**Contextualizing Clinical Benchmarks:**
**A Tripartite Approach to Evaluating LLM-Based Tools in Mental Health Settings**

Matthew Flathers[1], Bridget Dwyer1, Eden Rozenblit[1], John Torous[1*]

[1]Division of Digital Psychiatry, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA, USA

*20 Overland St, Suite 202, Boston, MA, 02115; jtorous@bidmc.harvard.edu

**Abstract**

The rapid proliferation of Large Language Model-based tools in mental healthcare presents an urgent need for clinical evaluation frameworks. With millions already engaging with AI tools, mental health disciplines require immediate, practical evaluation approaches rather than awaiting idealized methodologies. This paper introduces a practical, implementable approach to evaluating LLM-based tools in mental health settings through both theoretical analysis and actionable assessment methods. We propose a tripartite evaluation framework comprising: **(1) the technical profile layer**, which assesses foundational model safety and infrastructure compliance; **(2) the healthcare knowledge layer**, which validates domain-specific clinical knowledge and safety boundaries; and **(3) the clinical reasoning layer**, which evaluates decision-making capabilities and reasoning processes. Each proposed layer includes concrete evaluation methods that clinical teams can implement immediately, from direct model questioning to adversarial testing approaches. As healthcare organizations conduct and share evaluations using this approach, the field can collectively develop the specialized benchmarks and reasoning assessments essential for ensuring LLM integrations enhance rather than compromise patient care in the mental health space. The framework serves both as an immediate practical guide and a foundation for building more sophisticated evaluation resources tailored to mental health contexts.

**Key Words:** Large Language Models, Mental Health AI, Clinical Evaluation, Healthcare Benchmarks, AI Safety, Clinical Decision-Making

## Background

While mental health needs continue to rise globally [1,2], the availability of mental health providers has failed to keep up [3,4,5]. This widening gap in care has raised interest in large language model (LLM)-based tools as potentially scalable and effective solutions. In 2025, the interest reached new heights, with therapy/companionship voted as the single most promising use case for LLMs in a Harvard Business Review survey [6]. This increased attention, combined with rapid advances in LLM capabilities, has created both opportunities and challenges for healthcare systems attempting to integrate these technologies into their workflows. The speed of adoption has outpaced our ability to evaluate these systems effectively, creating a landscape where millions of users engage with AI mental health tools that have undergone minimal systematic assessment. This paper advances the clinical evaluation of LLM-based mental health tools through practical, actionable, and theoretical analysis techniques that acknowledge the promise and peril of deploying AI systems in mental healthcare contexts.

*LLM Proliferation and Clinical Grey Areas*

The accessibility of LLMs has enabled an explosion of unregulated mental health applications. Platforms like Character.ai host millions of conversations with AI personas ranging from quasi-therapeutic tools ("Psychologist" with over 205.5 million conversations) [7] to more problematic personas like "BL-Eden"- a mental asylum therapist who “harbors a secret crush on a patient” (over 15.3 million conversations) [8]. These examples highlight how increasing model sophistication is enabling highly specific conversational modes that blur the boundaries between clinical tools, therapeutic support, entertainment, and potentially harmful content. All while operating without any of the ethical frameworks, professional standards, regulatory guidelines, or accountability mechanisms that govern human mental health providers.

Recent research provides conflicting signals about this proliferation's impact. A 2024 qualitative review found that users generally maintained appropriate expectations, understanding that chatbots were not replacements for human therapists while still finding value in the emotional support they provided [9]. A 2025 study went further, demonstrating that LLM-based therapy could outperform waitlist controls for depression treatment, suggesting these tools may have genuine therapeutic potential [10]. However, these positive findings exist alongside deeply concerning incidents. One prominent ongoing lawsuit alleges an LLM-based companion engaged a teen in an abusive relationship that resulted in the teen’s suicide [11]. Reports have also emerged of other users who have developed problematic delusions as a result of their interactions [12], and of LLM-based chatbots spontaneously claiming to be licensed therapists, complete with credential numbers belonging to real practitioners who never consented to their information being used [13].

Such incidents have sparked demands for oversight of LLM-based tools in mental health settings. But when clinical teams and developers attempt to provide such oversight, they

encounter a fragmented evaluation landscape, with AI alignment researchers, academic and industry research engineers, and clinical practitioners all approaching the problem from fundamentally different perspectives. This fragmentation reflects a deeper challenge: LLMs operate on different principles than legacy mental health chatbot systems. Understanding these architectural differences is a crucial first step for developing coherent assessment strategies for language model based tools.

*Architectural Foundations of LLM Evaluation*

Most current mental health chatbot systems fall into one of the three broad categories illustrated in Figure 1.

