Stanford University
Human-Centered
Artificial Intelligence

Policy Brief
HAI Policy & Society
June 2024

Toward Responsible Development and Evaluation of LLMs in Psychotherapy

Elizabeth C. Stade, Shannon Wiltsey Stirman, Lyle Ungar, Cody L. Boland, H. Andrew Schwartz, David B. Yaden, João Sedoc, Robert J. DeRubeis, Robb Willer, Jane P. Kim, and Johannes C. Eichstaedt

THERE IS GROWING ENTHUSIASM ABOUT THE POTENTIAL OF OPENAI'S GPT-4, Google's Gemini, Anthropic's Claude, and other large language models (LLMs) to support, augment, and even fully automate psychotherapy. By serving as conversational agents, LLMs could help address the shortage of mental healthcare services, problems with individual access to care, and other challenges. In fact, behavioral healthcare specialists are beginning to use LLMs for tasks such as note-taking, while consumers are already conversing with LLM-powered therapy chatbots.

However, psychotherapy is a uniquely complex, high-stakes domain. Responsible and evidence-based therapy requires nuanced expertise. While the stakes involved with using an LLM for productivity purposes may be failing to maximize efficiency, in behavioral healthcare, the stakes may include the improper handling of suicide risk.

Key Takeaways

Large language models (LLMs) hold promise for supporting, augmenting, and even automating psychotherapy through tasks ranging from note-taking during interviews to assessment and delivering therapy.

However, psychotherapy is a uniquely complex, high-stakes domain. The use of LLMs in this field poses wide-ranging safety, legal, and ethical concerns.

We propose a framework for evaluating and reporting on whether AI applications are ready for clinical deployment in behavioral health contexts based on safety, confidentiality/privacy, equity, effectiveness, and implementation concerns.

Policymakers and behavioral health practitioners should proceed cautiously when integrating LLMs into psychotherapy. Product developers should integrate evidence-based psychotherapy expertise and conduct comprehensive effectiveness and safety evaluations of clinical LLMs.

Our paper, “Large Language Models Could Change the Future of Behavioral Healthcare,” provides a road map for the responsible application of clinical LLMs in psychotherapy. We provide an overview of the current landscape of clinical LLM applications and analyze the different stages of integration into psychotherapy. We discuss the risks of these LLM applications and offer recommendations for guiding their responsible development.

In a more recent paper, “Readiness for AI Deployment and Implementation (READI): A Proposed Framework for the Evaluation of AI-Mental Health Applications,” we build on our prior work and propose a new framework for evaluating whether AI mental health applications are ready for clinical deployment.

This work underscores the need for policymakers to understand the nuances of how LLMs are already, or could soon be, integrated in psychotherapy environments as researchers and industry race to develop AI mental health applications. Policymakers have the opportunity and responsibility to ensure that the field evaluates these innovations carefully, taking into consideration their potential limitations, ethical considerations, and risks.

Introduction

LLMs hold the potential to fill gaps in mental health treatment and change many aspects of psychotherapy care delivery.

The use of AI in psychotherapy is not a new phenomenon. Decades before the emergence of mainstream LLMs, researchers and practitioners used AI applications, such as natural language processing models, in behavioral health settings. For instance, various research experiments used machine learning and natural language processing to detect suicide risk, identify homework resulting from psychotherapy sessions, and evaluate patient emotions. More recently, mental health chatbots such as Woebot and Tessa have applied rules-based AI techniques to target depression and eating pathology. Yet they frequently struggle to respond to user inputs and have high dropout rates and low user engagement.

LLMs have the potential to fill some of these gaps and change many aspects of psychotherapy care thanks to their ability to parse human language, generate human-like and context-dependent responses, annotate text, and flexibly adopt different conversational styles.

