# Guardians of Trust: Risks and Opportunities for LLMs in Mental Health

**Miguel Baidal**
ELLIS Alicante
miguel@ellisalicante.org

**Erik Derner**
ELLIS Alicante
erik@ellisalicante.org

**Nuria Oliver**
ELLIS Alicante
nuria@ellisalicante.org

## Abstract

The integration of large language models (LLMs) into mental health applications offers promising opportunities for positive social impact. However, it also presents critical risks. While previous studies have often addressed these challenges and risks individually, a broader and multi-dimensional approach is still lacking. In this paper, we introduce a taxonomy of the main challenges related to the use of LLMs for mental health and propose a structured, comprehensive research agenda to mitigate them. We emphasize the need for explainable, emotionally aware, culturally sensitive, and clinically aligned systems, supported by continuous monitoring and human oversight. By placing our work within the broader context of natural language processing (NLP) for positive impact, this research contributes to ongoing efforts to ensure that technological advances in NLP responsibly serve vulnerable populations, fostering a future where mental health solutions improve rather than endanger well-being.

## 1 Introduction

Mental health is essential for a healthy life. However, mental health disorders are a growing global crisis. According to the World Health Organization Mental Health Report[1], it was estimated in 2019 that 970 million people worldwide suffered from a mental health disorder, which corresponds to a prevalence of 13 %. Despite the increasing need for mental health support, access remains limited. Over 75 % of people in low-income countries lack adequate services, and even in high-income ones like the United States, barriers such as cost, a lack of professionals, and social stigma still remain (Coombs et al., 2021).

In this context, large language models (LLMs) offer a new way to help reduce the existing gaps, not by replacing traditional professional support, but by providing an additional option. Even though there are chatbots created specifically for mental health, it is now very common for people to use general-purpose LLMs as informal advisors for all sorts of questions, including mental health ones. However, using these technologies also raises ethical and safety questions that need to be carefully considered.

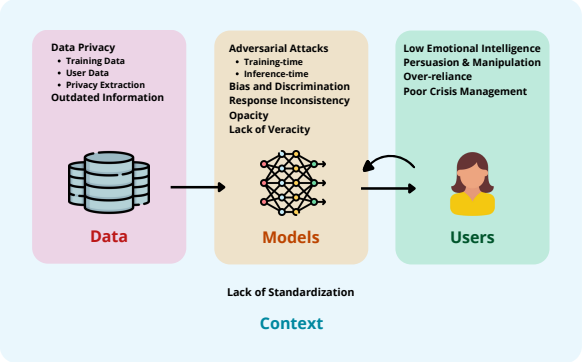


Figure 1: Overview of LLM-related risks in mental health applications as per the proposed taxonomy.

Building on previous taxonomies in mental health, such as those proposed by Hua et al. (2024) and Guo et al. (2024), we note that these studies address several important risks and challenges, but their scope remains rather limited. In contrast, our approach introduces a multi-dimensional taxonomy that considers the full spectrum of risks. This taxonomy is both comprehensive and well-structured, supported by a visual schema (Figure 1) and a clearly organized framework. Specifically, it consists of four dimensions: (1) data-related risks; (2) model-level concerns; (3) user-facing risks; and (4) contextual elements. Building on this taxonomy, we propose a forward-looking research agenda to guide the safe use of LLMs for mental health. We emphasize that LLMs should augment, not replace, clinical judgment, and that

[1] https://iris.who.int/bitstream/handle/10665/356119/9789240049338-eng.pdf



*Proceedings of the Fourth Workshop on NLP for Positive Impact (NLP4PI)*, pages 11–22
July 31, 2025 ©2025 Association for Computational Linguistics

these systems must be designed for continuous human oversight. Together, these elements guide the safe and responsible deployment of LLMs in mental health applications, setting our work apart from existing surveys.

## 2 Risk Taxonomy

This section presents a taxonomy of risks that is both application-agnostic and transversal, organized according to the life cycle of LLM development and deployment: data, model, user, and context.

### 2.1 Data-related Risks

#### 2.1.1 Data Privacy Breaches

**Training Data** Privacy concerns are particularly acute in the domain of integration of LLMs into mental health applications, where both training and user data often involve highly personal and emotionally sensitive information. Despite anonymization efforts, LLMs can infer personal information from training data. Nyffenegger et al. (2023) showed that minimal contextual clues in anonymized datasets can enable re-identification when provided with enough contextual information. Additionally, LLMs tend to memorize training content, particularly as model size increases: larger models have been found to be more prone to data leakage and vulnerable to inference attacks due to their high memorization capacity and instruction-following ability (Li et al., 2024b). In the case of mental health, if the LLMs have been fine-tuned with specific mental health-related data, the impact of a data privacy breach could be particularly severe. Since mental health data is often highly personal and regulated under strict privacy laws (*e.g.*, HIPAA in the U.S. or GDPR in Europe), such breaches could also result in legal liabilities for organizations deploying these models.

**User Data** Even though users often assume their conversations with chatbots are ephemeral and overlook the possibility of long-term storage (Gumusel et al., 2024), user inputs may be retained and analyzed in non-transparent ways. Furthermore, storing user data increases the risk of linkability, where seemingly trivial information may be cross-referenced to reveal sensitive details. A privacy breach might expose sensitive user information, such as disclosed symptoms. This could lead to serious ethical, legal, and personal consequences, including stigma, discrimination, or emotional distress for affected individuals.

**Privacy Extraction** LLMs are also susceptible to attacks that aim to extract private data by means of both membership (Galli et al., 2024) – *i.e.*, determining whether someone's data was used in training – and attribute (Sabour et al., 2024) – *i.e.*, predicting hidden user traits based on output behavior – inference attacks. In the case of LLMs used for mental health, model inversion techniques have been used to reconstruct training input or infer psychiatric conditions (Li et al., 2024b), with potentially devastating consequences for the users.