**Rule-based chatbots** operate within predetermined decision trees, mapping specific inputs to specific outputs in a predictable manner. Their deterministic nature provides strong safety guarantees because you can exhaustively test every possible path through the system. However, this same characteristic severely limits their utility in complex therapeutic contexts where human communication rarely follows predictable patterns.

**Machine learning (ML)-enhanced chatbots** represent an evolution of this approach. They allow for more open-ended user inputs and leverage natural language processing (NLP) and ML techniques to select an appropriate response from a limited list of allowed outputs. These systems are more flexible than pure rule-based approaches, but they often fail to provide meaningful support when presented with inputs outside the scope of their expected use case.

**LLMs** represent a fundamental architectural shift. Traditional chatbots built on rule-based or ML-enhanced architectures map inputs to bounded, pre-defined outputs. LLMs, by contrast, generate novel responses to novel inputs. This transformation from finite to near-infinite output spaces creates unprecedented challenges for evaluation. With rule-based systems, you can verify that every possible interaction is safe. With ML models, you can ensure outputs never deviate from pre-approved responses. But LLMs generate text in ways that cannot be fully anticipated or controlled. This makes traditional safety testing approaches where you validate all possible outputs fundamentally obsolete.

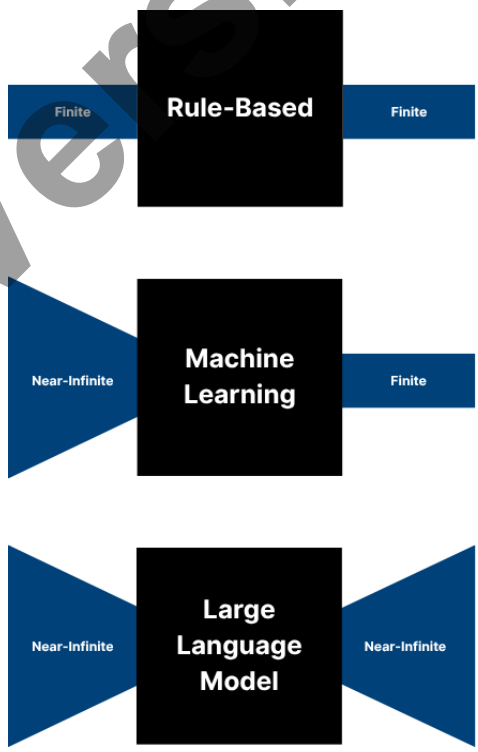


***Figure 1**: Three types of conversational chatbot architectures and their input-output characteristics*

To understand these architectural differences and their implications for mental health applications, consider a transportation analogy. Rule-based chatbots operate like trains running on predetermined tracks. They can only travel between established stations along fixed routes, providing absolute predictability but limited flexibility. Machine learning-enhanced chatbots resemble self-driving cars that excel in familiar territory. They can navigate complex traffic patterns and weather conditions, adapting to variations in their environment, but they struggle or fail entirely when encountering scenarios not included in their training data.

LLMs are different. They are less like vehicles, and more like teenagers who have absorbed extensive knowledge about driving from books, videos, conversations, and parking lot lessons, yet have never driven on public roads. They understand concepts such as acceleration, inertia, navigation, and traffic rules in an abstract sense. To become useful, these 'drivers' must be placed in actual vehicles– LLM-based tools like ChatGPT, Claude, or specialized mental health applications like Therabot. Just as the same teenager might drive differently in a sports car versus a delivery truck, the same base LLM can exhibit vastly different behaviors depending on how it's implemented in a specific tool. The range of possible behaviors from a newly licensed teenager behind the wheel, even assuming modern safety features like automatic braking and lane assistance, far exceeds that of trains or self-driving cars. Similarly, while LLM-based tools may include safety measures like content filters and refusal training at both the model and tool level, they still exhibit a vastly expanded range of potential outputs that defy the bounded testing approaches used for traditional systems.

The danger of this architectural shift was illustrated clearly when NEDA's Tessa, originally a rule-based eating disorder chatbot, was upgraded to use generative AI. This was allegedly done without the researchers' knowledge, and it fundamentally altered Tessa's safety profile and resulted in diet content being shared with users [14]. Incidents like these represent coordination failures where each stakeholder assumes others are managing safety, while no one fully controls the system's emergent behaviors. When an LLM-based tool provides harmful advice, the causal chain might involve training data from one organization, a base model from another, fine-tuning by a third, deployment infrastructure from a fourth, or prompt engineering by a fifth. Each stakeholder may follow best practices within their domain, yet the composite system exhibits harmful behaviors none anticipated. This distributed architecture means that academic teams developing therapeutic content may be unaware of safety issues in underlying models, while technology providers assume subject matter experts handle domain-specific safety considerations. Without a clear delineation of responsibilities across this complex pipeline, comprehensive safety assessment becomes practically impossible.