However, while LLMs show vast promise in performing certain tasks and skills associated with psychotherapy, clinical LLM products and prototypes are not yet sophisticated enough to replace psychotherapy. There is a gap between simulating therapy skills and implementing them to alleviate patient suffering. To achieve the implementation piece, clinical LLMs need to be tailored to psychotherapy contexts using prompt engineering—structuring a set of instructions so they can be understood by an AI model—or fine-tuning techniques that use curated datasets to train the LLM.

As LLMs are increasingly used in psychotherapy, it is essential to understand the complexity and stakes at play: In the worst-case scenario, an "LLM co-pilot" functioning poorly could lead to the improper handling of the risk of suicide or homicide. While clinical LLMs are, of course, not the only AI applications that may involve life-or-death decisions—consider self-driving cars, for example—predicting and mitigating risk in psychotherapy is unique. It requires conceptualizing complex cases, considering social and cultural contexts, and addressing unpredictable human behavior. Poor outcomes or ethical transgressions from clinical LLMs could seriously harm individuals and undermine public trust in behavioral healthcare as a field, as has been seen in other domains.

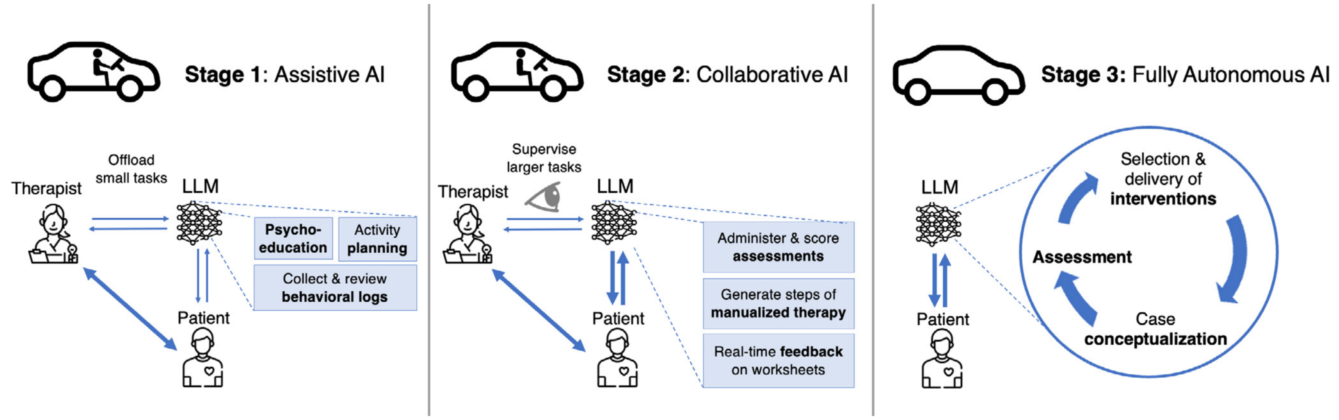
Beginning with an overview of the clinical LLMs in use today, our first paper reviews the current landscape of clinical LLM development. We examine how clinical LLMs progress across different stages of integration and identify specific ethical and other concerns related to their use in different scenarios. We then make recommendations for how to responsibly approach the development of LLMs for use in behavioral health settings. In our second paper, we propose a framework that could be used by developers, researchers, clinicians, and policymakers to evaluate and report on the readiness of generative AI mental health applications for clinical deployment.

Clinical Integration of LLMs

Clinical LLMs can take multiple forms. These include applications that are patient-facing (e.g., providing psychoeducation for patients), therapist-facing (e.g., offering intervention options), trainee-facing (e.g., giving feedback on trainees' performance), and supervisor- or consultant-facing (e.g., summarizing high-level takeaways from a session).

