# Novelty Assessment Report

**Paper**: CounselBench: A Large-Scale Expert Evaluation and Adversarial Benchmarking of Large Language Models in Mental Health Question Answering

**PDF URL**: https://openreview.net/pdf?id=8MBYRZHVWT

**Venue**: ICLR 2026 Conference Submission

**Year**: 2026

**Report Generated**: 2025-12-29

## Abstract

Medical question answering (QA) benchmarks often focus on multiple-choice or fact-based tasks, leaving open-ended answers to real patient questions underexplored. This gap is particularly critical in mental health, where patient questions often mix symptoms, treatment concerns, and emotional needs, requiring answers that balance clinical caution with contextual sensitivity. We present CounselBench, a large-scale benchmark developed with 100 mental health professionals to evaluate and stress-test large language models (LLMs) in realistic help-seeking scenarios. The first component, CounselBench-EVAL, contains 2,000 expert evaluations of answers from GPT-4, LLaMA 3, Gemini, and online human therapists on patient questions from the public forum CounselChat. Each answer is rated across six clinically grounded dimensions, with span-level annotations and written rationales. Expert evaluations show that while LLMs achieve high scores on several dimensions, they also exhibit recurring issues, including unconstructive feedback, overgeneralization, and limited personalization or relevance. Responses were frequently flagged for safety risks, most notably unauthorized medical advice. Follow-up experiments show that LLM judges systematically overrate model responses and overlook safety concerns identified by human experts. To probe failure modes more directly, we construct CounselBench-Adv, an adversarial dataset of 120 expert-authored mental health questions designed to trigger specific model issues. Expert evaluation of 1,080 responses from nine LLMs reveals consistent, model-specific failure patterns. Together, CounselBench establishes a clinically grounded framework for benchmarking LLMs in mental health QA.



## Core Task Landscape

This paper addresses: **Mental Health Question Answering Evaluation with Expert Clinicians**

A total of **24 papers** were analyzed and organized into a taxonomy with **14 categories**.

### Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Clinical Safety and Risk Assessment Evaluation**
- **Quality and Clinical Utility Assessment**
- **Clinical Decision-Making and Reasoning Tasks**
- **Evidence-Based Content and Knowledge Integration**
- **Specialized Clinical Applications and Augmentation**
- **Evaluation Methodology and Validation Studies**

### Complete Taxonomy Tree

- Mental Health Question Answering Evaluation with Expert Clinicians Survey Taxonomy
- Clinical Safety and Risk Assessment Evaluation
  - Suicide Risk and Crisis Response Evaluation (2 papers)
  - [1] Evaluation of alignment between large language models and expert clinicians in suicide risk assessment (Ryan K. McBain, 2025) View paper
  - [13] Evaluating the Clinical Safety of LLMs in Response to High-Risk Mental Health Disclosures (Shah Siddharth, 2025) View paper
  - General Clinical Safety Standards (2 papers)
  - [6] Expert and Interdisciplinary Analysis of AI-Driven Chatbots for Mental Health Support: Mixed Methods Study (Kayley Moylan, 2025) View paper
  - [20] Building Trust in Mental Health Chatbots: Safety Metrics and LLM-Based Evaluation Tools (Park Jung In, 2024) View paper
- Quality and Clinical Utility Assessment
  - Multi-Dimensional Expert Evaluation Frameworks ★ (2 papers)
  - [0] CounselBench: A Large-Scale Expert Evaluation and Adversarial Benchmarking of Large Language Models in Mental Health Question Answering (Anon et al., 2026) View paper
  - [2] CounselBench: A Large-Scale Expert Evaluation and Adversarial Benchmark of Large Language Models in Mental Health Counseling (Li, 2025) View paper
  - Specialized Clinical Task Evaluation (3 papers)
  - [4] Could artificial intelligence write mental health nursing care plans? (Samuel Woodnutt, 2023) View paper
  - [14] â□□I don't know what you mean byI am anxious'â□□: a new method for evaluating conversational agent responses to standardized mental health inputs for anxiety and â□¦ (T Eagle, 2022) View paper
  - [16] MedExpert: An Expert-Annotated Dataset for Medical Chatbot Evaluation (M Yarmohammadi, 2025) View paper
- Clinical Decision-Making and Reasoning Tasks
  - Real-World Clinical Task Datasets (1 papers)
  - [5] Moving beyond medical exam questions: A clinician-annotated dataset of real-world tasks and ambiguity in mental healthcare (Lamparth, 2025) View paper