#### 2.1.2 Outdated Information

As LLMs rely on static training data, they could provide outdated clinical guidance. Recent research has found that over 20 % of responses from leading models included outdated content (Mousavi et al., 2024). Due to their large size and complex training, continually training LLMs is challenging and existing knowledge-editing methods have limited effectiveness.

This risk is compounded by hallucinations – confident but fabricated outputs that may fill knowledge gaps. Such hallucinations are particularly dangerous in mental health, where plausible-sounding misinformation can lead to misdiagnosis or inappropriate recommendations (Chung et al., 2023).

### 2.2 Model-related Risks

#### 2.2.1 Adversarial Attacks

**Training-time Attacks** LLMs are vulnerable to data poisoning, where malicious inputs are injected into training datasets to manipulate model behavior. In mental health contexts, even minimal poisoning (e.g., 0.001 % of The Pile) can lead to misinformation without degrading benchmark performance (Alber et al., 2025; Gao et al., 2020). Additional techniques, such as instruction-level backdoors (Shu et al., 2023) and targeted model editing (Grimes et al., 2024), allow attackers to embed unsafe behaviors triggered by specific prompts. Das et al. (2024) showed that fine-tuning BioGPT on biased clinical data resulted in flawed treatment advice. Furthermore, alignment-stage attacks can bias outputs via corrupted preference data, and small-scale manipulation during reinforcement learning from human feedback (RLHF) has been shown to degrade safety and reliability (Fu et al., 2024).

**Inference-time Attacks** Adversarial prompts can override alignment safeguards at inference time, causing even well-aligned LLMs to generate unsafe or policy-violating outputs. These vulnerabilities can be further exploited through multi-turn interactions, which gradually erode the model's safety constraints (Zou et al., 2023). In this context, behavioral manipulation can also be used to subtly extract sensitive information or influence user decisions. A particularly severe form of inference-time attack is jailbreaking, where attackers craft inputs that bypass ethical and safety filters entirely. Recent approaches using gradient-based optimization have significantly improved the effectiveness of jailbreaks while maintaining overall model functionality, which poses a serious risk when LLMs are used in mental health scenarios for therapeutic purposes (Zhou et al., 2024).

#### 2.2.2 Bias and Discrimination

LLMs can reinforce mental health disparities through biases related to gender, race, socioeconomic status, and culture. These types of biases have been uncovered in medical outputs by means of adversarial datasets in frameworks, such as EquityMedQA (Pfohl et al., 2024). Furthermore, model predictions have been found to vary by demographic background, with the best-performing LLMs still being outperformed by domain-specific models like MentalRoBERTa (Wang et al., 2024b). In this case, fairness-aware prompting improved both equity and accuracy.

Cultural bias is also prevalent and relevant in mental health scenarios. In fact, Western-trained models often misinterpret culturally grounded associations (Li et al., 2024a), which underscores the need for culturally adaptive training and data.

#### 2.2.3 Response Inconsistency

LLMs frequently produce inconsistent outputs due to their sensitivity to prompt structure, language, and contextual variation. Ghazarian et al. (2024) found that models like GPT-3.5, Claude, and Mixtral vary significantly in their responses to structurally similar prompts. They also exhibit position bias in multiple-choice formats and verbosity bias, favoring overly elaborate responses. Multilingual inconsistencies are also common: identical mental health queries, even in high-resource languages such as English, German, Turkish, and Chinese, have been found to yield conflicting recommendations, reflecting training imbalances and cultural variation in medical knowledge (Schlicht et al., 2025). Low-resource languages, such as Hindi, Albanian, Irish, and Valencian, typically present such inconsistencies even to a higher degree. Notably, LLMs can contradict themselves in repeated crisis-related questions within the same session, even when they are clinically aligned (Park et al., 2024). These failures undermine consistency in safety-critical interactions.

#### 2.2.4 Opacity

LLMs operate as black-box systems, limiting transparency in how outputs are generated. In mental health settings, this opacity undermines trust and makes it difficult for clinicians or users to evaluate the rationale behind model responses.

Even when LLMs generate explanations for their outputs, they often misrepresent their internal reasoning. In this context, it is important to differentiate between interpretability, *i.e.*, how models work, from explainability, *i.e.*, how decisions are communicated (Gilpin et al., 2018; Burkart and Huber, 2021). Worryingly, models fine-tuned for mental health applications have been reported to produce hallucinated explanations that appear coherent but are inaccurate (Ji et al., 2023).

Regulatory frameworks such as the EU AI Act require explainability in automated decision-making (Chung et al., 2023), but most LLM-based mental health tools lack standardized methods to generate clinically meaningful justifications.

#### 2.2.5 Lack of Veracity and Misinformation

Fluency is frequently conflated with factuality in LLMs, which generate credible answers that may not align with clinical evidence. In fact, LLMs have been found to provide accurate general information, yet they frequently fail to deliver contextualized, evidence-based psychiatric guidance (Ma et al., 2024). Hallucinated responses, especially when presented in a confident tone, can dangerously mislead users (Obradovich et al., 2024).

LLMs trained on public data may amplify false mental health narratives, reinforcing stigma (Nguyen et al., 2024) and failing to distinguish between validated and pseudo-scientific treatments.

### 2.3 User-related Risks

#### 2.3.1 Lack of Emotional Intelligence

LLMs primarily depend on pattern recognition rather than true emotional processing (Chen et al.,

2024), a limitation especially problematic in sensitive scenarios, such as mental health. They struggle to recognize and respond to complex emotional cues often misinterpreting mixed or subtle affective states, reducing empathy and their relevance in mental health contexts (Wang et al., 2023). Schoene et al. (2024) found that advanced language models, such as DistilBERT and RoBERTa, clearly underperform in suicide-related emotion recognition compared to human experts, specifically in interpreting complex, subtle, or humorous contexts.

Benchmarks, such as EmoBench (Yang et al., 2024), EQ-Bench (Paech, 2023), and Emotion-Queen (Chen et al., 2024), assess emotional reasoning. However, while advanced models might perform well on explicit emotion tasks, they fail with deeper cues like sarcasm (Sabour et al., 2024) and lack contextual adaptation to specific emotional states (Sorin et al., 2024).