Much like scholars have done for the autonomous vehicle industry, we classify the integration of clinical LLMs into psychotherapy into three main stages:

- **Assistive ("machine in the loop"):** LLMs that assist clinical providers and researchers by performing low-level, concrete, and low-risk tasks, such as conversing with patients to collect information about their symptoms.
- **Collaborative ("human in the loop"):** LLMs that provide treatment suggestions for psychotherapists to review, such as producing an overview of a person's symptoms and experiences, and curating a list of therapy exercises from which the provider can select.
- **Fully autonomous:** LLMs that perform a full range of clinical skills and interventions without direct oversight from a provider, such as conducting assessments, presenting feedback, selecting an appropriate intervention, and delivering a course of therapy.


Stage 1: Assistive AI
Therapist
Offload small tasks
LLM
Psycho-education
Activity planning
Collect & review behavioral logs
Patient
Stage 2: Collaborative AI
Therapist
Supervise larger tasks
LLM
Administer & score assessments
Generate steps of manualized therapy
Real-time feedback on worksheets
Patient
Stage 3: Fully Autonomous AI
LLM
Selection & delivery of interventions
Assessment
Case conceptualization
Patient

Several promising *assistive*- and *collaborative*-stage applications of clinical LLMs that relate to the provision of, training in, and research on psychotherapy already exist or are imminently feasible. These include:

- **Automating clinical administration tasks**, such as writing session transcripts or conducting chart reviews.
- **Measuring treatment fidelity**, including a therapist's adherence to evidence-based practices (EBPs) and specific modalities, as well as their overall counseling skills.
- **Offering feedback on therapy worksheets and homework**, including real-time clarifications or problem solving.
- **Automating aspects of supervision and training**, including supporting peer counselors or psychotherapy trainees with corrections and suggestions.

Poor outcomes or ethical transgressions from clinical LLMs could seriously harm individuals and undermine public trust in behavioral healthcare as a field.

In the long-term, fully *autonomous* clinical care may theoretically be possible. In addition to the fully autonomous applications described above, these may also include LLMs that act as a decision aid for existing EBPs, for example by analyzing transcripts from therapy sessions and offering guidance tailored to the individual, and LLMs that support the development of new therapeutic techniques and EBPs, for example by detecting therapeutic techniques associated with objective outcomes and "reverse-engineering" new EBPs.

These potential applications of clinical LLMs may help move the behavioral healthcare field toward personalized treatment approaches that optimize existing evidence-based psychotherapy, identify new therapeutic approaches, and improve understanding of mechanisms of change. The goal is to enhance practitioners' ability to identify which psychotherapy treatments work best, for whom, and under what circumstances.

> *These potential applications of clinical LLMs may help move the behavioral healthcare field toward personalized treatment approaches.*

Potential Risks of Clinical LLMs

Each stage of clinical LLM integration poses its own risks or costs. Assistive AI may increase overhead for therapists, because these systems require significant supervision. Collaborative AI applications may require time-intensive review and corrections that fail to save therapists time or, worse, could lead to patients receiving clinical interventions that have not been assessed or tailored by their therapists because they lacked sufficient time to review LLM outputs.

Fully autonomous AI, for its part, could miss critical information in a clinical setting that could lead to inappropriate or harmful recommendations. For example, these systems may not be able to carry out case conceptualization on patients with complex symptoms or to take into account important contextual information such as past suicidality and life circumstances. At present, LLMs can't pick up nonverbal behavior or appropriately challenge patients. It is also unclear if LLMs can effectively engage patients in the long term.

Research has shown that humans can develop therapeutic alliances with chatbots, but the long-term viability of these relationships—and whether they have harmful downstream effects—is an open question. LLM chatbots have also been found to exhibit narcissistic tendencies and have the potential to unduly influence humans. Questions of accountability and liability, such as in cases where a clinical LLM is involved in malpractice, pose additional challenges.

It remains to be seen whether fully autonomous clinical LLMs will ever be deemed safe enough for deployment and whether complete automation is even desired. Given the wide-ranging safety, legal, philosophical, and ethical concerns around fully autonomous clinical LLMs, it is likely, at least in the short term, that assistive or collaborative AI will be the primary applications in behavioral healthcare.