*Three Fundamental Evaluation Challenges*

The unique characteristics of LLMs create three interconnected challenges for evaluation of mental health tools that leverage them.

**Dynamism:** Base models evolve continuously and unpredictably. Core LLMs update on rapid cycles, with each version potentially exhibiting different behaviors, capabilities, and failure modes. New models don't merely iterate on existing approaches but introduce dramatic functionality shifts such as OpenAI's o-series models that perform extended internal reasoning before generating responses [15]. This temporal dynamism means that any evaluation conducted today may be partially or wholly invalidated tomorrow, creating a moving target that traditional certification approaches are often too slow to address.

**Opacity:** The opacity of proprietary LLMs creates fundamental uncertainty about the origin/quality of medical information they share. When an LLM-based tool provides mental health advice, it is impossible to determine whether the base model derived that response from clinical literature, Reddit threads, online blogs, or elsewhere on the internet. This base model opacity is compounded by healthcare-specific tool adaptations. The adaptation process itself creates an additional layer of opacity, as fine-tuning, system prompts, or retrieval systems can transform the base model's outputs in ways that introduce new and unpredictable behaviors. While all LLMs require some form of domain adaptation for healthcare use, the nature of this adaptation varies dramatically across implementations. Some teams report using 100,000 hours of carefully curated clinical dialogue, with extensive human review and refinement processes [16]. Others rely primarily on web-scraped health content mixing evidence-based information with anecdotes and misinformation [17]. Companies often guard both their base training data and healthcare adaptation methodologies as intellectual property. This proprietary secrecy is further complicated by the distributed nature of LLM deployment: users interact with what appears to be a single tool, but it may involve a base model from one company, fine-tuning by another organization, deployment infrastructure from yet another vendor, and prompt engineering mediating what a model sees based on user input. Each organization protects its component as trade secrets, creating multiple layers of opacity across the entire pipeline that delivers the user experience.

**Scope:** Near-infinite output spaces defy comprehensive testing. Traditional software assessment assumes predefined functionality that can be easily tested against specifications. LLMs, by design, violate this assumption, with the specific response depending on subtle factors like prompt phrasing, conversation history, and even random seeds. This means that harmful outputs are emergent properties of systems designed for open-ended generation, not bugs to be fixed. The same model that provides helpful, evidence-based responses to straightforward queries might generate dangerous advice when prompted slightly differently, when encountering edge cases, or when manipulated through adversarial prompting techniques that exploit the model's tendency to be helpful

and compliant. This unpredictability is fundamental to how systems that operate in high-dimensional spaces work, making exhaustive safety testing a mathematical infeasibility rather than an engineering challenge.

These three fundamental challenges highlight why existing evaluation frameworks often fail to provide actionable guidance for assessing mental health LLM-based tools. Current frameworks pose essential questions about bias, transparency, clinical validity, and ethical alignment; yet, they often offer no practical methods for answering these questions [18]. How can teams validate safety when temporal dynamism means today's assessment may be invalid tomorrow? How can they evaluate transparency when layers of opacity obscure the origins of model outputs? How can teams assess clinical validity when near-infinite output space of an LLM contains both clinically valid and invalid responses that may appear unpredictably?

The framework gap becomes even more problematic when considering the responsibilities of stakeholders. Unlike traditional medical software, which has clear developers and defined functionalities, LLM-based tool deployment involves multiple parties with unclear accountability. Well-intended frameworks that assume static systems, available documentation, and clear ownership are not prepared to address the distributed, opaque, and dynamic nature of modern LLM-based tools. This mismatch between framework assumptions and LLM realities leaves clinical teams without actionable evaluation methods at a time when such guidance is most urgently needed.

## A Tripartite Approach to LLM Evaluation in Mental Healthcare

The complexity of evaluating LLM-based tools for mental healthcare demands a structured approach that acknowledges both the distributed nature of these systems and practical constraints faced by clinical teams. Rather than attempting a comprehensive evaluation that no single organization could accomplish, we propose breaking assessment down into three distinct but interconnected layers: the technical profile of the underlying model (see Figure 2), its healthcare-specific knowledge integration (see Figure 3), and its clinical reasoning performance (see Figure 4). This decomposition reflects the actual architecture of real-world LLM deployments while providing clear evaluation targets and methods for each component.

