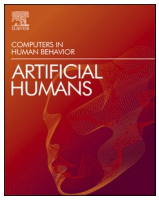
Computers in Human Behavior: Artificial Humans 6 (2025) 100212

Contents lists available at ScienceDirect

Computers in Human Behavior: Artificial Humans

journal homepage: www.journals.elsevier.com/computers-in-human-behavior-artificial-humans



Harnessing large language models for identification and treatment of obsessive-compulsive disorder

Inbar Levkovich *

Tel-Hai Academic College, Upper Galilee, 9977, Kiryat Shmona, 1220800, Israel

ARTICLE INFO

Keywords:
Artificial intelligence
ChatGPT
Large language models
Obsessive-compulsive disorder
Vignette study
Mental health professionals

ABSTRACT

Obsessive-Compulsive Disorder (OCD) is a mental health condition characterized by recurrent intrusive thoughts and repetitive behaviors, causing significant distress and disruption to daily life. Early identification and intervention are crucial for improving outcomes. This research compared the performance of four AI models (ChatGPT-3.5, ChatGPT-4, Claude 3.5 Sonnet, and Gemini 1.5 Pro) to that of human mental health professionals in recognizing OCD, recommending evidence-based therapies, and assessing stigma attributions. Using six vignettes, the study analyzed 480 AI evaluations and 315 assessments by mental health professionals. The AI models demonstrated superior performance in OCD recognition (100 % vs. 87 % accuracy for humans) and in recommending evidence-based interventions (60–100 % vs. 61.9 % for humans). Additionally, the AI models showed lower stigma and danger estimations. These findings indicating that AI demonstrated higher accuracy in vignette-based recognition tasks highlight the potential for AI integration into mental health care, particularly for enhancing OCD diagnosis and treatment.

1. Introduction

Obsessive-compulsive disorder (OCD) is a chronic and debilitating mental health condition that significantly impairs an individual's functioning and well-being (Cervin, 2023; Fineberg et al., 2019; Horwath & Weissman, 2022). The disorder is characterized by intrusive, repetitive thoughts (obsessions) and compulsive behaviors or mental acts aimed at reducing distress (Barnhill, 2022; Wairauch et al., 2024). This disorder is widely recognized as persistent and difficult to treat, with many patients experiencing lifelong symptoms if left untreated (Melkonian et al., 2022; Shams et al., 2024; Sharma et al., 2021). Despite the availability of effective evidence-based treatments, such as exposure and response prevention (ERP) and serotonin reuptake inhibitors (SRIs), individuals with OCD face persistent barriers to timely diagnosis and care (Fineberg et al., 2019; Font enelle et al., 2022; Liu et al., 2021). Studies have consistently demonstrated a striking delay averaging 12–17 years between the onset of symptoms and the initiation of appropriate treatment (Eisen et al., 2013; Perris et al., 2021; Pinto et al., 2006).

One major contributor to this delay is the misdiagnosis of OCD or the failure to recognize OCD symptoms. Patients with intrusive thoughts, particularly those with ego-dystonic obsessions of a taboo nature (e.g., sexual, religious, or aggressive content), are frequently misunderstood (Allely & Pickard, 2024; Bruce et al., 2018). For example, individuals who present to emergency departments with suicidal obsessions may be hospitalized as if they were at imminent risk, despite the fact that their thoughts are ego-dystonic and inconsistent with their intent (Bruce et al., 2018). Clinicians often conflate obsessions with excessive worries, leading to misdiagnosis as generalized anxiety disorder (GAD) or depression (Torres et al., 2007; Lafleur et al., 2017; Leichsenring et al., 2019).

Prior research has demonstrated that clinicians often have difficulty accurately diagnosing OCD, with misidentification rates varying depending on symptom presentation. Glazier et al. (2015) found that primary care physicians misdiagnosed half of OCD vignettes, with particularly high error rates for sexual (70.8 %–84.6 %) and aggressive (80.0 %) obsessions. Glazier et al. (2013) similarly reported high misidentification of atypical symptoms (up to 75 %) compared to contamination obsessions (15.8 %). Perez et al. (2022) showed that mental health providers misdiagnosed taboo thoughts (34.7 %–52.7 %) more often than contamination (11 %) or symmetry (6.9 %). Collectively, these findings highlight systematic blind spots, especially in the context of taboo or atypical presentations.

