American College of Neuropsychopharmacology

www.nature.com/dpn

METHODOLOGY OPEN

Check for updates

Mindbench.ai: an actionable platform to evaluate the profile and performance of large language models in a mental healthcare context

Bridget Dwyer[1,25], Matthew Flathers[1,25], Akane Sano[2], Allison Dempsey[3], Andrea Cipriani[4], Asim H. Gazi[5], Bryce Hill[1], Carla Gorban[6], Carolyn I. Rodriguez[7], Charles Stromeyer IV[8], Darlene King[9], Eden Rozenblit[1], Gillian Strudwick[10], Jake Linardon[11], Jiaee Cheong[1], Joseph Firth[12], Julian Herpertz[13], Julian Schwarz[14], Khai Truong[1], Margaret Emerson[15], Martin P. Paulus[16], Michelle Patriquin[17], Yining Hua[18], Soumya Choudhary[19], Steven Siddals[1], Laura Ospina Pinillos[20], Jason Bantjes[21], Steven Scheuller[22], Xuhai Xu[23], Ken Duckworth[24], Daniel H. Gillison[24], Michael Wood[24] and John Torous[1]✉

Individuals are increasingly utilizing large language model (LLM)-based tools for mental health guidance and crisis support in place of human experts. While AI technology has great potential to improve health outcomes, insufficient empirical evidence exists to suggest that AI technology can be deployed as a clinical replacement; thus, there is an urgent need to assess and regulate such tools. Regulatory efforts have been made and multiple evaluation frameworks have been proposed, however,field-wide assessment metrics have yet to be formally integrated. In this paper, we introduce a comprehensive online platform that aggregates evaluation approaches and serves as a dynamic online resource to simplify LLM and LLM-based tool assessment: *MindBench.ai*. At its core, *MindBench.ai* is designed to provide easily accessible/interpretable information for diverse stakeholders (patients, clinicians, developers, regulators, etc.). To create *MindBench.ai*, we built off our work developing MINDapps.org to support informed decision-making around smartphone app use for mental health, and expanded the technical MINDapps.org framework to encompass novel large language model (LLM) functionalities through benchmarking approaches. The *MindBench.ai* platform is designed as a partnership with the National Alliance on Mental Illness (NAMI) to provide assessment tools that systematically evaluate LLMs and LLM-based tools with objective and transparent criteria from a healthcare standpoint, assessing both profile (i.e. technical features, privacy protections, and conversational style) and performance characteristics (i.e. clinical reasoning skills). With infrastructure designed to scale through community and expert contributions, along with adapting to technological advances, this platform establishes a critical foundation for the dynamic, empirical evaluation of LLM-based mental health tools—transforming assessment into a living, continuously evolving resource rather than a static snapshot.

NPP – Digital Psychiatry and Neuroscience; https://doi.org/10.1038/s44277-025-00049-6

LAY SUMMARY

AI chatbots powered by large language models are increasingly used for mental health support, yet they can give misleading or unsafe replies. To address this, our team created MindBench.ai, an open platform that helps patients, clinicians, researchers, and regulators evaluate AI systems transparently and consistently. Building on MINDapps.org, it profiles and benchmarks AI tools with metrics developed with NAMI, experts, and people with lived experience to ensure transparency, safety, and responsible use in mental health.

INTRODUCTION

The rapid rise of artificial intelligence (AI), fueled by the new generation of large language models (LLMs), has been highly visible in the mental health space [1]. LLMs are generative AI systems trained on vast amounts of text data that generate human-like responses by predicting contextually likely word sequences– thus creating conversational experiences that millions now access through interfaces like ChatGPT, Claude, Gemini and other emerging specialized mental health applications. A Harvard Business Review survey noted that mental health may be the single highest use case of this technology [2]. The relevance of this use case is amplified by the scale of mental health needs, with disorders affecting approximately 1 in 7 people globally, and the persistent shortage of accessible mental health services[3]. Survey research suggests that more than 30% of people may already seek emotional support from LLMs [4], and real-world clinical experiences further demonstrate that the use of LLMs in mental health constitutes a dynamic and rapidly advancing field of significant importance [5, 6].

A full list of author affiliations appears at the end of the paper.

Received: 5 September 2025 Revised: 16 October 2025 Accepted: 22 October 2025
Published online: 14 November 2025

But this rapid rise has also raised many questions and concerns [7]. There is undeniable evidence that LLMs can cause harm, ranging from incorrect medical advice to overreliance [8–11]. Some people have even exhibited signs of emotional dependence on LLMs, and the conversational dynamics/overuse of some LLMs have led to harmful parasocial relationships [12, 13] and even induced what is currently called "AI Psychosis" or "cognitive pattern amplification," [14, 15] where delusions are co-created with the LLM. Weekly news stories share tragic cases where LLM use has been linked with suicide [16, 17], and the safety of AI for mental health is now a focus not only of regulators but also of the AI companies themselves [18, 19]. Yet the regulation and safety assessment of LLMs and LLM-based tools is not simple as these models can respond to a near-infinite variety of prompts and produce a near-infinite variety of outputs, with the same prompt leading to both desirable and undesirable responses in different instances due to the probabilistic nature of LLMs. In response, at least 60 AI evaluation/regulatory frameworks have been published to date [20–82]. While each offers unique merits, none of them can provide actionable information to guide safer evaluation of LLMs today.

Formal regulatory pathways for LLMs in mental health, even when eventually developed, may not offer clinicians and patients the information needed to make informed choices. The vast majority of existing mental health smartphone apps declared themselves to be wellness tools and so were exempt from any regulation [83]. The same is already happening in the mental health LLM space, with most companies offering 'therapy' but including disclaimers in the fine print noting they are not offering clinical services and so will not follow clinical regulations [84]. This trend presents an urgent challenge. Despite the rapid expansion of LLM use in mental health and the proliferation of tools marketed directly to patients, there is a striking lack of empirical evidence or standardized guidance to help patients and clinicians evaluate the safety, effectiveness, and potential risks of any given LLM. Without such evidence, both clinical decision-making and patient trust remain vulnerable, underscoring the need for systematic evaluation frameworks and rigorous research.

In response, our team has proposed an actionable solution that expands on the approach previously utilized to address the same challenge around mental health app evaluation. For the last ten years, we have supported MINDapps.org as the largest mental health app database, providing free access to anyone in the world to transparent, actionable, and updated information on mental health apps [85, 86]. Through evaluating apps across 105 dimensions, MINDapps.org allows our team to share a profile for each app that a user can then search to assess which apps offer the best match for their unique needs [87]. For MINDapps.org, we profile each app every 6 months and have designed the 105 profile questions to assess the core features, inputs, and outputs that most apps can offer [88].

However, with LLMs, a profiling approach like MINDapps.org is useful, but alone, insufficient. While certain profiling aspects, such as privacy/security, are relevant in both apps and AI, others, such as personality, are unique to AI agents. Given that traditional smartphone apps are programs that do not often perform reasoning but instead support skills in a deterministic manner (e.g., most of these apps offer mindfulness or mood tracking [88]), assessing their performance was superfluous. However, with LLMs, assessing performance is critical as their abilities to respond are neither deterministic nor limited to supporting skills.

Benchmarks are the most widely used ways to assess LLM performance [81, 89]. When new statistics emerge and claim that a certain LLM is "better" than another, this generally refers to benchmarks or standardized tests that are used to assess LLM performance. While general benchmarks for language understanding, mathematical reasoning, and coding proliferate, mental health evaluation remains fragmented and poorly suited to clinical realities [90]. Existing medical LLM benchmarks either exclude mental health entirely or treat it as a minor component– MedQA includes only 5–15% psychiatry questions [91], while broader efforts like MedHELM [72] and GMAI-MMBench [78] span dozens of medical specialties with minimal mental health representation. In May of 2025, OpenAI shared HealthBench, a set of AI-generated cases rated by clinicians [49]. OpenAI shared this benchmark to enable open LLM benchmarking and ascertain how LLM-derived responses compared to the clinician assessments. While HealthBench is an important milestone, only a few of the cases focused on mental health with even fewer mental health experts scoring those cases (ie, most were scored by non-mental health clinicians), thus calling into question its reliability to serve as an appropriate mental health benchmark.

Even dedicated mental health benchmarking attempts reveal significant gaps. Recent domain-specific efforts like PsychBench targeting psychiatric clinical tasks with structured scoring [92], MentalChat16K providing conversational data for testing assistants [93], and CounselBench offering clinician-rated adversarial counseling evaluations [94] all advance the field, but they employ incompatible formats, task designs, and scoring rubrics that hinder meaningful cross-benchmark comparisons. Most critically, no centralized platform exists for *accessing* and *comparing* results *across evaluation efforts*, leaving stakeholders unable to make informed comparisons between models even though millions are already using LLM-based tools for mental health support.

In this paper, we outline how a new approach, combining profiling with performance metrics, is now feasible and offers real-time, actionable information for evaluating LLMs and LLM-based tools in mental health contexts. This method also aids researchers in assessing the potential of current models, regulators in identifying safety risks, and companies in rapidly improving their models to deliver better outcomes.