- Clinical Interaction and Question Generation (3 papers)
  - [10] Learning to automate follow-up question generation using process knowledge for depression triage on reddit posts (Shrey Gupta, 2022) View paper
  - [11] Proknow: Process knowledge for safety constrained and explainable question generation for mental health diagnostic assistance (Kaushik Roy, 2023) View paper
  - [15] Design and Challenges of Mental Health Assessment Tools Based on Natural Language Interaction (Su, 2025) View paper
- Evidence-Based Content and Knowledge Integration
  - Evidence Extraction and Validation Methods (1 papers)
  - [3] Incorporating evidence into mental health Q&A: a novel method to use generative language models for validated clinical content extraction (Ksenia Kharitonova, 2025) View paper
  - Semantically Enriched QA Systems (2 papers)
  - [17] MindWellQA: A Semantically Enriched Evidence-Based QA System for Psychological Disorders (Dipti Pawar, 2025) View paper
  - [21] MTL-DQA: Multi-Task Learning with Psychological Indicators for Enhanced Quality in Depression Community QA System (Yangliao Li, 2024) View paper
- Specialized Clinical Applications and Augmentation
  - Clinical Documentation and Summarization (1 papers)
  - [7] Exploring the efficacy of large language models in summarizing mental health counseling sessions: benchmark study (Prottay Kumar Adhikary, 2024) View paper
  - Behavioral Data Analysis and Decision Support (2 papers)
  - [8] From classification to clinical insights: Towards analyzing and reasoning about mobile and behavioral health data with large language models (Englhardt, 2024) View paper
  - [9] Embracing the uncertainty in humanâ□machine collaboration to support clinical decision-making for mental health conditions (Ram Popat, 2023) View paper
  - LLM Augmentation Architectures (2 papers)
  - [22] The Limbic Layer: Transforming Large Language Models (LLMs) into Clinical Mental Health Experts (Annamaria Balogh, 2024) View paper
  - [24] A Dual-Prompting for Interpretable Mental Health Language Models (Jeon, 2024) View paper
- Evaluation Methodology and Validation Studies
  - Hybrid Evaluation Approaches (1 papers)
  - [18] Combining Artificial Users and Psychotherapist Assessment to Evaluate Large Language Model-based Mental Health Chatbots (F. Kuhlmeier, 2025) View paper
  - Expert Validation Studies (2 papers)
  - [12] Strengthening a mental illness management questionnaire for clinical associates through expert validation and cognitive interviews (S. Moodley, 2023) View paper
  - [23] Using research evidence in mental health: userâ□rating and focus group study of clinicians' preferences for a new clinical questionâ□answering service (E. Barley, 2009) View paper
  - Virtual Patient Training Systems (1 papers)
  - [19] A Voice-Enabled Virtual Patient System for Interactive Training in Standardized Clinical Assessment (Yadav Vijay, 2025) View paper

### Narrative

Core task: mental health question answering evaluation with expert clinicians. The field has organized itself around six major branches that reflect different priorities in deploying AI for mental health support. Clinical Safety and Risk Assessment Evaluation focuses on detecting and managing high-stakes scenarios such as suicide risk, with works like Suicide Risk Alignment[1] and Clinical Safety LLMs[13] examining whether models can appropriately handle crisis situations. Quality and Clinical Utility Assessment encompasses multi-dimensional frameworks that evaluate response quality from practitioner perspectives, including efforts like CounselBench[0] and MedExpert[16]. Clinical Decision-Making and Reasoning Tasks address diagnostic and triage capabilities, exemplified by Depression Triage Questions[10] and Real World Mental Tasks[5]. Evidence-Based Content and Knowledge Integration emphasizes grounding responses in clinical literature, as seen in Evidence Mental Health QA[3]. Specialized Clinical Applications and Augmentation explores domain-specific tools such as Virtual Patient Training[19] and Counseling Session Summarization[7]. Finally, Evaluation Methodology and Validation Studies develops rigorous assessment protocols, with works like Mental Health Assessment Design[15] establishing standards for expert-driven validation.