Recognition is further complicated by the diversity of OCD symptom

* Corresponding author.
E-mail address: levkovinb@telhai.ac.il.

https://doi.org/10.1016/j.chbah.2025.100212
Received 28 September 2024; Received in revised form 29 September 2025; Accepted 1 October 2025
Available online 2 October 2025

presentation (Shavitt et al., 2022). While common subtypes include contamination, symmetry/incompleteness, responsibility for harm, and taboo intrusive thoughts (Abramowitz et al., 2010; Canavan, 2024; Jacoby et al., 2018), many additional forms exist, among them somatic obsessions and "not just right" experiences. Empirical studies underscore this challenge: nearly 40 % of mental health professionals (Glazier et al., 2013) and over 50 % of primary care physicians (Glazier et al., 2015) failed to identify OCD correctly when presented with validated vignettes, particularly those involving taboo themes (Glazier & McGinn, 2015; Perez et al., 2022). Beyond the medical system, community gatekeepers such as clergy also misidentify OCD at high rates, particularly taboo-themed presentations, and frequently recommend interventions inconsistent with evidence-based care (Gouniai, Smith, & Glazier Leonte, 2022). Collectively, these findings highlight systemic barriers to human recognition of OCD, barriers that are rooted in stigma, symptom heterogeneity, and limited awareness of the full spectrum of the disorder (Dell'Osso et al., 2015; Thornicroft et al., 2019).

In view of these persistent diagnostic challenges, artificial intelligence (AI)—specifically large language models (LLMs)—may provide valuable complementary tools. LLMs, a subset of natural language processing systems, are trained on vast corpora of text to generate sophisticated predictions and responses (Zhang et al., 2024). Their design allows them to process nuanced linguistic input, potentially facilitating more consistent identification of psychopathological features than human judgment alone (Kim et al., 2024). AI models have already demonstrated promise in health-related applications, for example by streamlining administrative processes, increasing patient accessibility, and reducing stigma (Levkovich & Elyoseph, 2023a; Levkovich & Elyoseph, 2023b; Borna, 2024; Lee, 2023). In the specific case of OCD, patients often struggle to disclose intrusive or taboo obsessions due to shame or fear of judgment (Perris et al., 2021). AI-based tools may offer an accessible and less stigmatizing environment for initial recognition and triage (Borna et al., 2024; Thornicroft et al., 2019). Moreover, by applying consistent diagnostic heuristics, LLMs may reduce some of the human biases that often contribute to the misclassification of obsessions as worries or as spiritual struggles (Bruce et al., 2018; Glazier et al., 2013, 2015; Kim et al., 2024). Yet AI cannot and should not replace skilled clinicians; rather, it can serve as a supplementary tool to support recognition, referral, and psychoeducation (Motlagh et al., 2023; Rudolph et al., 2023; Gómez-Cabello, 2024).

Despite the rapid growth of AI applications across healthcare domains, research on their role in diagnosing OCD remains limited and has had mixed results. Recent work suggests that LLMs, such as ChatGPT-4 and Gemini Pro, can outperform clinicians in certain psychiatric diagnostic tasks, including the recognition of OCD symptom presentations, while providing consistent reasoning patterns (Motlagh et al., 2023; Rudolph et al., 2023). Simultaneously, content moderation policies and training biases have constrained AI's ability to accurately process sensitive or taboo obsessions, with some models refusing to engage with sexual or aggressive material (Bruce et al., 2018; Glazier & McGinn, 2015). Methodological inconsistencies across studies and ethical concerns regarding privacy, safety, and cultural bias also underscore the need for caution (Gomez-Cabello et al., 2024; Tal, 2023; Haber, 2024; Tornero-Costa, 2023).

Systematic reviews highlight the potential of LLMs in mental health for providing consistent diagnostic performance and treatment support while stressing the need for ethical safeguards (Levkovich & Omar, 2025, Omar & Levkovich, 2025). Evidence shows that although LLMs such as ChatGPT-4 may surpass professionals in treating conditions such as depression and PTSD, their accuracy is lower in complex cases such as early schizophrenia and their recovery predictions are more conservative (Levkovich, 2025). Recent findings further demonstrate that LLMs outperform clinicians in identifying OCD using standardized vignettes, with ChatGPT-4 achieving perfect accuracy, underscoring both their potential and limitations (Kim et al., 2024).