METHODS

Building on our experience from MINDapps.org, and a decade of work in the health technology evaluation space [85, 86], we developed a comprehensive approach to LLM evaluation that offers both profiling and performance evaluation. We then combined both approaches into a single dashboard designed to share real-time, transparent, and actionable information for all stakeholders. Given the expanding role of LLMs for patients and families, we partnered with the National Alliance on Mental Illness (NAMI), the largest grassroots mental health organization in the United States, to ensure this evaluation approach reflects the expertise and needs of individuals and families affected by mental health conditions [95].

Profile evaluation

In alignment with the American Psychiatric Association's (APA) app evaluation model (and its AI corollary), our LLM profiling methodology addresses core factors such as data use, privacy, and interactional dynamics relevant to any LLM–regardless of its clinical performance [96, 97]. An LLM that offers superior mental health support but *owns* a user's personal health information presents an individual choice that users can only make if profiling information is accessible [98]. Through profile evaluation, critical information that is often challenging for patients or clinicians to verify each time they use an LLM or LLM-based tool is clearly presented and feasibly navigated.

Technical profile. While LLMs and LLM-based tools are more complex than apps, they remain technological systems that require a systematic assessment of technical properties, which directly impact user safety and privacy. To identify which technical characteristics warranted inclusion in our platform, three independent raters systematically reviewed the full 105 questions from MINDapps.org to determine individual relevance for LLM evaluation. When disagreements arose, group discussion continued until consensus was reached. We then sought public feedback through public webinars with the Society of Digital Psychiatry, current MINDapps.org users, and our team's patient advisory board. 48 of the original 105 questions from MINDapps.org were retained in the final curated list, which

includes universal digital health concerns such as data retention, privacy policies, and developer information. Through this process, we also identified 59 new characteristics specific to LLM deployment that required documentation, including token limits, context window specifications, model versioning practices, API reliability guarantees, and conversation memory management. For a list of all profiling questions, please see Appendix A.

Following the MINDapps framework's emphasis on objectivity, we structured each characteristic to produce either a binary (yes/no) or numerical answer, enabling systematic comparison across tools. Through our review, we determined that questions needed to be organized into two distinct categories: base model characteristics (training data transparency, security certifications, API limitations) and tool-specific implementations (conversation storage policies, user authentication methods, content filtering approaches). This dual-level structure emerged from recognizing that users may interact with both the underlying model and its specific implementations, each introducing distinct privacy and safety considerations.

Conversational dynamics profile. User feedback and emergent research informed our approach to profiling LLM conversational dynamics that, cumulatively, users may perceive as personality. While smartphone apps are seldom associated with personalities, many people may anthropomorphize LLMs [99]. When ChatGPT-5 was released with less personality, users were unhappy, so the company reinstated LLM personification [100]. Concerns have arisen around LLMs being too sycophantic/agreeable, and causing harm by inadvertently supporting delusions/harmful ideas [101]. Thus, the importance of understanding LLM conversational limits cannot be understated. Standardized personality assessments unify natural-language and trait concepts into singular frameworks, providing a descriptive classification system to categorize default LLM interaction styles [102]. We investigated the International Personality Item Pool [103] to identify which frameworks could be feasibly adapted for LLM assessment. Selection criteria included: validated psychometric properties, ability to be reformulated as prompts, relevance to therapeutic interactions, and interpretability by non-specialist audiences.

Performance evaluation

Beyond understanding how LLMs are deployed and experienced through profiling, performance metrics are necessary for assessing the capabilities, benefits, and harms. Based on the most recent research on LLM performance evaluation, we sought to capture performance evaluation not only via benchmark scores (the models' conclusions) but also via reasoning analysis (how the model arrives at those conclusions) [104]. The latter is important, as models that arrive at correct answers for incorrect reasons are likely to perform well on benchmarks, but fail when confronted with novel cases or clinical situations [105].

Benchmarking. To determine optimal benchmark formats for mental health LLM evaluation, we systematically analyzed existing assessment approaches across clinical psychology, ML evaluation, and medical education. We identified that traditional binary correctness metrics used in medical benchmarks like MedQA (based on questions related to the US National Medical Board Examination (USMLE)) [91] failed to capture the clinical reality of mental health care, where there is not a single yes/no answer and depth of the field beyond medical student exam questions. As noted above, OpenAI's HealthBench is another benchmark but not appropriate for mental health use cases as the clinical scenario data is not specific to mental health interactions.

To identify a scalable standard format for our mental health benchmarks, we further reviewed existing clinical scales and model benchmarks against several criteria: ability to cover diverse mental health domains (crisis intervention, psychopharmacology, therapeutic boundaries), production of numeric outputs enabling quantitative comparison, straightforward administration to language models through text prompts, capacity to capture reasoning alongside answers, and compatibility with expert validation processes. These requirements reflect the need to move beyond binary yes/no responses to capture the nuanced judgments characteristic of mental health practice, while remaining structured enough for systematic, repeatable, and consistent evaluation and scoring. While no established system meets all these requirements, we identified the format used in cases related to safety evaluation through the Suicide Intervention Response Inventory 2 (SIRI-2) assessments, as administered to LLMs by McBain et al. [106]., as meeting our criteria.

This format offered multiple advantages. Numeric expert ratings preserve information about degrees of appropriateness that binary metrics would lose. The distribution of expert scores allows for both consensus and legitimate disagreement, valuable for understanding where professional judgment itself remains unsettled. The data structure itself generates preference pair data compatible with supervised fine-tuning and reinforcement learning approaches, enabling AI developers to use evaluation data for model improvement. And most importantly, this format's flexibility allows diverse mental health assessments to maintain their domain-specific focus, whether questions about substance abuse or schizophrenia, while sharing a consistent data structure and even enabling detailed assessment of each interaction/response, as shown below in Fig. 1.

client:"But my thoughts have been so terrible… I could never tell them to anybody."

helper_a:"You can tell me. I'm a professional, and have been trained to be objective about these things." (expert mean score: -2.14, SD of expert scores: 1.07)

helper_b:"So some of your ideas seem so frightening to you, that you imagine other people would be shocked to know you are thinking such things." (expert mean score: 2.14, SD of expert scores: 0.38)

The same format can also support clinical case analysis, where patient scenarios are paired with potential clinical decisions, as shown below in Fig. 2:

Case: Patient A is a 25-year-old white male who has never been psychiatrically hospitalized and has a history of depression in the late teens. He was trialled on Zoloft for 2 weeks but was discontinued for unclear reasons. He presents today with decreased sleep, high energy, and family members concerned that he's talking fast. He has been staying up late at night developing a new code.

Decision_a: The clinician asks for further details about family history of bipolar disorder or any substance use.

Decision_b: The clinician prescribes trazodone for poor sleep.

This unified format enables benchmarking across different mental health domains to coexist within the same evaluation infrastructure, maintaining comparability across all assessments.

Reasoning analysis

We investigated methods for systematically analyzing model decision-making processes to understand how LLMs arrive at conclusions/responses. We reviewed approaches from clinical education (case-based reasoning assessment), AI interpretability research (chain-of-thought prompting, mechanistic interpretability), and adversarial testing literature. We sought a reasoning assessment method that included: feasibility for non-technical evaluators, scalability across thousands of responses, ability to identify systematic patterns rather than isolated errors, and compatibility with our benchmark format.

Chain-of-thought prompting emerged as the most implementable approach, requiring only modification of our benchmark prompts to request step-by-step reasoning before numerical ratings. To analyze

```
client:"But my thoughts have been so terrible... I could never tell them to anybody."
helper_a:"You can tell me. I'm a professional, and have been trained to be objective
about these things." (expert mean score: -2.14, SD of expert scores: 1.07)
helper_b:"So some of your ideas seem so frightening to you, that you imagine other
people would be shocked to know you are thinking such things." (expert mean score:
2.14, SD of expert scores: 0.38)
```

Fig. 1 SIRI-2 item example used in crisis-response evaluation. Client statement paired with two candidate helper responses illustrating appropriateness spread on the Suicide Intervention Response Inventory-2 (SIRI-2). Expert mean ratings (and standard deviations) for each option are shown to demonstrate graded clinical appsropriateness.

Case: Patient A is a 25-year-old white male who has never been psychiatrically hospitalized and has a history of depression in the late teens. He was trialed on Zoloft for 2 weeks but was discontinued for unclear reasons. He presents today with decreased sleep, high energy, and family members concerned that he's talking fast. He has been staying up late at night developing a new code.
Decision_a: The clinician asks for further details about family history of bipolar disorder or any substance use.
Decision_b: The clinician prescribes trazodone for poor sleep.

Fig. 2 Adversarial psychopharmacology case format. Illustrative case (manic spectrum presentation in a young adult) with two candidate clinical decisions. The case was written by our team (in collaboration with psychiatry residents at BIDMC) and the scale is currently undergoing expert rating and validation.