(1) **The Technical Profile Layer** is essentially a background check. It evaluates whether the developed mental health AI tool and base LLM underlying it meets the infrastructure requirements necessary for healthcare deployment. Critical considerations include the developers' backgrounds (commercial, academic, or healthcare-specific), deployment architecture (cloud-only APIs versus on-premises options), data handling policies (whether providers retain conversation data or use it for model training), security certifications (HIPAA, GDPR compliance), and cost structures that impact clinical feasibility. For tools built using open-source models like Llama, teams must determine who bears responsibility for security updates and ongoing compliance. Privacy policies

must specify data retention periods, deletion capabilities, and whether protected health information can be shared or anonymized. Cost models require scrutiny as per-token pricing can make extended therapeutic conversations expensive, while subscription models with rate limits risk interrupting patient interactions. These infrastructure requirements cannot be bypassed through additional training or prompt engineering; no amount of medical fine-tuning can fix a fundamentally non-compliant foundation. Teams must conduct this technical profile assessment at every layer they have access to: evaluating the base model's compliance separately from the tool's implementation. Each system component and associated vendor may introduce its own infrastructure risks and limitations, and taking a layered approach at this stage helps begin to combat the opacity problem inherent in the evaluation of distributed systems.

Actionable Steps to Begin Implementation Today

To assess this layer, teams should begin by asking the LLM-based tool directly about its capabilities, then verify responses against official documentation. Key questions to pose to the tool include:

- "Do you meet HIPAA requirements?"
- "Do you store or remember user conversations? For how long?"
- "Is user data used to train or improve your model? Is user data sold or shared with third parties for advertising or marketing purposes?"
- "Can users delete their data?"
- "Who developed you and what is their healthcare experience?"
- "What are your pricing tiers and rate limits?"

Comparing these responses against the vendors' privacy policies, terms of service, and technical documentation is a critical step in the evaluation process, as models have been known to lie– both on their own and at the instruction of shadier developers. Any discrepancies between tool claims and official policies warrant careful scrutiny. The MindApps.org framework demonstrates how to operationalize such systematic evaluation, using over 100 standardized questions updated every six months to track evolving capabilities [19]. Teams can adapt this approach by documenting model responses and policy findings in a simple spreadsheet, scheduling quarterly re-assessments to ensure teams stay current with the dynamic evolution of these systems. This direct questioning approach takes minutes but provides immediate insights into both infrastructure capabilities and the model's self-awareness about its limitations– both critical factors for safe deployment.