Research on OCD-related misdiagnosis among mental health professionals other than psychiatrists further highlights the urgent need for reliable diagnostic tools. Psychologists, social workers, and primary care providers frequently misidentify OCD, particularly when it presents in less stereotypical or taboo-related forms and sometimes recommend interventions that are not evidence-based (Glazier et al., 2013, 2015; Pérez et al., 2022). These findings resonate with broader evidence that even trained professionals may overlook OCD when it manifests outside conventional symptom clusters, resulting in significant delays in treatment. Thus, while LLMs hold promise for addressing these systemic barriers, their effectiveness still must be rigorously evaluated in comparison with human clinicians.

1.1. Current study

The current study evaluated the diagnostic accuracy, treatment recommendations, and stigma attributions of four widely used large language models (LLMs): ChatGPT-3.5 (OpenAI, 2023a), ChatGPT-4 (OpenAI, 2023b), Claude (Claude 3.5 Sonnet), and Gemini (Gemini 1.5 Pro), in comparison with those of mental health professionals. The models were tested across multiple languages using standardized vignettes to examine whether LLMs can overcome persistent challenges in OCD recognition and whether their outputs are aligned with evidence-based standards of care.

1.2. Research questions

This study examined the following research question:

RQ1: How do various AI tools (ChatGPT-3.5, ChatGPT-4, Claude, and Gemini) compare to the evaluations of human mental health professionals in accurately recognizing OCD?
RQ2: What are the recommended therapies for individuals diagnosed with OCD according to various AI tools (ChatGPT-3.5, ChatGPT-4, Claude, and Gemini) compared with the recommendations of human evaluators (mental health professionals)?
RQ3: How do various AI tools (ChatGPT-3.5, ChatGPT-4, Claude, and Gemini) compare to the evaluations of human mental health professionals in their assessments of stigma attributions and treatment willingness for OCD?

2. Method

2.1. LLMs procedure

This study was approved by the Ethical Review Board of the XXX Ethics Committee (No. XXX) and was conducted in accordance with the Declaration of Helsinki. Four large language models (ChatGPT-3.5, ChatGPT-4, Claude 3.5 Sonnet, and Gemini 1.5 Pro) were evaluated in March 2024. These models were selected because they were the most widely accessible platforms at the time of data collection and represented distinct design philosophies and stages of technological development: ChatGPT-3.5-turbo served as the freely available baseline; GPT-4 was ChatGPT's premium version with advanced reasoning capabilities; Claude 3.5 Sonnet, Anthropic's Constitutional AI, emphasized safety; and Gemini 1.5 Pro was Google's multimodal system. Each vignette was entered into a new chat session (separate browser tab) to ensure that the models could not rely on the prior conversation history. In some cases, evaluations were conducted across different devices to minimize the influence of cached responses. To maintain independence between the assessments, no additional prompting or fine-tuning was applied; only the first response from each model was recorded. Other potential models were considered but excluded due to limited availability, licensing restrictions, or unstable performance during the study period.

2.2. Human participants

The present study also included a sample of 315 mental health professionals who met the following inclusion criteria: (1) certification as psychologists, social workers, or psychotherapists; (2) sufficient proficiency in English to respond to the vignettes; and (3) provision of written informed consent. Exclusion criteria were limited to incomplete questionnaire responses. Participants were recruited via professional networks and social media between March and June 2024.

The sample comprised 58 % psychologists, 32 % clinical social workers, and 10 % psychotherapists. The average age was 42.7 years (SD = 9.6, range 28–64). A slight majority of the respondents were women (62 % female, 38 % male). Most participants were married or in a long-term partnership (68 %), while the remainder were single (22 %) or divorced or widowed (10 %). Clinicians were employed in a variety of settings, including community mental health centers (40 %), hospitals (35 %), and private practice (25 %). They reported an average of 9.2 years of clinical experience (range 2–28 years). Each mental health professional completed four vignettes along with the accompanying assessment items, yielding 1260 vignette-level responses (315 participants × 4 vignettes each). This dataset was analyzed separately from the large language model (LLM) outputs. Fig. 1 provides an overview of the study design.

2.3. Input source: vignettes

Vignettes are widely used as a research tool that allows for the systematic examination of diagnostic reasoning, clinical decision-making, and professional attitudes in a controlled and standardized manner (Levkovich, 2025; Cvetkovic et al., 2025). Their primary advantage lies in their ability to present consistent clinical descriptions to different participants or systems, thereby enabling reliable comparisons of responses. In the present study, we employed six vignettes that had been previously developed and extensively validated in earlier studies (Glazier et al., 2015; Canavan, 2022). In Glazier et al. (2015), 97 % of physicians confirmed that the vignettes met the diagnostic criteria for OCD, providing strong validation.