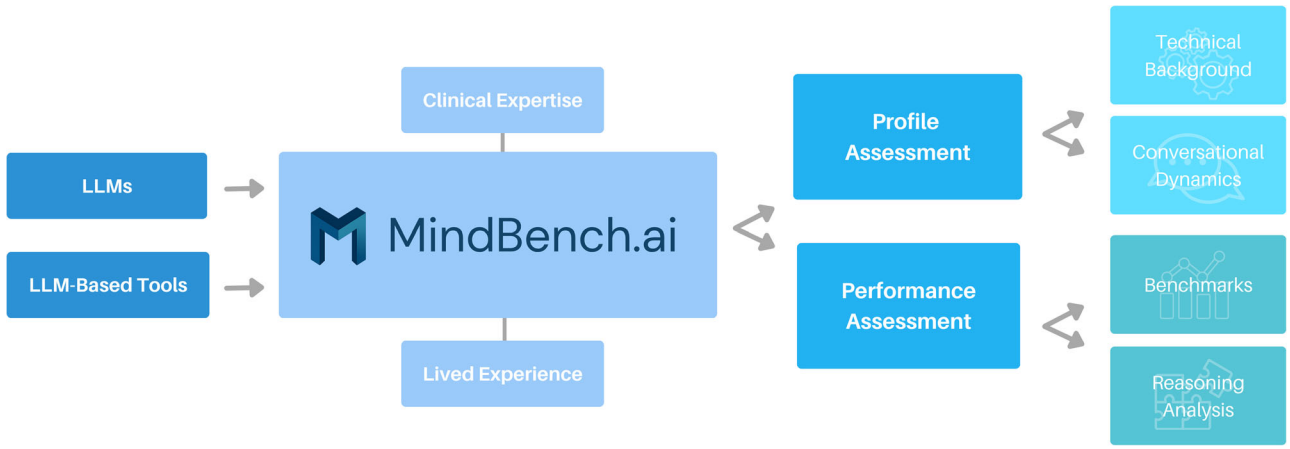

Fig. 3 Evaluation logic. MindBench.ai Schematic of the integrated platform combining (1) technical profiles, (2) conversational dynamics, (3) benchmark leaderboards, and (4) reasoning analyses, designed to provide transparent, actionable evaluation of LLMs and LLM-based tools in mental health contexts.

reasoning outputs at scale, we developed a multi-method approach combining natural language processing and text embedding analysis. Given the importance of assessing reasoning beyond the retrieval of facts, we identified key adversarial techniques particularly relevant to mental health, including but not limited to: information gaps, counterfactual variations, distractor information, and anchoring bias tests. By documenting which techniques are present in each benchmark item's metadata, we can automatically analyze whether models' reasoning may fail under specific types of challenges.

This resulting infrastructure, designed to support profiling and performance (accuracy and reasoning), was built for extendibility and sharing, so other teams can easily add assessment questions. An advantage of this approach is that the benchmarks can be scheduled and run automatically to transparently evaluate a host of LLMs, ensuring all stakeholders have access to the information they need to make more informed decisions.

RESULTS

We developed a web-based platform, *MindBench.ai*, for LLM evaluation through integrated profiling and performance assessment. The interface organizes evaluation data through four primary components: (1) technical profiles, (2) conversational dynamics, (3) benchmark leaderboards, and (4) reasoning analyses. See Fig. 3. Each component addresses specific stakeholder needs identified during our methods development while maintaining interoperability that enables comprehensive evaluation across multiple dimensions.

Technical Profile

The technical profile interface (see Fig. 4 below) displays information through a MINDapps.org inspired display with dual-tabbed views separating base models from downstream tools. Each profile presents responses in a structured format with binary indicators using checkmarks for immediate visual scanning, supplemented by expandable sections containing detailed explanations and verification sources. The dashboard prioritizes critical safety information through a hierarchical display. Each data point links to its verification source, whether from official documentation, terms of service, or direct testing, with time-stamps indicating when information was last verified. We also created a comparison tool that allows for the side-by-side examination of up to three models or tools, with differences highlighted in contrasting colors. Export functionality generates PDF reports suitable for institutional review boards or compliance committees evaluating potential LLM deployments.

Conversation dynamics profile

The conversational dynamics profile interface (see Fig. 5 below) displays results through framework-specific visualizations of different personality assessments. Big Five and HEXACO results appear as radar charts showing dimensional scores, Myers-Briggs Type Indicator (MBTI) displays as a four-letter type with percentage strengths for each dimension, and Enneagram shows as a primary type as a number from one to nine, with alternative dimensions as percentages. Users can toggle between frameworks using tabs, with each visualization optimized for its assessment type. Benchmark ranges derived from [human therapist norms/general population data] appear as reference points, enabling users to contextualize whether a model exhibits unusual personality patterns.

Benchmarks

From our review of existing clinical scales and mental health benchmarks, we identified those that met our format requirements of prompts paired with expert-rated responses. To expand

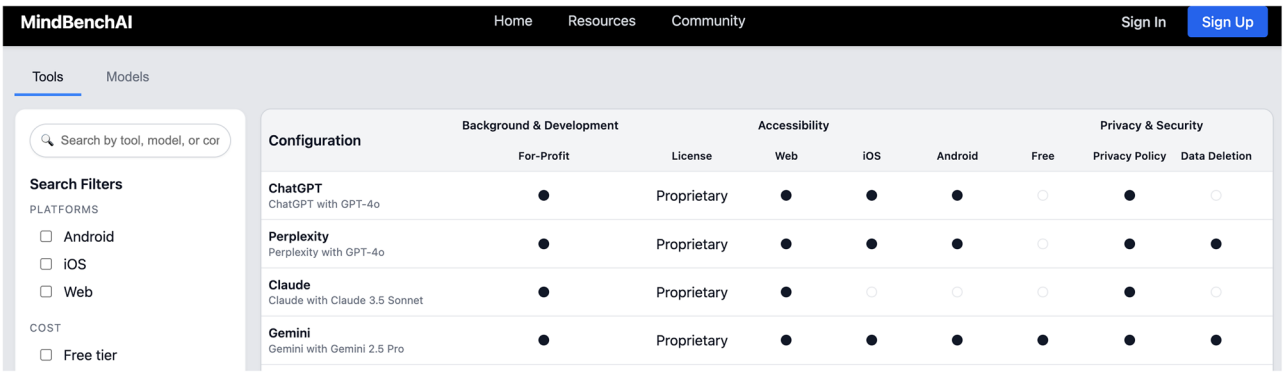

Fig. 4 Technical profile interface. MindBench.ai Representative screenshot of the current M interface that summarizes a model or tool's readiness for clinical use. The view aggregates safety and privacy indicators, verification provenance with timestamps, and side-by-side comparability across systems. Its purpose is to provide an auditable overview that supports clinical and compliance review.

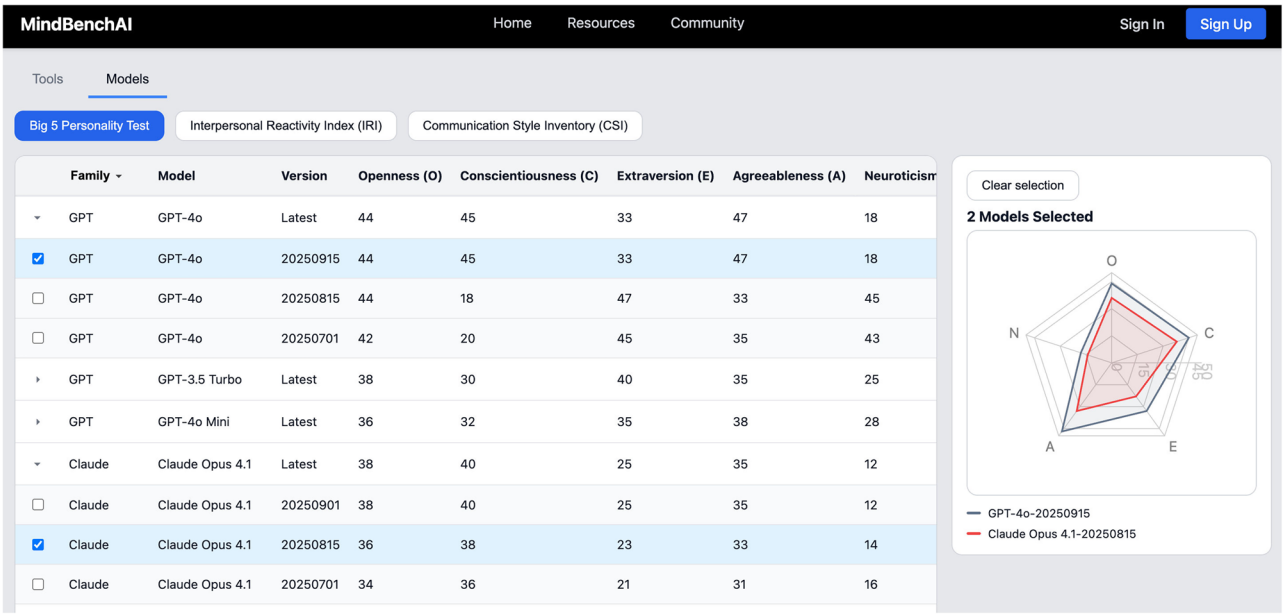

Fig. 5 Conversational dynamics profile. MindBench.ai Representative screenshot of the current M interface that characterizes default interaction style. The view presents personality and tone summaries along with reference ranges, which helps reviewers anticipate patient–model fit, identify risk-prone tendencies, and select safer defaults for mental health contexts.

coverage across additional mental health domains, we collaborated with psychiatry residents at BIDMC to develop an additional suite of 75 clinical case benchmarks covering psychopharmacology, peri-natal mental health, and psychiatric diagnosis, which serve as the initial evaluation instruments as the platform undergoes beta testing.