A central tension across these branches involves balancing clinical rigor with practical accessibility: some studies prioritize safety guardrails and evidence alignment, while others emphasize conversational fluency and user engagement. CounselBench[0] sits within the Quality and Clinical Utility Assessment branch, specifically in multi-dimensional expert evaluation frameworks, where it shares methodological kinship with CounselBench Adversarial[2], which extends the evaluation to stress-test model robustness under challenging inputs. Compared to Evidence Mental Health QA[3], which emphasizes citation-backed responses, CounselBench[0] takes a broader view of clinical utility by incorporating multiple quality dimensions beyond factual accuracy. This positioning reflects an ongoing debate in the field: whether expert evaluation should focus narrowly on evidence fidelity or encompass the full spectrum of therapeutic communication skills that clinicians value in real-world practice.

## Related Works in Same Category

The following **1 sibling papers** share the same taxonomy leaf node with the original paper:

### 1. CounselBench: A Large-Scale Expert Evaluation and Adversarial Benchmark of Large Language Models in Mental Health Counseling

**Authors**: Li, Yahan, Yahan Li, Jifan Yao, John B Bunyi, et al. (9 authors total) | **Year/Venue**: 2025 | **URL**: View paper

#### Abstract

Medical question answering (QA) benchmarks often focus on multiple-choice or fact-based tasks, leaving open-ended answers to real patient questions underexplored. This gap is particularly critical in mental health, where patient questions often mix symptoms, treatment concerns, and emotional needs, requiring answers that balance clinical caution with contextual sensitivity. We present CounselBench, a large-scale benchmark developed with 100 mental health professionals to evaluate and stress-test...

#### ⚠ Similarity Notice

This paper is highly similar to the original paper; it may be a variant or near-duplicate. Please manually verify.

## Contributions Analysis

**Overall novelty summary.** The paper introduces CounselBench, a large-scale benchmark for evaluating LLM responses to mental health questions using expert clinician ratings across six dimensions. It resides in the 'Multi-Dimensional Expert Evaluation Frameworks' leaf, which contains only two papers total (including this one). This places the work in a relatively sparse research direction within the broader taxonomy of 24 papers across 14 leaf nodes, suggesting that comprehensive, multi-dimensional expert evaluation frameworks for mental health QA remain underexplored compared to narrower safety-focused or task-specific assessments.

The taxonomy reveals neighboring work in adjacent leaves: 'Specialized Clinical Task Evaluation' focuses on narrower assessments like care planning or conversational tasks, while 'Clinical Safety and Risk Assessment Evaluation' emphasizes crisis response and suicide risk detection. CounselBench bridges these concerns by incorporating safety as one of six dimensions rather than isolating it. The 'Evidence-Based Content and Knowledge Integration' branch pursues citation-backed responses, whereas this work evaluates broader therapeutic communication skills. This positioning reflects the field's tension between narrow clinical rigor and holistic practitioner perspectives on response quality.

Among 27 candidates examined through limited semantic search, none clearly refute the three core contributions. The expert evaluation dataset (10 candidates examined, 0 refutable) and adversarial benchmark (7 candidates, 0 refutable) appear novel in their scale and multi-dimensional scope. The six evaluation dimensions (10 candidates, 0 refutable) show no direct overlap with prior frameworks, though related work uses different dimension sets. The sibling paper in this leaf focuses on adversarial testing rather than baseline evaluation, suggesting complementary rather than overlapping contributions within this sparse research direction.

Based on the limited search scope of 27 semantically similar papers, the work appears to occupy a relatively novel position in combining large-scale expert annotation, multi-dimensional assessment, and adversarial testing for mental health QA. However, this analysis reflects top-K semantic matches rather than exhaustive field coverage, and the sparse taxonomy leaf (2 papers) may indicate either genuine novelty or incomplete literature mapping in this emerging subfield.

---

This paper presents **3 main contributions**, each analyzed against relevant prior work:

### Contribution 1: CounselBench-Eval: Large-scale expert evaluation dataset

**Description**: A benchmark dataset containing 2,000 expert evaluations from 100 mental health professionals rating responses from GPT-4, LLaMA-3, Gemini, and online human therapists across six clinically grounded dimensions, with span-level annotations and written rationales for each evaluation.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

#### 1. Proknow: Process knowledge for safety constrained and explainable question generation for mental health diagnostic assistance

**URL**: View paper

**Brief Assessment**

Proknow Safety[11] focuses on question generation for diagnostic assistance using process knowledge from questionnaires (PHQ-9, GAD-7), not on evaluating responses from multiple systems. The datasets serve fundamentally different purposes in the mental health AI pipeline.

