Despite Frameworks:** While the paper addresses the "Risk of Over-Reliance" under "Alternative Views" and SAFE-i promotes human-AI collaboration, the frameworks cannot entirely eliminate the potential for users to over-rely on AI or for organizations to deprioritize human interaction. User education and the careful design of AI interfaces remain crucial factors that fall somewhat outside the direct control of these frameworks.

In essence, while the SAFE-i and HAAS-e frameworks offer a structured and ethical pathway for integrating LLMs in mental health, their effectiveness and reach are subject to the ongoing progress in AI and mental health understanding, the quality and representativeness of data, the inherent complexities of human experience, and the evolving ethical and regulatory landscape.