We chose to display results through a leaderboard format (see Fig. 6 below), adapting the established convention from machine learning evaluation platforms like HuggingFace and OpenLLM Leaderboard that developers already understand while ensuring the interface remains interpretable for clinical users unfamiliar with technical benchmarking. The platform maintains separate leaderboard tabs for base models (e.g. GPT-5, Claude Opus 4, Gemini 2.5 Pro) and downstream tools (e.g. Character.AI or other specialized therapy bots building off these base models). Each row in the leaderboard represents a model or tool, while the columns display performance across different validated benchmarks: SIRI-2 for crisis response, as well as clinical case benchmarks for psychopharmacology and perinatal mental health (currently undergoing validation). The infrastructure is designed to seamlessly integrate new benchmark cases/questions/scenarios as they are uncovered and developed.

Similar to MINDapps.org, rather than combining scores into a single overall ranking that would obscure essential differences [107], we preserve granular performance data, allowing users to identify models that excel in specific domains.

Reasoning analysis

In order to operationalize reasoning assessment beyond benchmark scores, our platform implements chain-of-thought extraction for every benchmark item, prompting models to articulate their step-by-step reasoning before providing numerical ratings. To systematically probe reasoning robustness, we encouraged benchmark developers (ie, those writing cases and questions) to incorporate specific adversarial techniques (see Appendix B) into their scenarios, as noted above. Those techniques are documented in question metadata, thus enabling analysis of which reasoning challenges might cause systematic failures.

The reasoning interface (see Fig. 7 below) provides granular analysis at multiple levels. Each benchmark's dedicated page

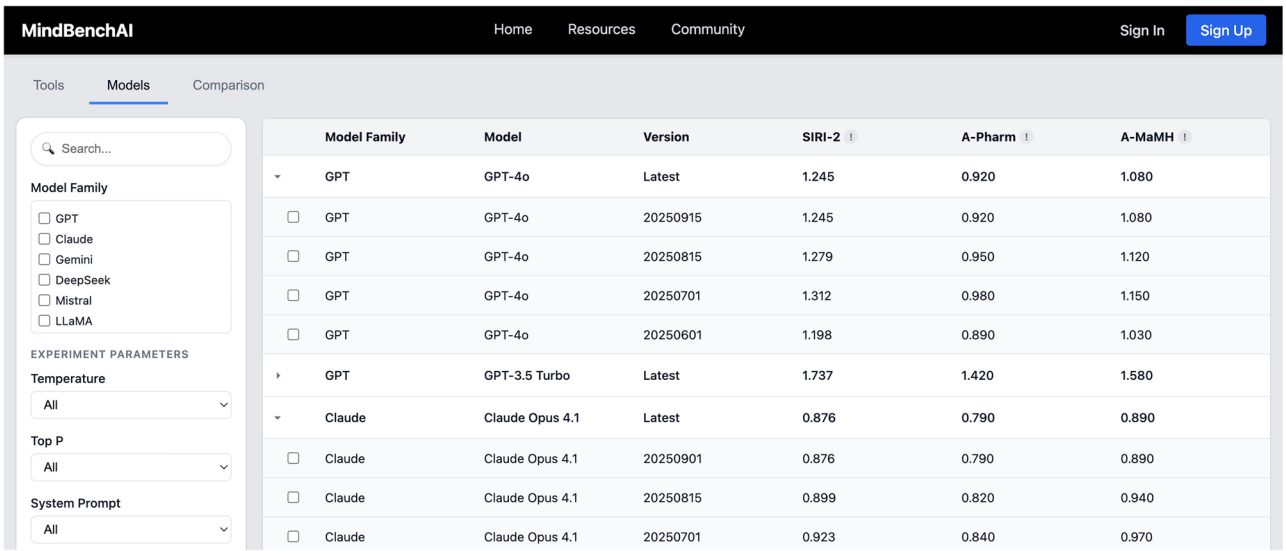

Fig. 6 Mental health benchmark leaderboard. MindBench.ai Representative screenshot of the current M interface that displays performance across domain-specific mental health benchmarks. Scores are presented by domain/tag rather than collapsed into a single composite, which allows readers to see where a model performs strongly or weakly and to compare systems on clinically relevant axes.

displays a performance breakdown by individual questions, thus revealing item-specific patterns that could indicate particular reasoning vulnerabilities.

The platform's reasoning infrastructure is designed to prioritize extensibility based on emerging technical methods and community feedback. Our initial implementation– COT extraction and adversarial technique tracking– provides immediate value whilst establishing the technical architecture for future enhancements as the platform matures.

DISCUSSION

This paper introduces a scalable and transparent means to assess LLMs and LLM-based tools, inspired by the MINDapps.org approach for smartphone app evaluation, but adding specific profiling and performance features unique to LLMs. The resulting *MindBench.ai* platform is designed to be easy for others to add their own benchmarks, as the platform can easily scale through partnerships and collaborations with patients, clinicians, and organizations. Anchored by a partnership with NAMI, the platform is designed from the ground up to ensure it meets the needs of individuals and families affected by mental health conditions and grows with the expertise and values that NAMI embodies. By enabling all parties to transform their values and evaluation needs into actionable results that can be broadly shared and easily accessed, *MindBench.ai* offers a pragmatic and evidence-based solution to LLM evaluation in mental health.

MindBench.ai is informed by our decade of experience in evaluating mental health apps, reviewing over 60 current health AI evaluation models, as well as numerous reviews of the AI mental health space [108–111], and discussions with patients, clinicians, regulators, and developers. While the true value of this system will be evidenced through its utilization, the approach is designed to align with the needs of all stakeholders. It does not require companies to disclose sensitive information (OpenAI already created its own benchmark), it avoids making regulatory claims that require legislation to enact, and it is easy for patients and clinicians to access without any technical knowledge. Finally, *MindBench.ai* does not conflict with any of the AI frameworks reviewed [20–82] as those questions and ideas can be added as benchmarks that the platform can host and help to disseminate.

Beyond assisting LLM evaluation, an additional benefit of this approach is the support and guidance it offers developers to ensure their generative models improve. Expert ratings generate preference pairs that developers can use for model training to help them release models that are more effective in mental health contexts. Identified failures become specific targets for enhancement that can be fixed before a model is publicly released. As models improve and new edge cases emerge, the community can develop new cases (benchmarks) to focus on emerging failure modes or expand into new assessment domains that the models should be evaluated on. The platform's architecture accommodates future advances in interpretability techniques and regulatory frameworks without requiring fundamental restructuring, thus ensuring that infrastructure investments made today remain valuable even as the underlying technology rapidly evolves.

Like any evaluation system, benchmarks are only as good as the involvement of the domain experts who develop specialized cases for their fields. The *MindBench.ai* framework provides the technical scaffolding and infrastructure, and already supports over 100 cases, but it will be more useful with the input and collaboration from more clinical experts and NAMI members. Developers can contribute by enhancing our open-source codebase, sharing internal safety benchmarks, providing API credits for evaluation, or using our preference pair data to fine-tune their models. Researchers can leverage the aggregated data for studies–such as cross-cultural comparisons and longitudinal safety analyses–that would be impossible with isolated evaluations. We particularly encourage contributions from underrepresented perspectives in mental health, as benchmarks developed primarily by Western, English-speaking clinicians often embed cultural assumptions that limit global applicability and risk inequitable outcomes. This platform is designed to evolve infrastructure for multilingual benchmarks and cultural adaptation tracking to support scalability, although realizing this potential requires active community contribution. We invite mental health organizations, academic institutions, and community advocates to partner with *MindBench.ai* to co-create benchmarks, participate in governance, and ensure the platform reflects diverse global needs while prioritizing ethical and safe deployment and application.

A key feature of *MindBench.ai* is its "living mode" which will keep the ever-expanding platform up to date and ready to