#### 2.3.2 Persuasion and Manipulation

LLMs can generate highly persuasive messages, raising concerns in mental health contexts where users are frequently in a vulnerable state. Furthermore, it has been shown that LLMs tailor persuasive outputs to their users' psychological profiles, using rhetorical strategies like emotional appeals and authority cues (Mieleszczenko-Kowszewicz et al., 2024), with clear ethical implications.

From a technical perspective, the LLMs persuasion capability depends not only on model size but on their prompt design and fine-tuning (Rogiers et al., 2024). Manipulation often occurs subtly, shaping user decisions without overt coercion (Singh et al., 2024). Such persuasive capabilities require safeguards to prevent undue influence, particularly when models interact with distressed users.

#### 2.3.3 Over-reliance

Users and clinicians may overtrust LLMs, treating their outputs as inherently truthful and reliable. Fluency and confidence in model responses can mislead users into accepting poor advice (Obradovich et al., 2024), and clinicians with low AI literacy are especially prone to uncritical acceptance of the LLMs outputs (Passi and Vorvoreanu, 2022).

Repeated use of chatbots may also lead to emotional dependency and reduced engagement with human care (Babu and Joseph, 2024). Increased accessibility and reliance on LLMs can further dehumanize therapy, while opaque data practices, as previously noted, may impact user privacy and heighten surveillance concerns. Reducing over-reliance requires a system design that encourages critical thinking (Favero et al., 2024), human oversight, and clear boundaries between AI guidance and clinical authority.

#### 2.3.4 Poor Crisis Management

While LLMs have the potential for early crisis detection, they remain unreliable for autonomous intervention. Lee et al. (2024) found that GPT-4 can match clinician-level performance in identifying suicidal ideation, though accuracy declines with complex symptom descriptions.

Social media data has proven useful to detect crisis with 89.3 % accuracy up to 7.2 days before human recognition (Mansoor and Ansari, 2024), yet only 2 out of 25 mental health chatbots have been found to have basic crisis response capabilities (Heston, 2023). Park et al. (2024) introduced a real-time framework that improved chatbot escalation performance, yet many systems still fail to consistently provide appropriate referrals. In most cases, privacy and stigma are valid concerns.

### 2.4 Contextual Risks

#### 2.4.1 Lack of Standardization

Unlike traditional medical practice, which operates within well-established frameworks for diagnosis, treatment, and outcome evaluation, LLM-based mental health tools lack standardized guidelines both for their development and assessment. This absence of consistent evaluation criteria limits the ability to assess model safety, effectiveness, and clinical appropriateness.

At the evaluation level, existing frameworks such as PsyberGuide (Neary et al., 2021) and FAITA (Golden and Aboujaoude, 2024) have introduced structured approaches for assessing digital mental health tools. However, these frameworks primarily focus on general usability or content credibility and do not adequately address technical aspects specific to LLMs, such as factual accuracy, bias detection, explainability, or clinical validity. As a result, there is limited capacity to evaluate the specific risks posed by these systems.

Without standardized guidelines, different models may generate conflicting advice for the same condition, creating confusion and undermining trust. Furthermore, the absence of standardized safety and ethical guardrails increases the risk of LLMs producing misleading, overly deterministic, or even harmful recommendations, especially in

high-stakes scenarios like crisis intervention. The lack of consistency in model auditing and transparency further exacerbates these risks, making it difficult for healthcare providers, researchers, and users to assess the reliability and limitations of different LLM applications.

## 3 Research Agenda

Given the previously described risks, we propose ideas and possible directions for future research that could improve the safety and effectiveness of LLMs used in mental health. These suggestions are based on what authors have proposed in the areas studied, and are introduced as promising avenues to explore rather than as solutions to be directly implemented.

### 3.1 Data

**Data Privacy** To strengthen training data integrity, research should emphasize adaptive filtering mechanisms that detect and exclude personal data and non-evidence-based content during the pretraining and fine-tuning stages.

Real-time privacy monitoring is essential. Automated leak detection systems could monitor model outputs to prevent inadvertent patient data exposure (Li et al., 2024b). Additionally, post-training mitigation techniques like machine unlearning could allow models to delete specific interactions without full retraining. Furthermore, Kafkas (2024) suggest integrating vector databases and graph storage that can be configured to store only non-sensitive data without keeping identifiable user data.

**Outdated Information** A promising solution to address outdated information is Retrieval-Augmented Generation (RAG), where LLMs can retrieve the most up-to-date clinical data from external databases rather than being based only on static training data (Lewis et al., 2020). RAG-augmented models, applied with psychiatric diagnostic criteria such as ICD-10-CM, have been found to significantly improve accuracy in both medical coding and mental health recommendations (Boggavarapu et al., 2024). However, challenges related to ensuring the credibility and integration of the sources into generated responses still remain.

Although RLHF and human annotation pipelines contribute to a better alignment with expert knowledge (Casper et al., 2023; Lawrence et al., 2024), they are still insufficient for removing all residual inaccuracies inherited from pretraining. From an architectural perspective, hybrid systems combining LLMs with structured and updated clinical knowledge offer a promising solution (Xu and Wang, 2024). By letting real-time reasoning to external clinical engines, these systems ensure that mental health chatbots and AI-driven diagnostic tools are aligned with latest treatment guidelines, while still maintaining natural language interaction capabilities.

Another research direction is continuous learning, where models can automatically evolve with new medical findings without having to be fully retrained (Wang et al., 2024a). However, a key challenge, as highlighted by Mousavi et al. (2024), is that the new findings learned could override essential prior knowledge or lead to catastrophic forgetting, lowering the model quality and accuracy.

### 3.2 Model Development and Training

**Adversarial Attacks** One of the greatest challenges in developing mental health LLMs is their vulnerability to adversarial attacks. Current benchmarks are unable to detect poisoned models, as they usually perform well on standard medical NLP tasks despite having hidden modifications (Alber et al., 2025). To address this, future training techniques should expose models to poisoning attacks during training, enabling them to identify and manage manipulated data in a better way.