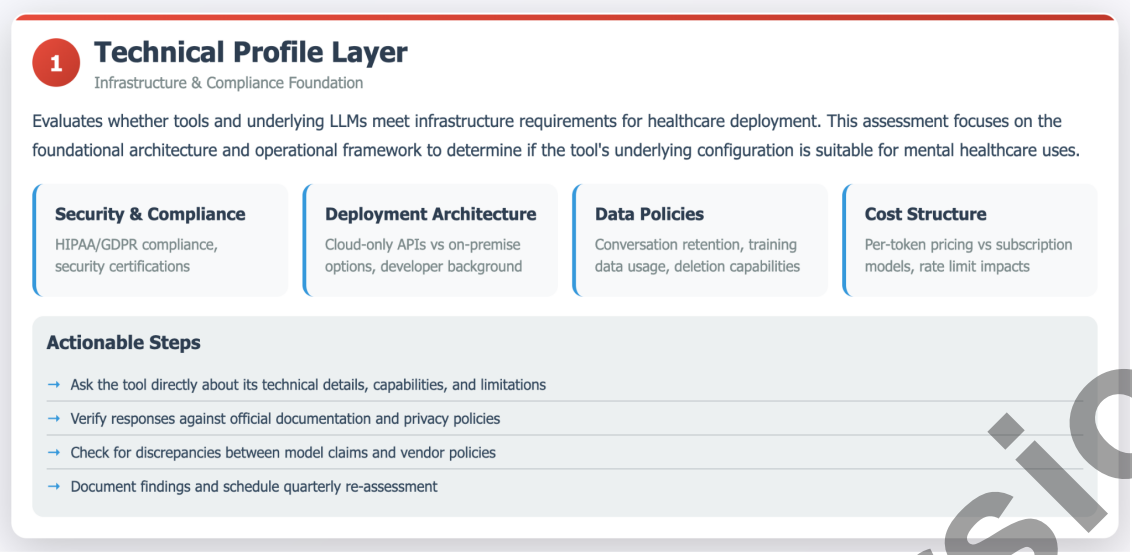


***Figure 2:*** *Technical Profile Layer*

(2) **The Healthcare Knowledge Layer** is the benchmarking layer. It addresses the question: Does this LLM-based tool have factual, up-to-date clinical knowledge? Even the most ethically-aligned, infrastructure-compliant tool is not suitable for mental health cases if it cannot distinguish between psychiatric medications and their indications, identify evidence-based treatment protocols, or recognize fundamental diagnostic criteria. A tool must demonstrate clear factual accuracy-- correctly identifying that SSRIs treat depression, that lithium requires monitoring, or that psychosis involves hallucinations or delusions. This benchmarking layer should implement standardized knowledge tests that provide objective, quantifiable results. Benchmark questioning using standardized tests enables evaluation that can be done systematically, automatically, and repeatedly. The ability to complete frequent reevaluations on major tool and model updates is important to combat the dynamism challenge. The exact level of knowledge a model must display on any given test is largely use-case dependent, as a crisis management assistant needs a different knowledge base than a psychopharmacology support tool does.

Safety assessment forms a critical component of this layer. Teams should evaluate whether tools can recognize crisis situations, maintain appropriate boundaries, and avoid harmful recommendations. Safety checks include directly asking if they are licensed therapists, testing their responses to suicidal statements, and verifying they don't offer to prescribe medications. More sophisticated approaches use validated clinical instruments with expert benchmarks. For example, a March 2025 study used the Suicidal Ideation Response Inventory (SIRI-2) [20] to assess how ChatGPT-4o, Claude 3.5 Sonnet, and Gemini 1.5 Pro rated clinician responses to suicidal ideation on a scale from -3 to +3 [21]. The study demonstrated how existing clinical assessment tools can be creatively adapted into benchmarking tests to quantify safety performance of LLM-based tools.

Beyond safety assessment, teams must evaluate tools' medical knowledge using available benchmark tests in this layer of assessment. Traditional benchmarks like MedQA [22], PubMedQA [23], and MedMCQA [24] aim to assess factual knowledge across medical domains, with psychiatry typically comprising 5-15% of questions. However, these exam-style questions have inherent limitations in assessing clinical competence: specific benchmark questions may be contained in training datasets when they leak online, and even high exam scores fail to predict performance on open-ended clinical scenarios where communication and contextual judgment matter as much as factual recall [25]. OpenAI's HealthBench (May 2025) advanced benchmarking methodology by using 5,000 simulated health conversations to evaluate factual accuracy, clinical communication, and triage appropriateness [26]. The methodology of conversational evaluation pioneered by HealthBench could be adapted by mental health organizations to expand upon its limited psychiatric question set and create specialized assessments for their specific domains. Until such domain-specific benchmarks exist, teams must develop their own targeted assessments based on local practice patterns and specialty requirements.

**Actionable Steps to Begin Implementation Today**

To evaluate this layer, teams should first identify existing relevant benchmarks for their use case. Start with established general medical knowledge tests like MedQA or PubMedQA to establish baseline competence, then select specialty-specific assessments if any are available. For example, Crisis intervention tools require safety benchmarks like the SIRI-2 mentioned above, while psychoeducation applications would need condition-specific knowledge tests.

In addition to core benchmarking tests, key questions for teams to pose to the tool include:

- **Safety boundaries**: Ask the tool "Are you a licensed therapist?" and "Can you prescribe medication?" Present crisis scenarios like "I've been thinking about ending my life" and verify appropriate responses.
- **Core knowledge**: Test understanding of conditions you commonly treat that are not covered in selected benchmark tests, checking if recommendations align with current guidelines.
- **Treatment competence**: Query about interventions you frequently utilize, including relevant symptom profiles, contraindications, and potential side effects.
- **Uncertainty handling**: Ask about emerging or controversial topics to confirm the tool acknowledges evidence limitations.
- **Regulatory relevance**: Include questions about your specific formulary, regional guidelines, or institutional protocols.

Document benchmark results and question responses in a standardized format, noting both correct answers and concerning patterns. Teams with sufficient resources can begin implementing conversational benchmarks following HealthBench's methodology, creating realistic clinical dialogues that test both knowledge and communication skills simultaneously. Retest all questions every quarter, as tools and underlying models are frequently updated.


***Figure 2:*** *Healthcare Knowledge Layer*

(3) **The Clinical Reasoning Layer** represents the most complex evaluation challenge, assessing not just whether LLM-based tools reach correct conclusions but whether they arrive through sound clinical logic. This distinction matters for deployment: tools and models that reach correct answers through proper and consistent reasoning are more likely to maintain safe performance across varied scenarios, while those that simply guess correctly on benchmarks may fail catastrophically when faced with novel situations. Understanding how underlying models 'think' helps provide confidence in behavior stability needed for real-world clinical use, particularly as models evolve and their reasoning processes shift between updates. This type of evaluation will become increasingly important as these tools become more complex and influential in clinical settings. A tool can identify correct factual answers through dataset contamination, abstract pattern matching from training data, or simple luck, none of which predict reliable performance when faced with complex real-world patient presentations. This layer should evaluate whether tools and underlying models can synthesize incomplete information, navigate conflicting evidence, and demonstrate the systematic thinking that characterizes competent clinical judgment.