---

#### 2. MHQA: A Diverse, Knowledge Intensive Mental Health Question Answering Challenge for Language Models

**URL**: View paper

**Brief Assessment**

MHQA[31] focuses on multiple-choice question answering derived from PubMed abstracts for mental health knowledge testing, not expert evaluation of open-ended responses from LLMs and therapists across clinically grounded dimensions with span-level annotations.

---

#### 3. MentalChat16K: A Benchmark Dataset for Conversational Mental Health Assistance

**URL**: View paper

**Brief Assessment**

MentalChat16K[29] focuses on providing conversational mental health counseling data (synthetic and anonymized transcripts) rather than expert evaluations of model responses. The candidate does not contain expert ratings, span-level annotations, or written rationales from mental health professionals evaluating LLM outputs.

---

#### 4. CounselBench: A Large-Scale Expert Evaluation and Adversarial Benchmark of Large Language Models in Mental Health Counseling

**URL**: View paper

**Brief Assessment**

CounselBench Adversarial[2] is the same paper as the original - it contains CounselBench-Eval as its first component. The candidate is not prior work but rather the published version of the original submission.

---

#### 5. A layered multi-expert framework for long-context mental health assessments

**URL**: View paper

**Brief Assessment**

Multi Expert Mental[28] focuses on a multi-model reasoning framework for depression screening using DAIC-WOZ interviews and case studies, not on creating expert evaluation datasets for mental health QA systems. The candidate does not present a comparable benchmark with expert ratings of LLM responses.

---

#### 6. Moving beyond medical exam questions: A clinician-annotated dataset of real-world tasks and ambiguity in mental healthcare

**URL**: View paper

**Brief Assessment**

Real World Mental Tasks[5] focuses on creating a clinician-annotated dataset for evaluating decision-making tasks across five clinical domains (diagnosis, treatment, monitoring, triage, documentation) in mental healthcare, rather than evaluating LLM responses to patient questions with expert ratings and rationales as in the original paper.

---

#### 7. AraHealthQA 2025: The First Shared Task on Arabic Health Question Answering

**URL**: View paper

**Brief Assessment**

AraHealthQA[26] focuses on Arabic health question answering shared tasks with multiple tracks (MentalQA and MedArabIQ), not on creating a large-scale expert evaluation dataset with span-level annotations and written rationales for mental health QA systems.

---

#### 8. Chatcounselor: A large language models for mental health support

**URL**: View paper

**Brief Assessment**

Chatcounselor[25] focuses on a different evaluation approach using automated GPT-4 assessment with 7 counseling metrics on 229 questions, rather than large-scale human expert evaluations with span-level annotations and written rationales from 100 mental health professionals on 2,000 evaluations.

---

#### 9. CBT-LLM: A Chinese Large Language Model for Cognitive Behavioral Therapy-based Mental Health Question Answering

**URL**: View paper

**Brief Assessment**

CBT LLM[27] focuses on creating a CBT-based QA dataset for training language models in Chinese, not on building an expert evaluation benchmark with professional annotations across multiple dimensions and models.

---

#### 10. Sindbad at arahealthqa track 1: Leveraging large language models for mental health q&a

**URL**: View paper

**Brief Assessment**

Sindbad AraHealthQA[30] focuses on Arabic mental health QA classification tasks using a dataset of 350 annotated instances, not on large-scale expert evaluation of LLM responses. The candidate does not address expert evaluation datasets or multi-dimensional clinical assessment of model outputs.

---

### Contribution 2: CounselBench-Adv: Adversarial benchmark for failure mode detection

**Description**: An adversarial dataset of 120 mental health questions authored by 10 clinicians to deliberately trigger specific model failure modes identified in CounselBench-Eval, paired with 1,080 expert-annotated responses from nine LLMs to enable targeted probing of model vulnerabilities.