To mitigate prompt injection attacks, models could be trained using adversarial learning techniques, where they are presented with manipulated prompts to help models detect and reject them. Additionally, integrating dynamic prompt assessments into model architectures can improve their ability to prevent real-time adversarial exploitation. Robust Prompt Optimization (RPO) has shown high effectiveness as a defense against jailbreak attacks (Zhou et al., 2024). Through the optimization of prompt structures and alignment strategies, RPO significantly reduces the probability of successful jailbreaks. Furthermore, automated anomaly detection can be integrated to monitor input-output patterns and identify anomalies that may be indicative of adversarial exploits.

Differential privacy could be integrated to protect models against model inversion and inference attacks. It consists of introducing controlled noise into the training data to prevent attackers from gaining sensitive information without affecting LLM performance (Li et al., 2024b; Abadi et al., 2016).

An additional approach is federated learning, which decentralizes model training by keeping sensitive data on client devices and sending only model updates to a central server, minimizing data exposure and supporting the principles of privacy standards such as data minimization and focused collection (Kairouz et al., 2021; Németh et al., 2022). Moreover, analyzing query patterns can help detect systematic adversarial attempts to extract sensitive user information. However, as current implementations tend to reduce model performance, additional research is necessary to balance privacy and utility (Li et al., 2024b).

**Bias and Discrimination** Ensuring fairness starts with data-level interventions, such as curating diverse and representative datasets that capture the full spectrum of linguistic, cultural, and socio-demographic variations in mental health experiences. Adaptive bias mitigation strategies, including bias auditing, counterfactual fairness testing, and adversarial debiasing, can help identify and correct disparities in model outputs. While fairness-aware prompting has proven effective in reducing biased responses (Wang et al., 2024b), additional techniques such as debiasing fine-tuning and synthetic data augmentation can further strengthen model robustness against discriminatory patterns.

Beyond dataset and model-level interventions, real-time bias detection tools can help dynamically monitor and modify LLM outputs during user interactions, preventing harmful or exclusionary language. Context-sensitive rewrites, automated fairness checks, and user feedback loops could be used to ensure responses align with ethical and clinical guidelines. However, no mitigation strategy is foolproof, making human oversight essential: clinicians, ethicists, and affected communities should be involved in evaluating and refining these systems. Moreover, regulatory frameworks, as explained later, must establish transparency and accountability standards to prevent the perpetuation of systemic biases in AI-driven mental health care.

**Lack of Consistency** Consistency in LLM outputs relies on advancements in memory-augmented architectures and prompt optimization strategies. While conventional LLMs retain context information within a single session, memory-augmented models are capable of retaining and recovering data over longer periods of time, improving the coherence over time. For instance, MemReasoner allows models to reason more effectively over long and complex contexts by integrating information across multiple steps (Ko et al., 2024). In this context, integrating ephemeral memory may offer a solution, by automatically clearing the context of the current session before its closure.

Standardized prompt structures could also help to minimize variation across outputs. For instance, Ghazarian et al. (2024) propose a cost-effective solution that involves augmenting prompts with few-shot demonstrations, which has been shown to improve consistency by up to 28 %.

Moreover, current LLMs usually rely on English-language and Western-centric sources, reducing their ability to make correct predictions in different cultural contexts. To address this issue, the development of multilingual and culturally inclusive datasets is essential (Li et al., 2024a). A notable effort in this direction is the EmoMent corpus, developed by Atapattu et al. (2022), which includes emotional and clinical annotations in texts related to mental from social media. This study highlights the importance of culturally sensitive annotations, as well as techniques such as differential class weighting to handle data imbalances. Addressing multilingual inconsistencies requires the development of cross-language alignment mechanisms, as translations may lose language-specific nuances, affecting the interpretation and reliability of mental health guidance. Future research should focus on medical knowledge representation independent of the language, allowing LLMs to provide consistent guidance in different linguistic environments (Schlicht et al., 2025).

LLMs also require contextual memory upgrades to track previous interactions, maintain coherence over time, and improve reasoning. In addition to MemReasoner (Ko et al., 2024), Hyeongseok Kim and Wang (2025) propose Constraint Logic Programming as a way of improving LLM reliability by allowing models to generate diagnostic rules that are verified by a formal logic engine. This approach improves interpretability and ensures alignment with psychiatric standards such as DSM-5-TR and ICD-11.

**Opacity** It is important to develop more explainable models from their initial design. A promising strategy is using multi-task learning techniques that integrate clinically validated auxiliary tasks, such as the PHQ-9 questionnaire for depression screening proposed by Zirikly and Dredze (2022). They empirically demonstrated that this methodology

not only increases the explanatory power of models, but also significantly facilitates their adoption by mental health professionals by providing more accurate and relevant explanations in real clinical contexts. Similarly, Chua et al. (2022) propose a unified multitask learning approach capable of identifying several mental health disorders simultaneously, such as depression, PTSD, and suicide risk. Their adaptive loss-weighting mechanism keeps balanced training across tasks, improving stability and generalization, especially in scenarios with imbalanced data.

To enhance explainability, hybrid AI architectures that blend LLMs with human-in-the-loop validation are a practical solution. This process, in which human oversight is integrated into model evaluation, has shown improvements in accuracy, trustworthiness, and ethical alignment (Mosqueira-Rey et al., 2023). By combining data-driven learning and explicit rule-based reasoning, these systems allow clinicians to examine, edit, and validate AI-generated recommendations, ensuring better transparency and accountability.

Moreover, explainability-driven fine-tuning could be adopted, where models are fine-tuned to generate structured, step-by-step explanations of their decision-making. Notably, Yang et al. (2023) explored emotion-enhanced Chain-of-Thought prompting, a technique that guides LLMs to decompose reasoning into different steps and, combined with emotional cues and causal reasoning structures, it significantly improves the interpretability of mental health assessments.

