

Enhancing Diagnostic Accuracy in Mental Health: : A Comparative Study of Human Therapists and AI-Assisted Assessment

SI: Ibrahim Bukhari | Yixin Ding | Mina Lu | Yutong Pan | Aneri Shah | Poorvi Sriraman | Victoria St. John | PI: Yiyang Xiong, PhD

INTRO

Study is to compare the diagnostic accuracy of licensed human therapists against three commercially available AI chatbots—ChatGPT, Meta LLaMA, and Claude Opus—when evaluating standardized fictional mental health case scenarios.

Research Question: How does the diagnostic accuracy of AI chatbots compare to that of licensed human therapists when evaluating standardized case scenarios for five common mental health conditions (Anxiety Disorders, Major Depressive Disorder, PTSD, Bipolar Disorder, Borderline Personality Disorder).

METHODS

- Phase 1 — Scenario Validation:** 3 expert consultants (≥5 yrs post-licensure) established consensus diagnoses for 5 fictional cases. Each case included intake forms, mental status exam, interview notes, and symptom scales (PHQ-9, GAD-7, C-SSRS).
- Phase 2 — Diagnostic Evaluation:** 80 licensed therapists (LPC, LCSW, LMFT, PhD/PsyD, MD/DO) completed a one-time Qualtrics survey (~30–60 min) with DSM-5-TR diagnoses + 0–10 confidence ratings. Licensure verified prior to enrollment.
- AI Evaluation:** ChatGPT, Meta LLaMA & Claude Opus — 10 independent runs per vignette under default parameters using a standardized prompt. No fine-tuning or iterative prompting.

RESULTS

 **Data Collection In Progress**

Enrollment is currently underway. Quantitative findings will be reported upon completion of data collection and analysis.

Measures & Analysis:

- Diagnostic accuracy (binary match with expert consensus)
- Confidence & case complexity (0–10 scales)
- Primary: Mixed-effects logistic regression (random intercepts for therapist & vignette)
- Secondary: Disorder-type subgroups, between-AI comparisons, differential pattern analyses

DISCUSSION

Depending on findings, results could inform responsible AI integration in mental health practice—as decision-support, not replacement—and reveal disorder types requiring heightened human oversight.

Implications for counselor training, ethical AI guidelines, and equity in mental health assessment.

Future work: AI–human collaborative diagnostic models; clinician attitudes toward AI-assisted assessment tools.



Do AI systems diagnose mental health conditions as accurately as licensed therapists?

Study in Progress — Five hypotheses being tested across ChatGPT, Meta LLaMA & Claude Opus against 80 licensed clinicians and expert consensus diagnoses.

Hypotheses

H1 Therapists will show higher overall accuracy than AI systems

H2 AI accuracy will vary by disorder (lower for overlapping profiles)

H3 AI responses will vary across 10 runs, especially for complex cases

H4 ChatGPT, LLaMA & Claude will differ significantly in accuracy

H5 Therapist confidence will correlate more strongly with accuracy

education.jhu.edu

ibukhar2@jh.edu

LinkedIn: [ibrahimbukhari0](#)