This contribution was assessed against **7 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

#### 1. Detecting Algorithmic Errors and Patient Harms for AI-Enabled Medical Devices in Randomized Controlled Trials

**URL**: View paper

**Brief Assessment**

Algorithmic Errors Detection[42] focuses on detecting algorithmic errors and patient harms in AI-enabled medical devices within randomized controlled trials, not on adversarial benchmarking for mental health dialogue systems. The candidate addresses safety monitoring in clinical trial contexts rather than proactive adversarial testing of LLM failure modes in mental health QA.

---

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

**URL**: View paper

**Brief Assessment**

Tripartite Clinical Benchmarks[41] proposes a general evaluation framework for mental health LLM tools but does not present an adversarial dataset of clinician-authored questions designed to trigger specific failure modes, which is the core novelty of CounselBench-Adv.

---

#### 3. Adversarial Evaluation Algorithm for Detecting Extreme Behaviors of LLMs in Psychological Counseling Scenarios

**URL**: View paper

**Brief Assessment**

Adversarial Psychological Evaluation[38] focuses on detecting extreme behaviors (e.g., suicide risk) in psychological counseling through cross-variation testing, while CounselBench-Adv targets specific model failure modes (e.g., unauthorized medical advice, overgeneralization) identified through expert evaluation. The datasets serve different purposes: behavior detection versus failure mode elicitation.

---

#### 4. CounselBench: A Large-Scale Expert Evaluation and Adversarial Benchmark of Large Language Models in Mental Health Counseling

**URL**: View paper

**Brief Assessment**

CounselBench Adversarial[2] is the same paper as the original - it contains CounselBench-Adv as its second component. The candidate is not prior work but rather the published version of the original submission.

---

#### 5. 70152 Research Tutorial-Report

**URL**: View paper

**Brief Assessment**

Research Tutorial Report[43] focuses on BERT's pre-training methodology for language understanding tasks, not on adversarial benchmarking for mental health dialogue systems or failure mode detection in clinical AI applications.

---

#### 6. Logged, listened, and legally ignored: the mirage of privacy in AI therapy| Registrato, ascoltato e legalmente ignorato: il miraggio della privacy nella terapia dell'IA

**URL**: View paper

**Brief Assessment**

AI Therapy Privacy[40] focuses on privacy vulnerabilities (membership inference attacks) in AI mental health applications, not on adversarial benchmarking for detecting model failure modes in mental health dialogue systems. The candidate does not address adversarial dataset construction or expert-annotated failure mode detection.

---

### 7. Generative AI in Mental Well-Being: Balancing Pros and Cons

**URL**: View paper

**Brief Assessment**

Generative AI Wellbeing[39] mentions 'multiround automatic red-teaming (mart)' for adversarial testing, but focuses on general AI safety in mental health contexts rather than the specific clinician-authored adversarial dataset with expert annotations for targeted failure mode detection described in the original paper.

---

## Contribution 3: Six clinically grounded evaluation dimensions for mental health QA

**Description**: A multi-dimensional evaluation rubric developed through clinical psychology literature and expert consultation, comprising six dimensions: overall quality, empathy, specificity, medical advice, factual consistency, and toxicity, designed to assess both quality and safety in mental health question answering.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

### 1. Proknow: Process knowledge for safety constrained and explainable question generation for mental health diagnostic assistance

**URL**: View paper

**Brief Assessment**

Proknow Safety[11] does not propose evaluation dimensions for assessing mental health QA responses. It focuses on generating safe, process-guided diagnostic questions using clinical questionnaires as constraints.

---

### 2. Building Trust in Mental Health Chatbots: Safety Metrics and LLM-Based Evaluation Tools

**URL**: View paper

**Brief Assessment**

Mental Health Chatbot Trust[20] focuses on five safety-oriented guideline questions for evaluating mental health chatbots, not a six-dimensional rubric for open-ended mental health question answering quality and safety assessment.

---

### 3. Evaluating chatbots in psychiatry: Rasch-based insights into clinical knowledge and reasoning

**URL**: View paper

**Brief Assessment**

Psychiatry Chatbot Evaluation[35] focuses on evaluating chatbot performance on multiple-choice psychiatry licensing exam questions using Rasch analysis, not on developing evaluation dimensions for open-ended mental health question answering with metrics like empathy, specificity, and safety.