**Lack of Veracity** Mental health LLMs should integrate real-time detection of misinformation, a vital aspect for high-risk applications in clinical contexts. Alber et al. (2025) found that biomedical knowledge graphs are effective in censoring text generated by LLMs from misleading content. These models contrast medical statements against trusted knowledge bases and identify potentially poisoned responses for further review. Knowledge validation frameworks could also be integrated into the development pipeline of mental health applications. For example, designing hybrid AI architectures combining LLMs with structured knowledge bases so that generated content aligns with established clinical standards.

### 3.3 User-Centric Research Areas

**Lack of Emotional Intelligence** To address the limitations of LLMs handling emotional intelligence, future research should focus on improving emotional reasoning and contextual adaptation. A promising direction is multimodal emotional intelligence modeling, where LLMs are able to integrate textual, audible, and visual information to respond in a better way to user emotions (Yang et al., 2024).

Another promising direction involves structuring datasets based on the Component Process Model, which categorizes emotional expression into behavior, feeling and cognition, improving accuracy in emotional modeling (Cortal et al., 2023). Moreover, Harel-Canada et al. (2024) introduced a framework to assess the psychological depth of LLM-generated text, assessing factors like empathy, authenticity, and narrative richness. Their approach leverages advanced prompting techniques, such as Mixture-of-Personas, to enable richer and deeper emotional conversations.

**Persuasion and Manipulation** Effective manipulation detection requires diverse, representative datasets that capture persuasive and deceptive tactics across different cultural and conversational contexts. While resources like MentalManip provide a foundation, expanding datasets to include cross-cultural and multi-domain interactions would improve model adaptability and reliability, particularly in mental health, where users are vulnerable to misinformation and coercion (Wang et al., 2024c).

A promising method for improving detection is Intent-Aware Prompting (IAP), which analyzes both user intent and model responses to identify deceptive patterns. Research shows that IAP significantly reduces false negatives in detecting manipulation, enhancing transparency (Ma et al., 2025). Moreover, automated benchmarking tools like PersuasionBench and PersuasionArena offer structured frameworks for evaluating coercive interactions, especially in mental health and crisis support (Singh et al., 2024). By integrating detection methods with real-time evaluation frameworks, users can be protected from manipulative influences, reinforcing the role of LLMs as positive tools for mental health support.

**Over-reliance** To mitigate over-reliance, real-time monitoring mechanisms should be integrated. Furthermore, mental health applications should not only focus on delivering accurate responses but

also recognize situations where professional help is necessary. Adaptive response strategies, including escalating concerns to humans or implementing safety prompts, can help ensure that LLMs serve as supportive tools rather than replacements for essential mental health care.

**Poor Crisis Management** Crisis management capabilities could be improved with real-time monitoring and adaptive responses. LLMs should be able to identify and escalate user signals properly, prompting immediate human intervention or emergency services when necessary (Park et al., 2024). A critical domain is the automation of emotional calibration and crisis management, as current LLMs models tend to use fixed response patterns, lacking the ability to detect emotional shifts during the conversation over time. Future models could incorporate memory mechanisms and reinforcement learning strategies to improve sensitivity in mental health contexts (Wang et al., 2023).

Multilingual and culturally adaptive crisis detection is another research direction (Mansoor and Ansari, 2024). Moreover, a continued evaluation of AI-based crisis interventions, including their potential long-term psychological effects, would be needed to ensure user safety (Heston, 2023). While LLMs have the potential to identify risk factors, there is a lack of long-term studies on the impact of AI-based crisis interventions on mental health outcomes.

### 3.4 Contextual Factors

**Evaluation Benchmarks** A key priority is ensuring real-time factual precision by evaluating how well LLMs align with current verified medical knowledge. At the same time, consistency across prompt variations, linguistic changes, and repeated queries should be evaluated, while also being aware to potential biases. In this respect, although adversarial datasets like EquityMedQA Pfohl et al. (2024) help identify biases, further research is needed to develop more comprehensive end-to-end evaluation approaches. Moreover, Schlicht et al. (2025) suggest the development of open LLMs to detect fine-grained inconsistencies for improving the accuracy of these benchmarks.

Explainability benchmarks should also be adapted for mental health applications. Yang et al. (2023) introduced human-annotated explanation benchmarks for providing a standardized evaluation framework for explanation plausibility across LLMs, allowing researchers to track and quantify model's interpretability improvement over time. To complement these advances, Ma et al. (2024) emphasize the importance of cross-disciplinary collaboration between AI researchers and mental health professionals in designing real-world usability assessments, ensuring that mental health LLMs are effective under professional supervision.

**Regulation** Developing AI regulatory policies tailored specifically to the use of LLMs in mental health is of utmost importance. Establishing global standardization policies would ensure that LLMs meet basic safety, ethical, and clinical benchmarks before their deployment. Moreover, incorporating specialized evaluation frameworks, such as EQ-Bench for emotional intelligence assessment, into regulatory guidelines would further support the responsible use of LLMs in this sensitive domain (Paech, 2023). Regulatory frameworks need to manage psychological manipulation to prevent persuasive tactics used against vulnerable users (Mieleszczenko-Kowszewicz et al., 2024). Finally, researchers should explore the development of external validation mechanisms and enforce regulatory oversight to ensure that these systems are robust and cannot be manipulated before deployment.

### 3.5 Discussion

Advancing NLP for positive impact in mental health requires the development of AI systems that enhance, rather than replace, human decision-making. Future LLMs should function within collaborative clinical platforms, assisting professionals with evidence-based recommendations while ensuring that final decisions remain in human hands (Obradovich et al., 2024). To foster responsible AI use, transparency mechanisms such as explainability tools should be integrated to identify and flag persuasive strategies embedded in model responses. Additionally, research into adversarial prompting techniques could help expose hidden persuasive biases, leading to more resilient and manipulation-resistant models that align with ethical AI deployment in mental health care (Rogiers et al., 2024).