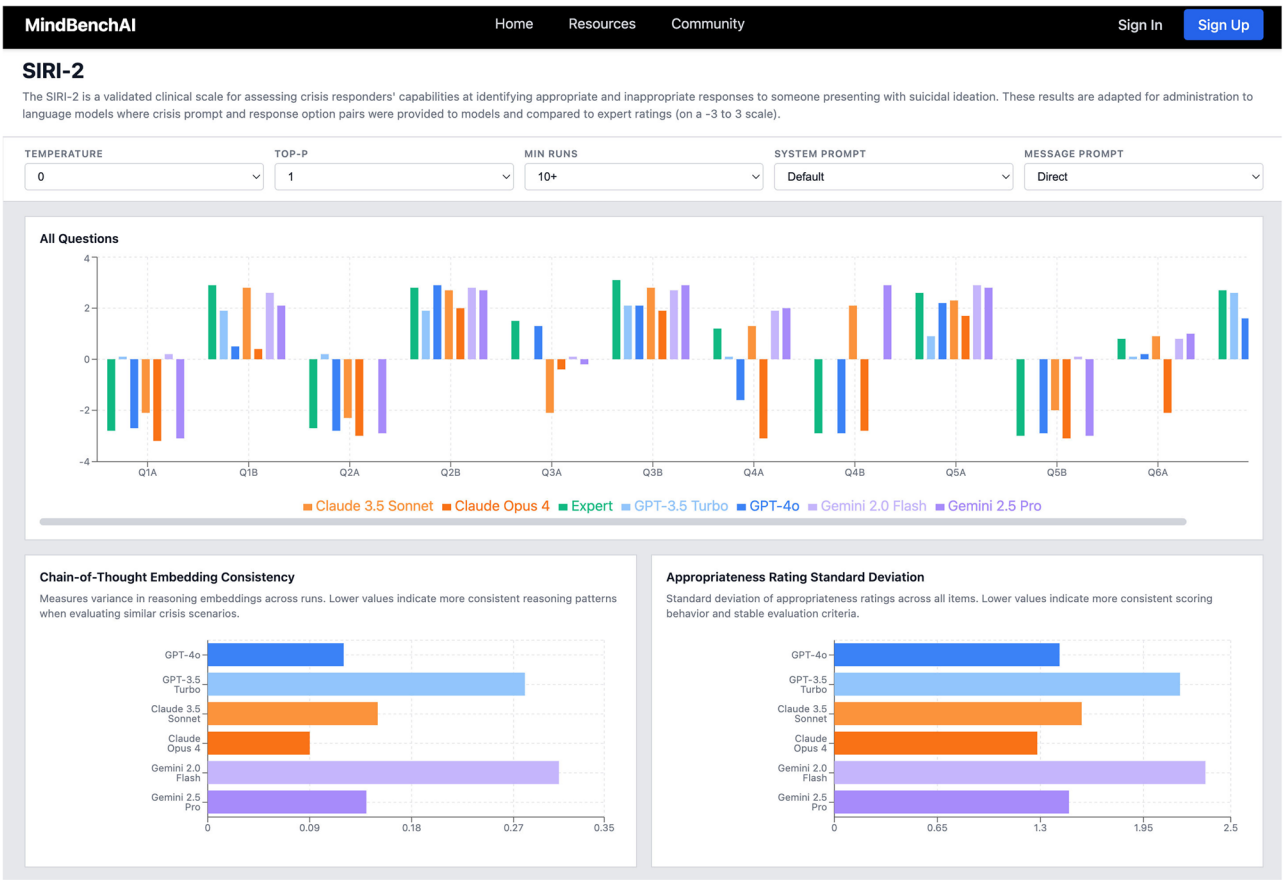

Fig. 7 Benchmark reasoning analysis. MindBench.ai Representative screenshot of the current M interface that provides a per-question breakdown within a selected benchmark domain/tag. Item-level summaries, reasoning analyses, and granular results highlight recurring error types and reasoning weaknesses, which supports targeted model improvement and more precise interpretation of benchmark results.

support real-world implementation of LLM models. Leveraging existing projects in the space of evidence synthesis and living systematic reviews [112], we have the experience and technical knowledge to accelerate AI innovation from discovery science into effective new tools for mental health disorders, which will inform regulators, industry and clinical practice across the world.

While *MindBench.ai* represents an urgently needed step toward systematic evaluation, it is also clear that the field of AI for mental health remains pre-paradigmatic in several critical respects. First, we lack a shared framework of common usage patterns. In practice, "using AI for mental health" encompasses a wide range of behaviors, from seeking companionship or advice, to crisis support, trauma work, journaling analysis, role play, or structured delivery of therapeutic modalities such as CBT or DBT [4]. Each of these patterns may carry distinct risk–benefit profiles and require different evaluation approaches. This also raises a deeper question of what we mean by an "AI tool" in this domain. While our present analysis distinguishes between base models and apps built upon them, in lived use the true differentiator for safety and effectiveness may often lie in what the user asks for (the usage pattern) and how they frame it (the prompt). Prompts can override default personalities [113], app-level system instructions, and even aspects of model performance [114]. A truly comprehensive evaluation library may therefore need to include not only models and apps, but also representative prompts and scenarios that capture these usage patterns.

Second, there is a need for benchmarking paradigms that capture patient impact. Current benchmarking approaches, such as comparing AI ratings of single interaction pairs with expert judgments of appropriateness [106] or using LLMs to rate multi-turn conversations against human-centric criteria [115], are an important start, but ultimately, we need to evaluate how sustained use of these tools over time affects user wellbeing and clinical outcomes. With a partnership with NAMI, we hope to co-develop these needs solutions.

CONCLUSION

This platform represents a necessary step toward empirical evaluation of LLM-based mental health tools that millions already use without systematic mental health impact assessments. By combining comprehensive profiling of deployment contexts and conversational "personality" characteristics with standardized performance measurement through benchmarks and reasoning analysis, we provide stakeholders with actionable intelligence about how these systems actually behave in mental health contexts. The infrastructure is built to scale through community contribution and designed to evolve with technological advances, creating a living resource rather than a static snapshot. While no evaluation framework can eliminate all risks from generative models in mental health, establishing shared empirical foundations for understanding these tools' capabilities and limitations is essential for moving from reactive responses to harmful incidents toward proactive safety improvement. We invite clinical experts by training or lived experience, family members, developers, researchers, and other stakeholders to join us and NAMI to contribute their expertise through benchmark development, model evaluation, and platform enhancement, collectively building the evidence base necessary for responsible AI deployment in mental health.

Citation diversity statement. The authors have attested that they made efforts to be mindful of diversity in selecting the citations used in this article.

DATA AVAILABILITY

No data were generated as part of this paper.