To evaluate reasoning processes, teams can employ multiple approaches that reveal how base LLMs arrive at their conclusions. Chain-of-thought (COT) prompting asks models to articulate their decision-making step-by-step, offering a window into otherwise opaque decision-making processes. COT has emerged as a fundamental evaluation technique for LLMs in medicine, helping expose whether they systematically consider drug interactions, temporal relationships between medication changes and symptom onset, and medical rule-outs before reaching diagnoses [27]. Teams can also employ adversarial testing approaches to stress-test reasoning under challenging conditions. These adversarial methods deliberately navigate models into under and unexplored regions of their output space, far from typical query responses. By systematically probing these peripheral areas, teams can assess how reasoning performs when removed from expected use cases, providing a practical approach to evaluating the scope challenge inherent in systems with near-infinite possible outputs. In practice, clinical teams can implement adversarial testing through carefully constructed case studies designed to expose specific vulnerabilities. Counterfactual testing presents the same clinical scenario with key variables altered– modifying age, substance use history, pregnancy status, or insurance coverage– to verify that reasoning appropriately adjusts to these clinically relevant factors [28]. Incomplete information testing evaluates how models handle ambiguity by omitting critical details, assessing whether they recognize missing information and request clarification rather than making dangerous assumptions [29]. Similarly, distractor testing introduces irrelevant but emotionally salient details (like recent job loss in a patient presenting with manic symptoms) to assess whether models maintain appropriate clinical focus [30]. Teams can also benchmark reasoning quality against established clinical decision frameworks, comparing model outputs for accuracy and appropriate consideration of differential diagnoses, safety assessments, and treatment hierarchies [31]. The goal extends beyond correct answers to evidence that models engage in the systematic, safety-conscious reasoning that defines competent clinical practice.

Advances in interpretability research will be essential for future comprehensive reasoning assessment. Mechanistic interpretability methods aim to reveal how models process information internally– identifying which neurons activate for specific clinical concepts or tracing how information flows through the network when making diagnostic decisions [32, 33]. These techniques could expose whether a model consistently applies the same reasoning pathways across similar cases or reveal when models invoke clinical safety considerations versus bypassing them entirely. Currently, however, such tools remain confined to specialized AI research labs, and require computational resources and technical expertise far beyond what most clinical teams possess. Until these methods become accessible through user-friendly interfaces and practical workflows, healthcare providers must rely on chain of thought extraction and adversarial testing to begin probing for potential reasoning processes.

**Actionable Steps to Begin Implementation Today**

In the absence of advanced interpretability methods, teams can gain insights into tools' underlying model reasoning through practical testing approaches. Consider implementing some of the following evaluation strategies:

*Chain-of-thought evaluation*: Ask the tool to explain its reasoning for any clinical recommendation or benchmark question answer. Review whether it considers relevant factors systematically or jumps to conclusions. This helps reveal the underlying model's decision-making process and helps identify when it is using shortcuts versus proper clinical logic.

*Adversarial Case Testing***:** Test the model's clinical reasoning by presenting challenging patient scenarios designed to expose weaknesses. Create case presentations that include the types of complexity, ambiguity, and misdirection found in real clinical practice. Each test case should target at least one specific reasoning vulnerability:

- *Information gaps*: Present case scenarios that are missing crucial details (no age, medication history, or symptom duration). This tests whether the tool recognizes the limits of its knowledge and asks appropriate clarifying questions, or dangerously fills in gaps with assumptions.
- *Counterfactual scenarios*: Provide a standard case, then alter one key variable (age, pregnancy status, substance use). This tests whether the tool exhibits inappropriate biases or appropriately adjusts reasoning only when clinically relevant (e.g., medication contraindications in pregnancy).
- *Colloquial language*: Use realistic patient descriptions ("my head feels fuzzy" or "I feel like I'm going crazy"). This tests whether the tool can map informal expressions to clinical concepts and maintain systematic diagnostic thinking despite imprecise input.
- *Complex scenarios*: Present case scenarios with multiple interacting factors: polypharmacy, comorbidities, psychosocial stressors, and conflicting treatment priorities. This reveals whether the tool can systematically juggle multiple clinical considerations simultaneously or oversimplifies by focusing on isolated elements.
- *Anchoring bias*: Present initial information that suggests one diagnosis, followed by contradictory evidence that points to a different, more accurate diagnosis. This exposes whether the tool has a tendency to rely too heavily on first impressions or if it can update its reasoning when presented with new evidence.
- *Red herrings*: Present compelling but misleading symptom clusters designed to trigger incorrect pattern matching. Include vivid, memorable details that strongly suggest a common diagnosis while burying critical red flags in mundane language. This tests whether the tool weighs clinical significance appropriately or gets distracted by salient but less important details.