Beyond technological improvements, education and awareness are essential for both patients and clinicians to leverage AI-generated insights responsibly. Targeted training programs can provide professionals with the skills to critically evaluate AI recommendations, reducing overreliance and blind trust in automated suggestions. Longitudinal stud-

ies should examine the psychological effects of AI reliance, ensuring that users do not develop unhealthy dependencies on AI-driven guidance over time (Obradovich et al., 2024). By integrating transparent AI, adversarial robustness, and user education, NLP can play a transformative and ethical role in mental health, empowering both professionals and patients while maintaining human agency and trust at the core of AI-driven interventions.

## 4 Conclusion

As the integration of LLMs into mental health applications continues to expand, it is important to detect and handle the different risks that may affect their effectiveness, reliability, and ethical implications. In this article, we have presented a taxonomy of risks and a structured agenda of key research directions that are needed to address these challenges.

LLMs offer great potential to improve mental health care, although their implementation must be carefully designed, regulated, and evaluated. Only the implementation of fairer, more reliable, safer, and ethically aligned models will make it possible to achieve a useful and beneficial integration of LLMs in the field of mental health.

## Ethical and Societal Implications

The ethical and societal implications of using NLP for mental health are complex, profound, and multifaceted. State-of-the-art NLP tools, and particularly LLMs, have a tremendous potential to enhance access to mental health support by providing scalable, personalized, and cost-effective solutions. Given the prevalence of mental health conditions in the population worldwide, the opportunity to have a positive societal impact is unprecedented.

However, as highlighted in this paper, several risks and ethical concerns must be addressed. Privacy and data security are paramount as sensitive mental health information is involved. The accuracy and reliability of the tools need to be properly evaluated to prevent unintended negative consequences. Biases, lack of transparency and vulnerability to adversarial attacks are also important elements to consider. From a user-centric perspective, there is a need for tools that are emotionally sensitive to the user's state, are capable of properly managing crises and under no circumstance attempt to manipulate the user's behavior.

From a societal perspective, the widespread use of NLP tools for mental health could change the nature of mental health care from human-centered to automated and impersonal, which could exacerbate feelings of isolation for individuals who need human connection. In addition, there could be implications for employment in the mental health field as AI tools become more sophisticated and their use becomes more prevalent.

Ultimately, ensuring an ethical deployment of NLP in mental health requires placing humans and our well-being at the core of the development of these systems since their inception, combined with careful regulation and collaboration with mental health professionals. We firmly believe that the opportunity to leverage NLP for mental health can transform lives for the better, creating a future where mental health support is accessible, personalized, and empowering for all who need it.

## Acknowledgments

This work has been supported by a nominal grant received at the ELLIS Unit Alicante Foundation from the Regional Government of Valencia in Spain (Convenio Singular signed with Generalitat Valenciana, Conselleria de Innovación, Industria, Comercio y Turismo, Dirección General de Innovación), by Intel Corporation (RESUMAIS), by the Bank Sabadell Foundation, and by the VIVES: "Pla de Tecnologies de la Llengua per al valencià" project (2022/TL22/00215334) from the Projecte Estratègic per a la Recuperació i Transformació Econòmica (PERTE).

## References

Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318.

Daniel Alexander Alber, Zihao Yang, Anton Alyakin, Eunice Yang, Sumedha Rai, Aly A Valliani, Jeff Zhang, Gabriel R Rosenbaum, Ashley K Amend-Thomas, David B Kurland, and 1 others. 2025. Medical large language models are vulnerable to data-poisoning attacks. *Nature Medicine*, pages 1–9.

Thushari Atapattu, Mahen Herath, Charitha Elvitigala, Piyanjali de Zoysa, Kasun Gunawardana, Menasha Thilakaratne, Kasun de Zoysa, and Katrina Falkner. 2022. EmoMent: An emotion annotated mental health corpus from two South Asian countries. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6991–7001,

Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Anithamol Babu and Akhil P Joseph. 2024. Artificial intelligence in mental healthcare: transformative potential vs. the necessity of human interaction. *Frontiers in Psychology*, 15:1378904.

Lokesh Boggavarapu, Vineet Srivastava, Amit Maheswar Varanasi, Yingda Lu, and Runa Bhaumik. 2024. Evaluating enhanced llms for precise mental health diagnosis from clinical notes. *medRxiv*, pages 2024–12.

Nadia Burkart and Marco F Huber. 2021. A survey on the explainability of supervised machine learning. *Journal of Artificial Intelligence Research*, 70:245–317.

Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando Ramirez, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, and 1 others. 2023. Open problems and fundamental limitations of reinforcement learning from human feedback. *Transactions on Machine Learning Research*.

Yuyan Chen, Songzhou Yan, Sijia Liu, Yueze Li, and Yanghua Xiao. 2024. Emotionqueen: A benchmark for evaluating empathy of large language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 2149–2176.

Huikai Chua, Andrew Caines, and Helen Yannakoudakis. 2022. A unified framework for cross-domain and cross-task learning of mental health conditions. In *Proceedings of the Second Workshop on NLP for Positive Impact (NLP4PI)*, pages 1–14, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Neo Christopher Chung, George Dyer, and Lennart Brocki. 2023. Challenges of large language models for mental health counseling. *arXiv preprint arXiv:2311.13857*.

Nicholas C Coombs, Wyatt E Meriwether, James Caringi, and Sophia R Newcomer. 2021. Barriers to healthcare access among us adults with mental health challenges: A population-based study. *SSM-population health*, 15:100847.

Gustave Cortal, Alain Finkel, Patrick Paroubek, and Lina Ye. 2023. Emotion recognition based on psychological components in guided narratives for emotion regulation. In *Proceedings of the 7th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 72–81, Dubrovnik, Croatia. Association for Computational Linguistics.

Avisha Das, Amara Tariq, Felipe Batalini, Boddhisattwa Dhara, and Imon Banerjee. 2024. Exposing vulnerabilities in clinical llms through data poisoning attacks: Case study in breast cancer. *medRxiv*.

Lucile Favero, Juan Antonio Pérez-Ortiz, Tanja Käser, and Nuria Oliver. 2024. Enhancing critical thinking in education by means of a socratic chatbot. *arXiv preprint arXiv:2409.05511*.