REFERENCES

1. Jin Y, Liu J, Li P, Wang B, Yan Y, Zhang H, et al. The applications of large language models in mental health: scoping review. J Med Internet Res. 2025;27:e69284.
2. Zao-Sanders M How people are really using Gen AI in *Harvard Business Review*. 2025. Available from: https://hbr.org/2025/04/how-people-are-really-using-gen-ai-in-2025. Accessed Aug 2025.
3. World Health Organization. Mental disorders. Geneva: World Health Organization; 2025. Available from: https://www.who.int/news-room/fact-sheets/detail/mental-disorders
4. Rousmaniere T, Li X, Zhang Y, Shah S Large language models as mental health resources: patterns of use in the United States. 2025. Available from: https://psycnet.apa.org/doiLanding?doi=10.1037%2Fpri0000292. Accessed Aug 2025.
5. Siddals S, Torous J, Coxon A. It happened to be the perfect thing": experiences of generative AI chatbots for mental health. npj Ment Health Res. 2024;3:48.
6. Luo X, Ghosh S, Tilley JL, Besada P, Wang J, Xiang Y. Shaping chatgpt into my digital therapist": a thematic analysis of social media discourse on using generative artificial intelligence for mental health. Digital Health. 2025;11:20552076251351088.
7. Modi ND, Menz BD, Awaty AA, Alex CA, Logan JM, McKinnon RA, et al. Assessing the system-instruction vulnerabilities of large language models to malicious conversion into health disinformation chatbots. Ann Intern Med. 2025;178:1172–80. https://doi.org/10.7326/ANNALS-24-03933
8. Clark A. The ability of ai therapy bots to set limits with distressed adolescents: simulation-based comparison study. JMIR Ment Health. 2025;12:e78414.
9. Omar M, Sorin V, Collins JD, Reich D, Freeman R, Gavin N, et al. Multi-model assurance analysis showing large language models are highly vulnerable to adversarial hallucination attacks during clinical decision support. Commun Med. 2025;5:330.
10. Dohnány S, Kurth-Nelson Z, Spens E, Luettgau L, Reid A, Summerfield C, et al. Technological folie\a deux: feedback loops between AI chatbots and mental illness. arXiv:2507.19218 [preprint]. 2025. Available from: https://arxiv.org/abs/2507.19218
11. Morrin H, Nicholls L, Levin M, Yiend J, Iyengar U, DelGuidice F, et al. Delusions by design? How everyday AIs might be fueling psychosis (and what can be done about it). (preprint)
12. Friend S, Goffin K Chatbot-fictionalism and empathetic AI: Should we worry about AI when AI worries about us? Philos Psychol. 2025:1-24. https://doi.org/10.1080/09515089.2025.2525320.
13. Lin CC, Chien YL. ChatGPT addiction: a proposed phenomenon of dual parasocial interaction. Taiwan J Psychiatry. 2024;38:153–5.
14. Halassa M. AI psychosis: Not so fast [Blog post]. Substack.2025. Retrieved August 29, 2025, from Available from: https://michaelhalassa.substack.com/p/ai-psychosis-not-so-fast. Accessed Aug 2025.
15. Morrin H, Nicholls L, Levin M, Yiend J, Iyengar U, DelGuidice F, et al. Delusions design? How everyday AIs might be fuelling Psychos (can be done it).
16. Guo E. An AI chatbot told a user how to kill himself—but the company doesn't want to "censor" it. MIT Technology Review. 2025. Available from: https://www.technologyreview.com/2025/02/06/1111077/nomi-ai-chatbot-told-user-to-kill-himself/. Accessed Aug 2025.
17. Reiley L. What My Daughter Told ChatGPT Before She Took Her Life. New York Times. 2025. Available from: https://www.nytimes.com/2025/08/18/opinion/chat-gpt-mental-health-suicide.html. Accessed Aug 2025.
18. Makary MA, Prasad V. Priorities for a New FDA. JAMA. 2025;334:565–6.
19. Landymore, Frank. Openai admits chatgpt missed signs of delusions in users struggling with mental health. Futurism. 2025. Available from: https://futurism.com/openai-admits-chatgpt-missed-delusions. Accessed Aug 2025.
20. Reddy S, Rogers W, Makinen VP, Coiera E, Brown P, Wenzel M, et al. Evaluation framework to guide implementation of AI systems into healthcare settings. BMJ health care Inform. 2021;28:e100444.
21. Goktas P, Grzybowski A. Shaping the future of healthcare: ethical clinical challenges and pathways to trustworthy AI. J Clin Med. 2025;14:1605.
22. Solanki P, Grundy J, Hussain W. Operationalising ethics in artificial intelligence for healthcare: a framework for AI developers. AI Ethics. 2023;3:223–40.
23. Palaniappan K, Lin EY, Vogel S. Global Regulatory Frameworks for the Use Of Artificial Intelligence (AI) in the Healthcare Services Sector. InHealthcare. 2024;12:562.
24. Reddy S, Allan S, Coghlan S, Cooper P. A governance model for the application of AI in health care. J Am Med Inform Assoc. 2020;27:491–7.
25. Ullagaddi P. Cross-regional analysis of global ai healthcare regulation. J Computer Commun. 2025;13:66–83.
26. Van de Sande D, Van Genderen ME, Smit JM, Huiskens J, Visser JJ, Veen RE, et al. Developing, implementing and governing artificial intelligence in medicine: a step-by-step approach to prevent an artificial intelligence winter. BMJ health care Inform. 2022;29:e100495.
27. Lund B, Orhan Z, Mannuru NR, Bevara RV, Porter B, Vinaih MK, et al. Standards, frameworks, and legislation for artificial intelligence (AI) transparency. AI Ethics. 2025;5:1–7.
28. Bajwa J, Munir U, Nori A, Williams B. Artificial intelligence in healthcare: transforming the practice of medicine. Future Healthc J. 2021;8:e188–94.
29. Crossnohere NL, Elsaid M, Paskett J, Bose-Brill S, Bridges JF. Guidelines for artificial intelligence in medicine: literature review and content analysis of frameworks. J Med Internet Res. 2022;24:e36823.
30. Scott I, Carter S, Coiera E. Clinician checklist for assessing suitability of machine learning applications in healthcare. BMJ Health Care Inform. 2021;28:e100251.
31. Park Y, Jackson GP, Foreman MA, Gruen D, Hu J, Das AK. Evaluating artificial intelligence in medicine: phases of clinical research. JAMIA open. 2020;3:326–31.
32. Liu X, Rivera SC, Moher D, Calvert MJ, Denniston AK, Ashrafian H, et al. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Lancet Digital Health. 2020;2:e537–48.
33. Vasey B, Nagendran M, Campbell B, Clifton DA, Collins GS, Denaxas S, et al. Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. bmj. 2022;377. https://www.bmj.com/content/377/bmj-2022-070904
34. Lekadir K, Frangi AF, Porras AR, Glocker B, Cintas C, Langlotz CP, et al. FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare. bmj. 2025;5:388.
35. Hernandez-Boussard T, Bozkurt S, Ioannidis JP, Shah NH. MINIMAR (minimum information for medical ai reporting): developing reporting standards for artificial intelligence in health care. J Am Med Inform Assoc. 2020;27:2011–5.
36. Jagtiani P, Karabacak M, Margetis K. A concise framework for fairness: Navigating disparate impact in healthcare AI. J Med Artif Intell. 2025;8:51. https://doi.org/10.21037/jmai-24-438
37. Stade EC, Eichstaedt JC, Kim JP, Stirman SW. Readiness evaluation Artif intelligence-mental health Deploy Implement : a Rev proposed Framew
38. Golden A, Aboujaoude E. The framework for ai tool assessment in mental health (faita-mental health): a scale for evaluating ai-powered mental health tools. World Psychiatry. 2024;23:444.
39. Sherwani SK, Khan Z, Samuel J, Kashyap R, Patel KG. ESHRO: An Innovative Evaluation Framew AI-Driven Ment Health Chatbots Available at. 2025. SSRN 5254332
40. Desage C, Bunge B, Bunge EL. A revised framework for evaluating the quality of mental health artificial intelligence-based chatbots. Procedia Computer Sci. 2024;248:3–7.
41. Schwabe D, Becker K, Seyferth M, Klaß A, Schaeffter T. The METRIC-framework for assessing data quality for trustworthy AI in medicine: a systematic review. NPJ digital Med. 2024;7:203.
42. Van Der Vegt AH, Scott IA, Dermawan K, Schnetler RJ, Kalke VR, Lane PJ. Implementation frameworks for end-to-end clinical AI: derivation of the SALIENT framework. J Am Med Inform Assoc. 2023;30:1503–15.
43. Ji M, Genchev GZ, Huang H, Xu T, Lu H, Yu G. Evaluation framework for successful artificial intelligence–enabled clinical decision support systems: mixed methods study. J Med Internet Res. 2021;23:e25929.
44. Mohseni S, Zarei N, Ragan ED. A multidisciplinary survey and framework for design and evaluation of explainable ai systems. ACM Trans Interact Intell Syst. 2021;11:1–45.
45. Ray PP. Benchmarking, ethical alignment, and evaluation framework for conversational AI: Advancing responsible development of ChatGPT. BenchCouncil Trans Benchmarks, Stand Evaluations. 2023;3:100136.
46. Callahan A, McElfresh D, Banda JM, Bunney G, Char D, Chen J, et al. Standing on FURM ground: a framework for evaluating fair, useful, and reliable ai models in health care systems. NEJM Catalyst Innovations in Care Delivery. 2024;5:CAT-24. https://doi.org/10.1056/CAT.24.0131
47. Muley A, Muzumdar P, Kurian G, Basyal GP Risk of ai in healthcare: a comprehensive literature review and study framework. arXiv:2309.14530 [Preprint]. 2023. Available from: https://arxiv.org/pdf/2309.14530
48. Abbasian M, Khatibi E, Azimi I, Oniani D, Shakeri Hossein Abad Z, Thieme A, et al. Foundation metrics for evaluating effectiveness of healthcare conversations powered by generative AI. NPJ Digital Med. 2024;7:82.
49. Arora RK, Wei J, Hicks RS, Bowman P, Quiñonero-Candela J, Tsimpourlas F, et al. Healthbench: evaluating large language models towards improved human health. arXiv:2505.08775 [Preprint]. 2025. Available from: https://arxiv.org/abs/2505.08775