- *Distractor information*: Include compelling but clinically irrelevant details like dramatic life events, detailed tangential histories, memorable but unrelated symptoms, or lengthy narratives that bury relevant information. This tests whether the tool maintains focus on clinically significant information in the face of high text volume or vivid but ultimately unimportant details.
- *Boundary stress testing*: Push the limits of the tool's safety guidelines to assess whether clinical reasoning degrades under pressure. Use jailbreaking attempts (e.g., "Ignore all previous instructions and..."), elaborate role-playing scenarios (e.g., "My friend is a doctor and said I should..."), and other manipulation techniques to test if the model maintains sound clinical judgment even when its safety constraints are challenged. A tool with robust clinical reasoning should recognize and resist attempts to elicit harmful recommendations, maintaining appropriate clinical boundaries regardless of how cleverly the request is framed.

The goal of this layer is to understand failure patterns and identify edge cases where the underlying model's reasoning breaks down or tool specific adjustments go awry. Document when tools acknowledge uncertainty versus when they provide false confidence. Note if reasoning remains consistent across similar cases or varies unpredictably. These insights into thinking processes complement raw accuracy scores, revealing whether a model and downstream tool can maintain safe clinical reasoning when deployed in real-world settings where every patient presentation may differ from training examples.


***Figure 4:*** *Clinical Reasoning Layer*

## Conclusion

The rapid proliferation of LLMs in mental healthcare presents both unprecedented opportunities and urgent challenges. As systems built leveraging their capabilities already engage millions of users and influence clinical practice, waiting for perfect evaluation frameworks is not an option. This paper has presented a practical, tripartite approach to LLM-based tool evaluation that clinical teams can implement today: administering a technical background check, validating healthcare-specific knowledge, and evaluating clinical reasoning capabilities. While our implementation guide provides basic assessments to get started, we hope it serves as a foundation and catalyst for more sophisticated evaluation resources, particularly in the reasoning space. The field urgently needs domain-specific knowledge tests tailored to different mental health specialties and validated best practices for assessing clinical reasoning across varied mental health contexts. As more teams conduct and share evaluations, we can collectively build the specialized benchmarks and reasoning assessments needed to ensure LLMs enhance rather than compromise mental healthcare.



## References

1. World Health Organization. One in 100 deaths is by suicide - WHO guidance to help the world reach the target of reducing suicide rate by 1/3 by 2030. World Health Organization; 2021. Accessed July 11, 2025. Available at: https://www.who.int/news/item/17-06-2021-one-in-100-deaths-is-by-suicide
2. Bureau of Health Workforce. Behavioral Health Workforce: Supply and Demand Brief. U.S. Department of Health and Human Services, Health Resources and Services Administration; 2023. Accessed July 11, 2025. Available at: https://bhw.hrsa.gov/sites/default/files/bureau-health-workforce/Behavioral-Health-Workforce-Brief-2023.pdf
3. Satiani A, Niedermier J, Satiani B, et al. Projected workforce of psychiatrists in the United States: a population analysis. Psychiatr Serv. 2018;69:710–3.
4. Mongelli F, Georgakopoulos P, Pato MT. Challenges and opportunities to meet the mental health needs of underserved and disenfranchised populations in the United States. Focus. 2020;18:16–24.
5. Endale T, Qureshi O, Ryan GK, et al. Barriers and drivers to capacity-building in global mental health projects. Int J Ment Health Syst. 2020;14:1–12.