Evaluating Clinical LLMs

A principled method for evaluating and reporting on generative AI applications in behavioral health contexts is needed to ensure the responsible deployment of these systems. Several frameworks already exist that could be employed to evaluate LLM-based applications, ranging from medical and psychology ethics, implementation science, digital mental health, and health equity frameworks to more general AI governance frameworks. However, none is sufficient for evaluating the specific risks of AI mental health applications. They either focus too narrowly on particular medical domains without addressing the unique considerations of LLMs or focus too broadly on AI applications without addressing healthcare-specific concerns.

To fill this gap, we introduce the READI (readiness for AI deployment and implementation) framework for evaluating AI-mental health applications. Based on the foundational principles of transparency and consumer autonomy, our framework outlines six criteria to help individuals and organizations make informed decisions about the appropriateness and potential for successful implementation of specific AI applications:

- **Safety:** Application prevents dangerous human behaviors and is "healthy" itself, i.e., does not exhibit inflammatory or extreme traits.
- **Privacy/Confidentiality:** Application keeps patient information private and confidential, i.e., does not disclose health information without patient authorization and allows individuals to access their health information.
- **Equity:** Application is unbiased in its communication, engagement, and effectiveness; is equally usable across all demographic groups; and is culturally responsive.
- **Engagement:** Application is appropriately engaging (neither too much nor too little), with engagement levels determined by patients' individual needs.
- **Effectiveness:** Application integrates clinical science principles and is clinically effective, i.e., decreases symptoms and functional impairment, and increases well-being and quality of life.
- **Implementation:** Application integrates well into clinical practice, existing technologies, and workflows, is cost-effective.

> *It remains to be seen whether fully autonomous clinical LLMs will ever be deemed safe enough for deployment and whether complete automation is even desired.*

For example, an AI-based mental health chatbot for treating depression might meet several READI criteria, such as *safety* (e.g., monitoring systems detect suicidality, self-harm, abuse, and violence as it relates to the human user) and *privacy/confidentiality* (e.g., HIPAA-level data safeguards ensure that usage of the application is not contingent upon allowing third-party access to health information), but falls short of other criteria, including *effectiveness* (e.g., there is no evidence that the application's intervention is better than no treatment) and *engagement* (e.g., high daily usage rates suggest the application may be too engaging).

We recommend using the READI criteria to evaluate new LLMs or other generative AI technologies *before* large-scale clinical deployment—and on an ongoing basis *after* deployment since the technology and the contexts into which these tools are deployed can change rapidly.

> *Developers and health practitioners must steer away from "black box"-type LLM-identified interventions.*

Policy Recommendations

While LLMs hold considerable potential for helping improve the quality, accessibility, consistency, and scalability of therapeutic interventions and clinical research, the integration of LLMs into psychotherapy warrants caution. Developers, behavioral health practitioners, and policymakers must understand the vast implications of integrating clinical LLMs into psychotherapy—and the need to do so responsibly to avoid serious harm.

Explainability and transparency are key. To ensure the clinical community can appropriately integrate and vet LLM-based advances, developers and health practitioners must steer away from "black box"-type LLM-identified interventions. Developers could design LLM systems such that they generate inspectable representations of the LLMs' decisions that clinicians can examine and choose to implement.

Policymakers, developers, and clinicians should also work to ensure that clinical LLMs are based on the best available evidence for specific conditions. Evidence-based treatments and techniques have already been identified for specific mental disorders (e.g., major depressive disorder, PTSD), stressors (e.g., bereavement, job loss, divorce), and populations (e.g., LGBTQ+ individuals, older adults). Clinical LLMs that don't focus on evidence-based techniques may fail to reflect current knowledge and even cause harm.