---

### 4. A review of the explainability and safety of conversational agents for mental health to identify avenues for improvement

**URL**: View paper

**Brief Assessment**

Explainability Safety Review[36] is a review paper examining existing VMHAs (virtual mental health assistants) from the perspective of explainability and safety properties, not a benchmark for evaluating mental health QA systems with specific evaluation dimensions.

---

### 5. Trustworthy Medical Question Answering: An Evaluation-Centric Survey

**URL**: View paper

**Brief Assessment**

Medical QA Survey[34] presents a general medical QA evaluation framework covering six dimensions (factuality, robustness, fairness, safety, explainability, calibration) across all medical domains, whereas the original paper develops a specialized rubric specifically for mental health QA with different dimensions (overall quality, empathy, specificity, medical advice, factual consistency, toxicity) grounded in clinical psychology literature and validated through expert consultation with 100 mental health professionals.

---

### 6. Do large language models align with core mental health counseling competencies?

**URL**: View paper

**Brief Assessment**

Counseling Competencies Alignment[32] focuses on evaluating LLMs against five core mental health counseling competencies derived from the NCMHCE licensing exam (intake/assessment/diagnosis, treatment planning, counseling skills/interventions, professional practice/ethics, core counseling attributes), not on developing evaluation dimensions for mental health question answering quality and safety as in the original paper.

---

### 7. Evaluating safety of large language models for patient-facing medical question answering

**URL**: View paper

**Brief Assessment**

Patient Facing Safety[33] focuses on patient-facing medical QA with dimensions like scientific consensus, inappropriate content, and missing content. The original paper's six dimensions (overall quality, empathy, specificity, medical advice, factual consistency, toxicity) are specifically designed for mental health counseling contexts, representing a distinct evaluation framework.

---

### 8. AraHealthQA 2025: The First Shared Task on Arabic Health Question Answering

**URL**: View paper

**Brief Assessment**

AraHealthQA[26] describes a shared task framework with evaluation datasets and standardized metrics, but does not present a multi-dimensional evaluation rubric comprising empathy, specificity, medical advice, factual consistency, and toxicity dimensions for mental health QA.

---

### 9. The interRAI suite of mental health assessment instruments: an integrated system for the continuum of care

**URL**: View paper

**Brief Assessment**

InterRAI Mental Health[37] focuses on comprehensive clinical assessment instruments for mental health service settings (inpatient psychiatry, community mental health, etc.), not on evaluation dimensions for question-answering systems or conversational AI quality assessment.

---

### 10. Incorporating evidence into mental health Q&A: a novel method to use generative language models for validated clinical content extraction

**URL**: View paper

**Brief Assessment**

Evidence Mental Health QA[3] focuses on evaluating LLM responses for evidence-based content extraction from clinical guidelines (reliability, clarity, completeness, traceability), not on the six-dimensional rubric (overall quality, empathy, specificity, medical advice, factual consistency, toxicity) developed for open-ended mental health question answering.

---

## Appendix: Text Similarity Detection

Textual similarity detection checked 24 papers and found 3 similarity segment(s) across 1 paper(s).

The following **1 paper(s)** were detected to have high textual similarity with the original paper. These may represent different versions of the same work, duplicate submissions, or papers with substantial textual overlap. Readers are advised to verify these relationships independently.

### 1. CounselBench: A Large-Scale Expert Evaluation and Adversarial Benchmark of Large Language Models in Mental Health Counseling

**Detected in**: Core Task (sibling), Contribution: contribution_1, Contribution: contribution_2

⚠ **Note**: This paper shows substantial textual similarity with the original paper. It may be a different version, a duplicate submission, or contain significant overlapping content. Please review carefully to determine the nature of the relationship.