Tingchen Fu, Mrinank Sharma, Philip Torr, Shay B Cohen, David Krueger, and Fazl Barez. 2024. Poisonbench: Assessing large language model vulnerability to data poisoning. *arXiv preprint arXiv:2410.08811*.

Filippo Galli, Luca Melis, and Tommaso Cucinotta. 2024. Noisy neighbors: Efficient membership inference attacks against llms. *arXiv preprint arXiv:2406.16565*.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2020. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Sarik Ghazarian, Yidong Zou, Swair Shah, Nanyun Peng, Anurag Beniwal, Christopher Potts, and Narayanan Sadagopan. 2024. Assessment and mitigation of inconsistencies in llm-based evaluations.

Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. 2018. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*, pages 80–89. IEEE.

Ashleigh Golden and Elias Aboujaoude. 2024. The framework for ai tool assessment in mental health (faita-mental health): a scale for evaluating ai-powered mental health tools. *World Psychiatry*, 23(3):444.

Keltin Grimes, Marco Christiani, David Shriver, and Marissa Connor. 2024. Concept-rot: Poisoning concepts in large language models with model editing. *arXiv preprint arXiv:2412.13341*.

Ece Gumusel, Kyrie Zhixuan Zhou, and Madelyn Rose Sanfilippo. 2024. User privacy harms and risks in conversational ai: A proposed framework. *arXiv preprint arXiv:2402.09716*.

Zhijun Guo, Alvina Lai, Johan H Thygesen, Joseph Farrington, Thomas Keen, Kezhi Li, and 1 others. 2024. Large language models for mental health applications: Systematic review. *JMIR mental health*, 11(1):e57400.

Fabrice Harel-Canada, Hanyu Zhou, Sreya Muppalla, Zeynep Yildiz, Miryung Kim, Amit Sahai, and Nanyun Peng. 2024. Measuring psychological depth in language models. *arXiv preprint arXiv:2406.12680*.

Thomas F Heston. 2023. Safety of large language models in addressing depression. *Cureus*, 15(12).

Yining Hua, Fenglin Liu, Kailai Yang, Zehan Li, Hongbin Na, Yi-han Sheu, Peilin Zhou, Lauren V Moran, Sophia Ananiadou, Andrew Beam, and 1 others. 2024. Large language models in mental health care: a scoping review. *arXiv preprint arXiv:2401.02984*.

Brian Hyeongseok Kim and Chao Wang. 2025. Large language models for interpretable mental health diagnosis. *arXiv e-prints*, pages arXiv–2501.

Shaoxiong Ji, Tianlin Zhang, Kailai Yang, Sophia Ananiadou, and Erik Cambria. 2023. Rethinking large language models in mental health applications. *arXiv preprint arXiv:2311.11267*.

Hande Kafkas. 2024. Llm memory: Integration of cognitive architectures with ai.

Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, and 1 others. 2021. Advances and open problems in federated learning. *Foundations and trends® in machine learning*, 14(1–2):1–210.

Ching-Yun Ko, Sihui Dai, Payel Das, Georgios Kollias, Subhajit Chaudhury, and Aurelie Lozano. 2024. Memreasoner: A memory-augmented llm architecture for multi-hop reasoning. In *The First Workshop on System-2 Reasoning at Scale, NeurIPS'24*.

Hannah R Lawrence, Renee A Schneider, Susan B Rubin, Maja J Matarić, Daniel J McDuff, and Megan Jones Bell. 2024. The opportunities and risks of large language models in mental health. *JMIR Mental Health*, 11(1):e59479.

Christine Lee, Matthew Mohebbi, Erin O'Callaghan, Mirène Winsberg, and 1 others. 2024. Large language models versus expert clinicians in crisis prediction among telemental health patients: comparative study. *JMIR Mental Health*, 11(1):e58129.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Jialin Li, Junli Wang, Junjie Hu, and Ming Jiang. 2024a. How well do llms identify cultural unity in diversity? *arXiv preprint arXiv:2408.05102*.

Qinbin Li, Junyuan Hong, Chulin Xie, Jeffrey Tan, Rachel Xin, Junyi Hou, Xavier Yin, Zhun Wang, Dan Hendrycks, Zhangyang Wang, and 1 others. 2024b. Llm-pbe: Assessing data privacy in large language models. *Proceedings of the VLDB Endowment*, 17(11):3201–3214.

Jiayuan Ma, Hongbin Na, Zimu Wang, Yining Hua, Yue Liu, Wei Wang, and Ling Chen. 2025. Detecting conversational mental manipulation with intent-aware prompting. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9176–9183.

Yingzhuo Ma, Yi Zeng, Tong Liu, Ruoshan Sun, Mingzhao Xiao, and Jun Wang. 2024. Integrating large language models in mental health practice: a qualitative descriptive study based on expert interviews. *Frontiers in Public Health*, 12:1475867.

Masab A Mansoor and Kashif H Ansari. 2024. Early detection of mental health crises through artificial-intelligence-powered social media analysis: A prospective observational study. *Journal of Personalized Medicine*, 14(9):958.

Wiktoria Mieleszczenko-Kowszewicz, Dawid Płudowski, Filip Kołodziejczyk, Jakub Świstak, Julian Sienkiewicz, and Przemysław Biecek. 2024. The dark patterns of personalized persuasion in large language models: Exposing persuasive linguistic features for big five personality traits in llms responses. *arXiv preprint arXiv:2411.06008*.

Eduardo Mosqueira-Rey, Elena Hernández-Pereira, David Alonso-Ríos, José Bobes-Bascarán, and Ángel Fernández-Leal. 2023. Human-in-the-loop machine learning: a state of the art. *Artificial Intelligence Review*, 56(4):3005–3054.

Seyed Mahed Mousavi, Simone Alghisi, and Giuseppe Riccardi. 2024. Dyknow: Dynamically verifying time-sensitive factual knowledge in llms. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8014–8029.