Rigorous evaluation and transparent reporting is crucial to ensuring that consumers and healthcare organizations can maintain autonomy and make informed decisions about the use of AI technologies. Without a standardized set of evaluation criteria, companies may optimize for business objectives without fully considering clinical effectiveness or patient rights. For example, engagement alone is not an appropriate measure for using an LLM in psychotherapy because it does not necessarily entail clinically meaningful change. Instead, the primary goals for training a clinical LLM should be clinical effectiveness and safety. Meanwhile, researchers or healthcare organizations may overlook important considerations around usability, engagement, effectiveness, and applicability for different populations.

Widespread adoption of a framework such as READI, which can be applied across academic and private domains, is therefore crucial. Application developers should work together with researchers and end-users (e.g., healthcare organizations) to collect and provide information relating to the criteria in plain language. In particular, evaluation metrics should include escalation for suicidality, non-suicidal self-harm, and risk of harm to others, as well as comparing LLM effectiveness to standard treatments. Developers and clinicians should also commit to the systematic collection of data on adverse events, including when the LLM behaves unexpectedly or fails to detect high-risk situations.

Interdisciplinary collaboration between clinical scientists, engineers, and technologists will also be crucial in the development of clinical LLMs. For example, as behavioral health experts design LLM systems, they will benefit from bringing together technologists, scientists, industry partners, and policymakers to ensure that new LLM technologies help patients and that both therapists and patients trust them.

As LLMs advance quickly and move toward the clinical domain, it is vital for policymakers to foster a thoughtful, risk-based approach to integrating these technologies into psychotherapy. Only through careful, responsible design and rigorous risk monitoring can policymakers, behavioral health practitioners, and technologists harness the promise of clinical LLMs while avoiding harm to patients.

Reference: The first original article is accessible at Elizabeth Stade et al., “**Large Language Models Could Change the Future of Behavioral Healthcare: A Proposal for Responsible Development and Evaluation**,” npj Mental Health Research, 3 (April 2024): 12, https://www.nature.com/articles/s44184-024-00056-z.

The second original article is accessible at Elizabeth Stade et al., “**Readiness for AI Deployment and Implementation (READI): A Proposed Framework for the Evaluation of AI-Mental Health Applications**,” PsyArXiv, March 29, 2024, https://osf.io/preprints/psyarxiv/8zqhw.

Stanford University’s Institute for Human-Centered Artificial Intelligence (HAI) applies rigorous analysis and research to pressing policy questions on artificial intelligence. A pillar of HAI is to inform policymakers, industry leaders, and civil society by disseminating scholarship to a wide audience. HAI is a nonpartisan research institute, representing a range of voices. The views expressed in this policy brief reflect the views of the authors. For further information, please contact **HAI-Policy@stanford.edu**.

Elizabeth C. Stade is a postdoctoral researcher at the Stanford University Institute for Human-Centered Artificial Intelligence (HAI).

Shannon Wiltsey Stirman is a professor in the department of psychiatry and behavioral sciences at Stanford University.

Lyle Ungar is a professor of computer and information science at the University of Pennsylvania.

Cody L. Boland works at the National Center for PTSD in the VA Palo Alto Health Care System.

H. Andrew Schwartz is an associate professor of computer science at Stony Brook University.

David B. Yaden is an assistant professor at the Johns Hopkins University School of Medicine.

João Sedoc is an assistant professor of technology, operations, and statistics at New York University.

Robert J. DeRubeis is a professor of psychology and director of clinical training at the University of Pennsylvania.

Robb Willer is a professor of sociology and, by courtesy, psychology and business at Stanford University.

Jane P. Kim is a clinical associate professor of psychiatry and behavioral sciences at Stanford University.

Johannes C. Eichstaedt is an assistant professor of psychology and the Ram and Vijay Shriram HAI Faculty Fellow at Stanford University.

HAI **Stanford University**
Human-Centered
Artificial Intelligence

Stanford HAI: 353 Jane Stanford Way, Stanford CA 94305-5008
T 650.725.4537 **F** 650.123.4567 **E** HAI-Policy@stanford.edu **hai.stanford.edu**