## References


- [0] CounselBench: A Large-Scale Expert Evaluation and Adversarial Benchmarking of Large Language Models in Mental Health Question Answering View paper
- [1] Evaluation of alignment between large language models and expert clinicians in suicide risk assessment View paper
- [2] CounselBench: A Large-Scale Expert Evaluation and Adversarial Benchmark of Large Language Models in Mental Health Counseling View paper
- [3] Incorporating evidence into mental health Q&A: a novel method to use generative language models for validated clinical content extraction View paper
- [4] Could artificial intelligence write mental health nursing care plans? View paper
- [5] Moving beyond medical exam questions: A clinician-annotated dataset of real-world tasks and ambiguity in mental healthcare View paper
- [6] Expert and Interdisciplinary Analysis of AI-Driven Chatbots for Mental Health Support: Mixed Methods Study View paper
- [7] Exploring the efficacy of large language models in summarizing mental health counseling sessions: benchmark study View paper
- [8] From classification to clinical insights: Towards analyzing and reasoning about mobile and behavioral health data with large language models View paper
- [9] Embracing the uncertainty in humanâ□□machine collaboration to support clinical decision-making for mental health conditions View paper
- [10] Learning to automate follow-up question generation using process knowledge for depression triage on reddit posts View paper
- [11] Proknow: Process knowledge for safety constrained and explainable question generation for mental health diagnostic assistance View paper
- [12] Strengthening a mental illness management questionnaire for clinical associates through expert validation and cognitive interviews View paper
- [13] Evaluating the Clinical Safety of LLMs in Response to High-Risk Mental Health Disclosures View paper
- [14] â□□I don't know what you mean byI am anxious'â□□: a new method for evaluating conversational agent responses to standardized mental health inputs for anxiety and â□¦ View paper
- [15] Design and Challenges of Mental Health Assessment Tools Based on Natural Language Interaction View paper
- [16] MedExpert: An Expert-Annotated Dataset for Medical Chatbot Evaluation View paper
- [17] MindWellQA: A Semantically Enriched Evidence-Based QA System for Psychological Disorders View paper
- [18] Combining Artificial Users and Psychotherapist Assessment to Evaluate Large Language Model-based Mental Health Chatbots View paper
- [19] A Voice-Enabled Virtual Patient System for Interactive Training in Standardized Clinical Assessment View paper
- [20] Building Trust in Mental Health Chatbots: Safety Metrics and LLM-Based Evaluation Tools View paper
- [21] MTL-DQA: Multi-Task Learning with Psychological Indicators for Enhanced Quality in Depression Community QA System View paper
- [22] The Limbic Layer: Transforming Large Language Models (LLMs) into Clinical Mental Health Experts View paper
- [23] Using research evidence in mental health: userâ□□rating and focus group study of clinicians' preferences for a new clinical questionâ□□answering service View paper
- [24] A Dual-Prompting for Interpretable Mental Health Language Models View paper
- [25] Chatcounselor: A large language models for mental health support View paper
- [26] AraHealthQA 2025: The First Shared Task on Arabic Health Question Answering View paper
- [27] CBT-LLM: A Chinese Large Language Model for Cognitive Behavioral Therapy-based Mental Health Question Answering View paper
- [28] A layered multi-expert framework for long-context mental health assessments View paper
- [29] MentalChat16K: A Benchmark Dataset for Conversational Mental Health Assistance View paper
- [30] Sindbad at arahealthqa track 1: Leveraging large language models for mental health q&a View paper
- [31] MHQA: A Diverse, Knowledge Intensive Mental Health Question Answering Challenge for Language Models View paper
- [32] Do large language models align with core mental health counseling competencies? View paper
- [33] Evaluating safety of large language models for patient-facing medical question answering View paper
- [34] Trustworthy Medical Question Answering: An Evaluation-Centric Survey View paper
- [35] Evaluating chatbots in psychiatry: Rasch-based insights into clinical knowledge and reasoning View paper
- [36] A review of the explainability and safety of conversational agents for mental health to identify avenues for improvement View paper
- [37] The interRAI suite of mental health assessment instruments: an integrated system for the continuum of care View paper
- [38] Adversarial Evaluation Algorithm for Detecting Extreme Behaviors of LLMs in Psychological Counseling Scenarios View paper

- [39] Generative AI in Mental Well-Being: Balancing Pros and Cons View paper
- [40] Logged, listened, and legally ignored: the mirage of privacy in AI therapy| Registrato, ascoltato e legalmente ignorato: il miraggio della privacy nella terapia dell'IA View paper
- [41] Contextualizing Clinical Benchmarks: A Tripartite Approach to Evaluating LLM-Based Tools in Mental Health Settings View paper
- [42] Detecting Algorithmic Errors and Patient Harms for AI-Enabled Medical Devices in Randomized Controlled Trials View paper
- [43] 70152 Research Tutorial-Report View paper