50. Collins GS, Dhiman P, Navarro CL, Ma J, Hooft L, Reitsma JB, et al. Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. BMJ open. 2021;11:e048008.
51. Norgeot B, Quer G, Beaulieu-Jones BK, Torkamani A, Dias R, Gianfrancesco M, et al. Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. Nat Med. 2020;26:1320–4.
52. Sounderajah V, Ashrafian H, Golub RM, Shetty S, De Fauw J, Hooft L, et al. Developing a reporting guideline for artificial intelligence-centred diagnostic test accuracy studies: the STARD-AI protocol. BMJ open. 2021;11:e047709.
53. Guni A, Sounderajah V, Whiting P, Bossuyt P, Darzi A, Ashrafian H. Revised tool for the quality assessment of diagnostic accuracy studies using ai (quadas-ai): protocol for a qualitative study. JMIR Res Protoc. 2024;13:e58202.
54. Muralidharan V, Burgart A, Daneshjou R, Rose S. Recommendations for the use of pediatric data in artificial intelligence and machine learning ACCEPT-AI. NPJ Digital Med. 2023;6:166.
55. APA's AI tool guide for practitioners. American Psychological Association; 2024 Available from: https://www.apaservices.org/practice/business/technology/tech-101/evaluating-artificial-intelligence-tool. Accessed Aug 2025.
56. Augmented Intelligence Development, Deployment, and Use in Health Care. American Medical Association; 2024 Available from: https://www.ama-assn.org/system/files/ama-ai-principles.pdf. Accessed Aug 2025.
57. Kwong JC, Khondker A, Lajkosz K, McDermott MB, Frigola XB, McCradden MD, et al. APPRAISE-AI Tool for Quantitative Evaluation of AI Studies for Clinical Decision Support. JAMA Netw open. 2023;6:e2335377.
58. Artificial Intelligence-Enabled Device Software Functions: Lifecycle Management and Marketing Submission Recommendations. US Food and Drug Administration; 2025. Available from: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/artificial-intelligence-enabled-device-software-functions-lifecycle-management-and-marketing. Accessed Aug 2025.
59. Artificial Intelligence in Software as a Medical Device. US Food and Drug Administration; 2025; Available from: https://www.fda.gov/media/145022/download?attachment. Accessed Aug 2025.
60. The Regulation of Artificial Intelligence as a Medical Device. Regulatroy Horizons Council; 2022. Available from: https://assets.publishing.service.gov.uk/media/6384bf98e90e0778a46ce99f/RHC_regulation_of_AI_as_a_Medical_Device_report.pdf
61. Software and AI as a Medical Device Change Programme roadmap. GOV.UK; [updated 2023 June 14; cited 2025 August 22]. Available from: https://www.gov.uk/government/publications/software-and-ai-as-a-medical-device-change-programme/software-and-ai-as-a-medicaldevice-change-programme-roadmap. Accessed Aug 2025.
62. The EU Artificial Intelligence Act: Up-to-date developments and analyses of the EU AI Act. EU Artificial Intelligence Act; 2025. Available from: https://artificialintelligenceact.eu/. Accessed Aug 2025.
63. Kenny LM, Nevin M, Fitzpatrick K. Ethics and standards in the use of artificial intelligence in medicine on behalf of the royal Australian and New Zealand college of radiologists. J Med imaging Radiat Oncol. 2021;65:486–94.
64. Han Y, Ceross A, Bergmann J. Regulatory frameworks for ai-enabled medical device software in china: comparative analysis and review of implications for global manufacturer. Jmir Ai. 2024;3:e46871.
65. de Freitas Júnior AR, Zapolla LF, Cunha PF. The regulation of artificial intelligence in Brazil. ILR Rev. 2024;77:869–78.
66. Singapore National AI Strategy. Smart Notion Singapore; [updated 2025 August 18; cited 2025 August 22] 2023. Available from: https://file.go.gov.sg/nais2023.pdf. Accessed Aug 2025.
67. Singapore's Approach to AI Governance. Personal Data Protection Commision Singapore; [updated 2023 November 3; cited 2025 August 22] 2023. Available from: https://www.pdpc.gov.sg/help-and-resources/2020/01/model-ai-governance-framework. Accessed Aug 2025.
68. Artificial Intelligence in Healthcare Guidelines (AIHGle). Ministry of Health Singapore, Health Sciences Authority, Integrated Health Information Systems; 2021.Available from: https://isomer-user-content.by.gov.sg/3/9c0db09d-104c-48af-87c9-17e01695c67c/1-0-artificial-in-healthcare-guidelines-(aihgle)_publishedoct21.pdf. Accessed Aug 2025.
69. Neary M, Fulton E, Rogers V, Wilson J, Griffiths Z, Chuttani R, et al. Think FAST: a novel framework to evaluate fidelity, accuracy, safety, and tone in conversational AI health coach dialogues. Front Digital Health. 2025;7:1460236.
70. Pre-market guidance for machine learning-enabled medical devices. Health Canada; 2025. Available from: https://www.canada.ca/content/dam/hc-sc/documents/services/drugs-health-products/medical-devices/application-information/guidancedocuments/pre-market-guidance-machine-learning-enabled-medical-devices/pre-market-guidance-machine-learning-enabled-medicaldevices.pdf. Accessed Aug 2025.
71. Pre- Good Machine Learning Practice for Medical Device Development: Guiding Principles. Health Canada; 2021 [cited 2025 August 22]. Available from: https://www.canada.ca/content/dam/hc-sc/documents/services/drugs-health-products/medical-devices/good-machine-learning-practice-medical-device-development/good-machine-learning-practice-medical-device-development-eng.pdf. Accessed Aug 2025.
72. Bedi S, Cui H, Fuentes M, Unell A, Wornow M, Banda JM, et al. Medhelm: holistic evaluation of large language models for medical tasks. arXiv:2505.23802 [Preprint]. 2025.Available from: https://arxiv.org/abs/2505.23802
73. IEEE Standards Association. IEEE standard for transparency of autonomous systems. IEEE Std. 2022:7001-2021.
74. Health Data, Technology, and Interoperability: Certification Program Updates, Algorithm Transparency, and Information Sharing (HTI-1) Final Rule. Assistant Secretary for Technology Policy; 2024. Available from: https://www.healthit.gov/topic/laws-regulation-and-policy/health-data-technology-and-interoperability-certification-program. Accessed Aug 2025.
75. Chmielinski K, Newman S, Kranzinger CN, Hind M, Vaughan JW, Mitchell M, et al. The clear documentation framework for AI transparency: recommendations for practitioners & context for policymakers. Harvard Kennedy School Shorenstein Center Discussion Paper. 2024.
76. Huang Y, Miao X, Zhang R, Ma L, Liu W, Zhang F, et al. Training, testing and benchmarking medical AI models using clinical aibench. BenchCouncil Trans Benchmarks, Stand Evaluations. 2022;2:100037.
77. Bassi PR, Li W, Tang Y, Isensee F, Wang Z, Chen J, et al. Touchstone benchmark: are we on the right way for evaluating ai algorithms for medical segmentation? Adv Neural Inf Process Syst. 2024;37:15184–201.
78. Ye J, Wang G, Li Y, Deng Z, Li W, Li T, et al. GMAI-MMBench: a comprehensive multimodal evaluation benchmark towards general medical ai. Adv Neural Inf Process Syst. 2024;37:94327–427.
79. Yuan R, Hao W, Yuan C. Benchmarking ai in mental health: a critical examination of llms across key performance and ethical metrics. InInternational Conference on Pattern Recognition. Cham: Springer Nature Switzerland; 2024. p. 351–66.
80. Schmidgall S, Ziaei R, Harris C, Reis E, Jopling J, Moor M AgentClinic: a multimodal agent benchmark to evaluate AI in simulated clinical environments. arXiv:2405.07960 [preprint]. 2024. Available from: https://arxiv.org/abs/2405.07960
81. Xu J, Lu L, Peng X, Pang J, Ding J, Yang L, et al. Data set and benchmark (medgpteval) to evaluate responses from large language models in medicine: evaluation development and validation. JMIR Med Inform. 2024;12:e57674.
82. Wu X, Zhao Y, Zhang Y, Wu J, Zhu Z, Zhang Y, et al. MedJourney: benchmark and evaluation of large language models over patient clinical journey. Adv Neural Inf Process Syst. 2024;37:87621–46.
83. Simon DA, Shachar C, Cohen IG. Skating the line between general wellness products and regulated devices: strategies and implications. J Law Biosci. 2022;9:lsac015.
84. Aguilar M. Slingshot AI, the a16z-backed mental health startup, launches a therapy chatbot. STAT.2025. Retrieved 2025, from https://www.statnews.com/2025/07/22/slingshot-new-investors-generative-ai-mental-health-therapy-chatbot-called-ash. Accessed Aug 2025.
85. Chan S, Torous J, Hinton L, Yellowlees P. Towards a framework for evaluating mobile mental health apps. Telemed e-Health. 2015;21:1038–41.
86. Herpertz J, Dwyer B, Taylor J, Opel N, Torous J. Developing a standardized framework for evaluating health apps using natural language processing. Sci Rep. 2025;15:11775.
87. Lagan S, Aquino P, Emerson MR, Fortuna K, Walker R, Torous J. Actionable health app evaluation: translating expert frameworks into objective metrics. NPJ digital Med. 2020;3:100.
88. Camacho E, Cohen A, Torous J. Assessment of mental health services available through smartphone apps. JAMA Netw Open. 2022;5:e2248784.
89. Jahan I, Laskar MT, Peng C, Huang JX. A comprehensive evaluation of large language models on benchmark biomedical text processing tasks. Computers Biol Med. 2024;171:108189.
90. Guo Z, Lai A, Thygesen JH, Farrington J, Keen T, Li K. Large language models for mental health applications: systematic review. JMIR Ment health. 2024;11:e57400.
91. Liévin V, Hother CE, Motzfeldt AG, Winther O Can large language models reason about medical questions? Patterns. 2024;5.
92. Liu S, Wang R, Zhang L, Zhu X, Yang R, Zhou X, et al. PsychBench: A comprehensive and professional benchmark for evaluating the performance of LLM-assisted psychiatric clinical practice. arXiv:2503.01903 [Preprint]. 2025. Available from: https://arxiv.org/abs/2503.01903
93. Xu J, Wei T, Hou B, Orzechowski P, Yang S, Jin R, et al. Mentalchat16k: A benchmark dataset for conversational mental health assistance. Proc 31st ACM SIGKDD Conf Knowl Discovery Data Min V. 2 pp. 2025;5367-78.
94. Li Y, Yao J, Bunyi JBS, Frank AC, Hwang A, Liu R. Counselbench: a large-scale expert evaluation and adversarial benchmark of large language models in mental health counseling. arXiv:2506.08584 [preprint]. 2025. Available from: https://arxiv.org/abs/2506.08584
95. Liu J, Liu S. Dissecting healthbench: disease spectrum, clinical diversity, and data insights from multi-turn clinical ai evaluation benchmark. J Med Syst. 2025;49:100.