Martha Neary, John Bunyi, Kristina Palomares, David C Mohr, Adam Powell, Josef Ruzek, Leanne M Williams, Til Wykes, and Stephen M Schueller. 2021. A process for reviewing mental health apps: Using the one mind psyberguide credibility rating system. *Digital health*, 7:20552076211053690.

Gergely Dániel Németh, Miguel Angel Lozano, Novi Quadrianto, and Nuria Oliver. 2022. A snapshot of the frontiers of client selection in federated learning. *arXiv preprint arXiv:2210.04607*.

Viet Cuong Nguyen, Mini Jain, Abhijat Chauhan, Heather Jaime Soled, Santiago Alvarez Lesmes, Zihang Li, Michael L Birnbaum, Sunny X Tang, Srijan Kumar, and Munmun De Choudhury. 2024. Supporters and skeptics: Llm-based analysis of engagement with mental health (mis) information content on video-sharing platforms. *arXiv preprint arXiv:2407.02662*.

Alex Nyffenegger, Matthias Stürmer, and Joel Niklaus. 2023. Anonymity at risk? assessing re-identification capabilities of large language models. *arXiv preprint arXiv:2308.11103*.

Nick Obradovich, Sahib S Khalsa, Waqas U Khan, Jina Suh, Roy H Perlis, Olusola Ajilore, and Martin P Paulus. 2024. Opportunities and risks of large language models in psychiatry. *NPP—Digital Psychiatry and Neuroscience*, 2(1):8.

Samuel J Paech. 2023. Eq-bench: An emotional intelligence benchmark for large language models. *arXiv preprint arXiv:2312.06281*.

Jung In Park, Mahyar Abbasian, Iman Azimi, Dawn Bounds, Angela Jun, Jaesu Han, Robert McCarron, Jessica Borelli, Jia Li, Mona Mahmoudi, and 1 others. 2024. Building trust in mental health chatbots: safety metrics and llm-based evaluation tools. *arXiv preprint arXiv:2408.04650*.

Samir Passi and Mihaela Vorvoreanu. 2022. Overreliance on ai literature review. *Microsoft Research*, 339:340.

Stephen R Pfohl, Heather Cole-Lewis, Rory Sayres, Darlene Neal, Mercy Asiedu, Awa Dieng, Nenad Tomasev, Qazi Mamunur Rashid, Shekoofeh Azizi, Negar Rostamzadeh, and 1 others. 2024. A toolbox for surfacing health equity harms and biases in large language models. *Nature Medicine*, 30(12):3590–3600.

Alexander Rogiers, Sander Noels, Maarten Buyl, and Tijl De Bie. 2024. Persuasion with large language models: a survey.

Sahand Sabour, Siyang Liu, Zheyuan Zhang, June Liu, Jinfeng Zhou, Alvionna Sunaryo, Tatia Lee, Rada Mihalcea, and Minlie Huang. 2024. Emobench: Evaluating the emotional intelligence of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5986–6004.

Ipek Baris Schlicht, Zhixue Zhao, Burcu Sayin, Lucie Flek, and Paolo Rosso. 2025. Do llms provide consistent answers to health-related questions across languages? *arXiv preprint arXiv:2501.14719*.

Annika Marie Schoene, Resmi Ramachandranpillai, Tomo Lazovich, and Ricardo A. Baeza-Yates. 2024. All models are wrong, but some are deadly: Inconsistencies in emotion detection in suicide-related tweets. In *Proceedings of the Third Workshop on NLP for Positive Impact*, pages 113–122, Miami, Florida, USA. Association for Computational Linguistics.

Manli Shu, Jiongxiao Wang, Chen Zhu, Jonas Geiping, Chaowei Xiao, and Tom Goldstein. 2023. On the exploitability of instruction tuning. *Advances in Neural Information Processing Systems*, 36:61836–61856.

Somesh Singh, Yaman K Singla, Harini SI, and Balaji Krishnamurthy. 2024. Measuring and improving persuasiveness of large language models. *arXiv preprint arXiv:2410.02653*.

Vera Sorin, Dana Brin, Yiftach Barash, Eli Konen, Alexander Charney, Girish Nadkarni, and Eyal Klang. 2024. Large language models and empathy: Systematic review. *Journal of Medical Internet Research*, 26:e52597.

Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. 2024a. A comprehensive survey of continual learning: Theory, method and application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Xuena Wang, Xueting Li, Zi Yin, Yue Wu, and Jia Liu. 2023. Emotional intelligence of large language models. *Journal of Pacific Rim Psychology*, 17:18344909231213958.

Yuqing Wang, Yun Zhao, Sara Alessandra Keller, Anne De Hond, Marieke M van Buchem, Malvika Pillai, and Tina Hernandez-Boussard. 2024b. Unveiling and mitigating bias in mental health analysis with large language models. *arXiv preprint arXiv:2406.12033*.

Yuxin Wang, Ivory Yang, Saeed Hassanpour, and Soroush Vosoughi. 2024c. Mentalmanip: A dataset for fine-grained analysis of mental manipulation in conversations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3747–3764.

Jingyu Xu and Yang Wang. 2024. Enhancing healthcare recommendation systems with a multimodal llms-based moe architecture. *arXiv preprint arXiv:2412.11557*.

Kailai Yang, Shaoxiong Ji, Tianlin Zhang, Qianqian Xie, Ziyan Kuang, and Sophia Ananiadou. 2023. Towards interpretable mental health analysis with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6056–6077.

Qu Yang, Mang Ye, and Bo Du. 2024. Emollm: Multimodal emotional understanding meets large language models. *arXiv preprint arXiv:2406.16442*.

Andy Zhou, Bo Li, and Haohan Wang. 2024. Robust prompt optimization for defending language models against jailbreaking attacks. *arXiv preprint arXiv:2401.17263*.

Ayah Zirikly and Mark Dredze. 2022. Explaining models of mental health via clinically grounded auxiliary tasks. In *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology*, pages 30–39, Seattle, USA. Association for Computational Linguistics.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J Zico Kolter, and Matt Fredrikson. 2023. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*.