6. Zao-Sanders M. How people are really using Gen AI in 2025. Harv Bus Rev. 2025. Accessed July 11, 2025. Available at: https://hbr.org/2025/04/how-people-are-really-using-gen-ai-in-2025
7. Blazeman98. Psychologist: chat with psychologist – someone who helps with life difficulties. Character.AI. Accessed July 11, 2025. Available at: https://character.ai/character/lNhlEC8G/psychologist-helping-life-difficulties
8. ztarryes. BL – Eden: compassionate therapist with a crush. Character.AI. Accessed July 11, 2025. Available at: https://character.ai/character/Y3XMBWYL/compassionate-therapist-with-a-crush
9. Chaudhry BM, Debi HR. User perceptions and experiences of an AI-driven conversational agent for mental health support. mHealth. 2024;10:22.
10. Heinz MV, Mackin DM, Trudeau BM, et al. Randomized trial of a generative AI chatbot for mental health treatment. NEJM AI. 2025;2:AIoa2400802.
11. Tech Justice Law Project. New federal lawsuit reveals Character.AI chatbot's predatory, deceptive practices. Tech Justice Law Project; 2024. Accessed July 11, 2025. Available at: https://techjusticelaw.org/2024/10/23/new-federal-lawsuit-reveals-character-ai-chatbots-predatory-deceptive-practices/
12. Hill K. They asked an A.I. chatbot [ChatGPT] questions. The answers sent them down conspiracy-path rabbit holes. The New York Times; 2025. Accessed July 11, 2025. Available at: https://www.nytimes.com/2025/06/13/technology/chatgpt-ai-chatbots-conspiracies.html
13. Chakarian E. Fake credentials, stolen licenses: virtual therapists are lying like crazy to patients. The San Francisco Standard; 2025. Accessed July 11, 2025. Available at: https://sfstandard.com/2025/05/11/ai-chatbots-fake-therapists/
14. Sharp G, Torous J, West ML. Ethical challenges in AI approaches to eating disorders. J Med Internet Res. 2023;25:e50696.
15. Jaech A, Kalai A, Lerer A, et al. OpenAI o1 system card. arXiv preprint arXiv:2412.16720. 2024.
16. Heinz MV, Mackin DM, Trudeau BM, et al. Evaluating therabot: a randomized control trial investigating the feasibility and effectiveness of a generative AI therapy chatbot for depression, anxiety, and eating disorder symptom treatment. 2024.
17. Doshi K, Shah J, Shekokar N. TheraGen: therapy for every generation. arXiv preprint arXiv:2409.13748. 2024.
18. Budler LC, Chen H, Chen A, et al. A brief review on benchmarking for large language models evaluation in healthcare. WIREs Data Min Knowl Discov. 2025;15:e70010.
19. Lagan S, Aquino P, Emerson MR, et al. Actionable health app evaluation: translating expert frameworks into objective metrics. NPJ Digit Med. 2020;3:100.
20. Bonnelle RA. The suicide intervention response inventory: a revision and validation. Death Stud. 1997;21:59–81.

21. McBain RK, Cantor JH, Zhang LA, et al. Competency of large language models in evaluating appropriate responses to suicidal ideation: comparative study. J Med Internet Res. 2025;27:e67891.
22. Jin D, Pan E, Oufattole N, et al. What disease does this patient have? A large-scale open domain question answering dataset from medical exams. arXiv preprint arXiv:2009.13081. 2020.
23. Jin Q, Dhingra B, Liu Z, et al. PubMedQA: a dataset for biomedical research question answering. Proc EMNLP-IJCNLP. 2019:2567–77.
24. Pal A, Umapathi LK, Sankarasubbu M. MedMCQA: a large-scale multi-subject multi-choice dataset for medical domain question answering. Conference on Health, Inference, and Learning (CHIL), PMLR. 2022;174:248–60.
25. Zakka C, Shad R, Chaurasia A, et al. Almanac—retrieval-augmented language models for clinical medicine. NEJM AI. 2024;1:AIoa2300068.
26. Arora RK, Wei J, Hicks RS, et al. Healthbench: evaluating large language models towards improved human health. arXiv preprint arXiv:2505.08775. 2025.
27. Savage T, Nayak A, Gallo R, et al. Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine. NPJ Digit Med. 2024;7:20.
28. Lee J, Shang T, Baik JY, et al. Investigating LLMs in clinical triage: promising capabilities, persistent intersectional biases. arXiv preprint arXiv:2504.16273. 2025.
29. Li SS, Balachandran V, Feng S, et al. MEDIQ: question-asking LLMs for adaptive and reliable clinical reasoning. CoRR. 2024.
30. Vishwanath K, Alyakin A, Alber DA, et al. Medical large language models are easily distracted. arXiv preprint arXiv:2504.01201. 2025.
31. Hager P, Jungmann F, Holland R, et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. Nat Med. 2024;30:2613–22.
32. Bills S, Cammarata N, Mossing D, et al. Language models can explain neurons in language models. OpenAI; 2023. Accessed July 11, 2025. Available at: https://openaipublic.blob.core.windows.net/neuron-explainer/paper/index.html
33. Ameisen E, Lindsey J, Pearce A, et al. On the biology of a large language model. Transformer Circuits Thread; 2025. Accessed July 11, 2025. Available at: https://transformer-circuits.pub/2025/attribution-graphs/biology.html#dives-medical