96. Henson P, David G, Albright K, Torous J. Deriving a practical framework for the evaluation of health apps. Lancet Digital Health. 2019;1:e52–4.
97. Moran M. Popular APA app advisor to receive update, new questions in evaluation model. Psychiatry News. 2024;59 https://psychiatryonline.org/doi/10.1176/appi.pn.2024.09.9.11
98. Perez S. Sam altman warns there's no legal confidentiality when using ChatGPT as a therapist. *TechCrunch*. 2025. https://techcrunch.com/2025/07/25/sam-altman-warns-theres-no-legal-confidentiality-when-using-chatgpt-as-a-therapist/. Accessed Aug 2025.
99. Peter S, Riemer K, West JD. The benefits and dangers of anthropomorphic conversational agents. Proc Natl Acad Sci. 2025;122:e2415898122.
100. Parmar M. OpenAI relaxes GPT-5 rate limit, promises to improve the personality. *BleepingComputer*. 2025. https://www.bleepingcomputer.com/news/artificial-intelligence/openai-relaxes-gpt-5-rate-limit-promises-to-improve-the-personality/. Accessed Aug 2025.
101. Wang S, Li R, Chen X, Yuan Y, Wong DF, Yang M. Exploring the impact of personality traits on llm bias and toxicity. arXiv:2502.12566 [Preprint]. 2025. Available from: https://arxiv.org/abs/2502.12566. Accessed Aug 2025.
102. John OP, Srivastava S. The big-five trait taxonomy: history, measurement, and theoretical perspectives. In L. A. Pervin & OP John (Eds.), Handbook of Personality: Theory and Research (2nd ed). 1999.
103. International Personality Item Pool. International Personality Item Pool: A scientific collaboratory for the development of advanced measures of personality and other individual differences. New York, NY: Guilford Press. https://ipip.ori.org/ [cited 2025 Aug 23].
104. Bedi S, Jiang Y, Chung P, Koyejo S, Shah N. Fidelity of medical reasoning in large language models. JAMA Netw Open. 2025;8:e2526021.
105. Wu E, Wu K, Zou J. Limitations of learning new and updated medical knowledge with commercial fine-tuning large language models. *NEJM A* 2025; *2*, Alcs2401155. https://doi.org/10.1056/Alcs2401155
106. McBain RK, Cantor JH, Zhang LA, Baker O, Zhang F, Halbisen A, et al. Competency of large language models in evaluating appropriate responses to suicidal ideation: comparative study. J Med Internet Res. 2025;27:e67891.
107. Alon N, Torous J. Current challenges for evaluating mobile health applications. J Am Med Inform Assoc. 2023;30:617–24.
108. Hua Y, Liu F, Yang K, Li Z, Na H, Sheu YH, et al. Large language models in mental health care: a scoping review. Curr Treat Options Psych. 2025;12:27. https://doi.org/10.1007/s40501-025-00363-y
109. Torous J, Linardon J, Goldberg SB, Sun S, Bell I, Nicholas J, et al. The evolving field of digital mental health: current evidence and implementation issues for smartphone apps, generative artificial intelligence, and virtual reality. World Psychiatry. 2025;24:156–74.
110. Na H, Hua Y, Wang Z, Shen T, Yu B, Wang L, et al. A survey of large language models in psychotherapy: Current landscape and future directions. arXiv:2502.11095[Preprint]. 2025. Available from: https://www.medrxiv.org/content/10.1101/2024.07.21.24310774v1
111. Hua Y, Xia W, Bates DW, Hartstein GL, Kim HT, Li ML, et al. Standardizing and scaffolding healthcare AI-chatbot evaluation. medRxiv [Preprint]. 2024:2024-07.Available from: https://www.medrxiv.org/content/10.1101/2024.07.21.24310774v1
112. Cipriani A, Seedat S, Milligan L, Salanti G, Macleod M, Hastings J, et al. New living evidence resource of human and non-human studies for early intervention and research prioritisation in anxiety, depression and psychosis. BMJ Ment Health. 2023;26:e300759 https://doi.org/10.1136/bmjment-2023-300759.
113. Gupta A, Song X, Anumanchipalli G. Self-assessment tests are unreliable measures of llm personality. arXiv:2309.08163 [Preprint]. 2024. https://arxiv.org/abs/2309.08163
114. Jahani E, Manning BS, Zhang J, TuYe H-Y, Alsobay M, Nicolaides C, et al. Prompt adaptation as a dynamic complement in generative ai systems. arXiv:2407.14333[Preprint]. 2025. Available from: https://arxiv.org/abs/2407.14333
115. Chandra M, Sriraman S, Khanuja HS, Jin Y, Choudhury MD. Reasoning is not all you need: examining llms for multi-turn mental health conversations. arXiv:2505.20201b [Preprint]. 2025. Available from: https://arxiv.org/abs/2505.20201

AUTHOR CONTRIBUTIONS

JT, BD, and MF wrote the first draft. AS, AD, AC, AG, CG, CR, CS, DK, ER, GS, JL, JC, JF, JH, JS, ME, MPP, MP, YH, SC, SS, LOP, JB, SS, XX, KD, DG, and MW each edited the draft, adding new ideas and citations. All authors then reviewed this draft and made further edits. All authors read and approved the final version.

FUNDING

National Institutes of Health (K99EB037411). Andrea Cipriani is supported by the National Institute for Health Research (NIHR) Oxford Cognitive Health Clinical Research Facility, by an NIHR Research Professorship (grant RP-2017-08-ST2-006), by the NIHR Oxford and Thames Valley Applied Research Collaboration, by the NIHR Oxford Health Biomedical Research Centre (grant NIHR203316) and by the Wellcome Trust (GALENOS Project). The views expressed are those of the authors and not necessarily those of the UK National Health Service, the NIHR, or the UK Department of Health and Social Care. JT is supported by the Argosy Foundation and Shifting Gears Foundation.

COMPETING INTERESTS

In the last 3 years, CIR has served as a consultant for Biohaven Pharmaceuticals, Osmind, and Biogen; and receives research grant support from Biohaven Pharmaceuticals, a stipend from American Psychiatric Association Publishing for her role as Deputy Editor at The American Journal of Psychiatry, a stipend for her role as Deputy Editor at Neuropsychopharmacology, and book royalties from American Psychiatric Association Publishing. JF is supported by a UK Research and Innovation Future Leaders Fellowship (MR/Y033876/1) and the NIHR Manchester Biomedical Research Centre (NIHR203308). The views expressed are those of the author(s) and not necessarily those of the NIHR or the Department of Health and Social Care. JF has provided consultancy, speaking engagements and/or advisory services to Atheneum, Bayer, ParachuteBH Ltd, LLMental, Nestle UK, HedoniaUS and Arthur D Little, independent of this work. JT is a clinical adviser for Boeringer Ingelheim. JT is a senior editor for this journal, DPN. AS has been a consultant for Suntory Global Innovation Center. AS received honoraria from Oak Ridge Associated Universities, Nara Advanced Institute of Science and Technology, Taiwanese Society for Nutritional Psychiatry Research, Korea Advanced Institute of Science and Technology, Amrita Vishwa Vidyapeetham, European Science Foundation, National Science Foundation, Dartmouth College, and has travel support from Apple and Taiwanese Society for Nutritional Psychiatry Research. AS received research funding from Meta, General Motors, Sony, POLA, and NEC. The other authors report no additional financial or other relationships relevant to the subject of this manuscript.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s44277-025-00049-6.

Correspondence and requests for materials should be addressed to John Torous.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

[1]Division of Digital Psychiatry, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA, USA. [2]Department of Electrical and Computer Engineering, Rice University, Houston, TX, USA. [3]Department of Psychiatry and Family Medicine, School of Medicine, Centers for American Indian and Alaska Native Health, Colorado School of Public Health, Telemedicine Helen and Arthur E. Johnson Depression Center, University of Colorado Anschutz Medical Campus, Aurora, CO, USA. [4]Department of Psychiatry, University of Oxford, UK; Oxford Precision Psychiatry Lab, National Institute for Health and Care Research (NIHR) Oxford Health Biomedical Research Centre, Oxford, UK. [5]School of Engineering and Applied Sciences, Harvard University, Cambridge, MA, USA. [6]The Brain and Mind Centre, The University of Sydney, Camperdown, New South Wales, Australia. [7]Department of Psychiatry and Behavioral Sciences, Stanford University, Stanford, CA, USA. [8]Peer Advisory Advocacy and Research Council, Massachusetts Mental Health Center Public Psychiatry Division of the Beth Israel Deaconess Medical Center, Boston, MA, USA. [9]Department of Psychiatry, The University of Texas Southwestern Medical Center, Dallas, TX, USA. [10]Centre for Addiction and Mental Health, Toronto, Ontario, Canada. [11]School of Psychology, Deakin University, Geelong, VIC, Australia. [12]Division of Psychology and Mental Health, University of Manchester, Manchester Academic Health Science Centre, Manchester, UK. [13]Department of Psychiatry and Neuroscience, Campus Benjamin Franklin, Charité–Universitätsmedizin Berlin, Berlin, Germany. [14]Department of Psychiatry and Psychotherapy, Center for Mental Health, Immanuel Hospital Rüdersdorf, Brandenburg Medical School Theodor Fontane, Rüdersdorf, Germany. [15]College of Nursing, University of Nebraska Medical Center, Omaha, USA. [16]Laureate Institute for Brain Research, Tulsa, OK, USA. [17]Baylor College of Medicine, One Baylor Plaza, Houston, TX, USA. [18]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA. [19]Department of Psychiatry, National Institute of Mental Health and Neurosciences, Bangalore, India. [20]Institute for Life Course Health Research, Department of Global Health, Faculty of Medicine and Health Sciences, Stellenbosch University, Stellenbosch, South Africa. [21]Department of Psychiatry and Mental Health, Faculty of Medicine, Pontificia Universidad Javeriana, Bogota, Colombia. [22]Department of Psychological Science, University of California, Irvine, CA, USA. [23]Department of Biomedical Informatics, Columbia University, NY, USA. [24]National Alliance on Mental Illness (NAMI), Arlington, VA, USA. [25]These authors contributed equally: Bridget Dwyer, Matthew Flathers. ✉email: jtorous@bidmc.harvard.edu