These vignettes were originally drafted by OCD specialists, refined iteratively, and validated through expert review, pilot testing, and repeated use in clinician samples. They have since been applied in multiple studies, demonstrating the expected recognition patterns among clinicians. Some of these vignettes were later translated into Spanish for use in an additional study (Perez et al., 2022), although only the original English versions were used in the present analysis.

The six vignettes covered the following OCD presentations: four taboo-themed intrusive thoughts—pedophilia-themed obsessions (POCD), sexual orientation OCD (SO-OCD), aggressive obsessions, and scrupulosity (religious)—as well as contamination and symmetry. Each vignette was constructed in both male and female versions (e.g., John or Lorraine). Apart from gender, all demographic details were held constant across the vignettes to minimize potential content bias (Glazier et al., 2015).

Ten separate evaluations were conducted for each vignette within new tabs in ChatGPT-3.5, GPT-4, Claude, and Gemini, yielding 480 outputs. These repeated runs represented non-independent outputs from the same model. They were included to assess within-model consistency and did not serve as evaluations from distinct raters. Each vignette described an initial therapy session with a client presenting OCD symptoms that had persisted for five years and were causing marked distress and impairment. All vignettes were standardized to 125–155 words. The full set of vignettes is provided in Appendix 1. One example vignette is described below for illustration purposes:

> Lorraine, a middle-aged woman, constantly worries about dirt and germs in her surroundings. She is unable to complete many of her daily activities because she tries at all costs to avoid touching things

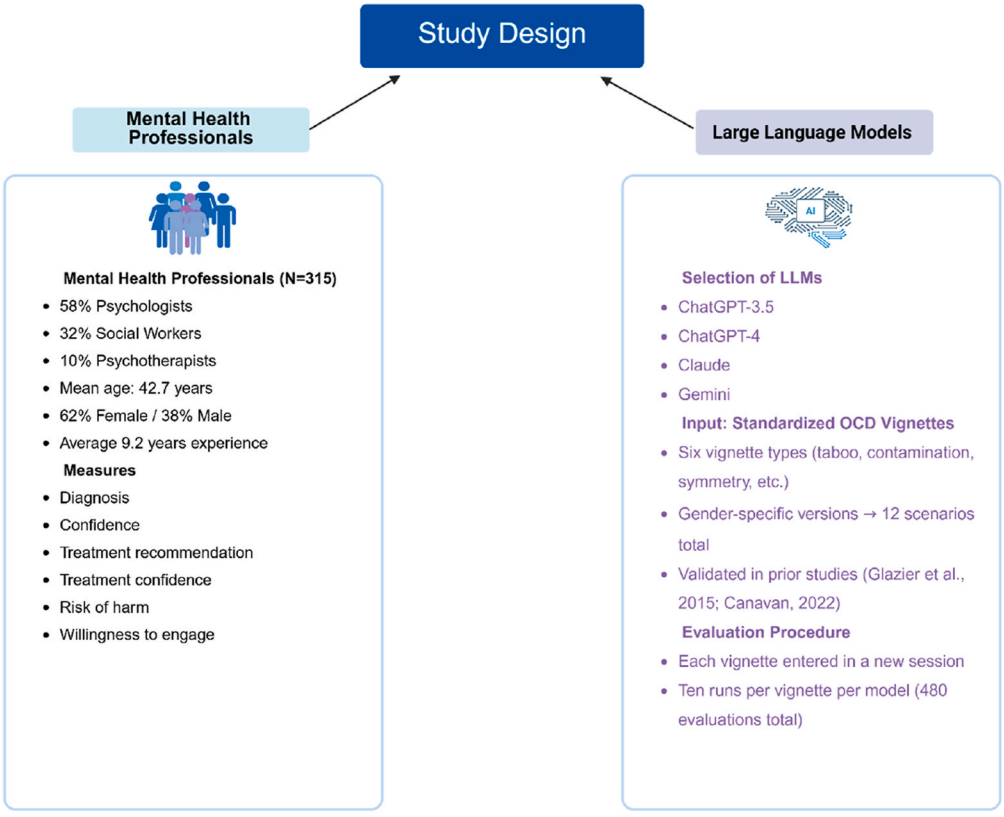


Fig. 1. Study design overview comparing Mental Health Professionals (MHPs) and Large Language Models (LLMs).
Note: MHPs = Mental Health Professionals; LLMs = Large Language Models.

she thinks may be dirty. If Lorraine does touch a "dirty" object, she will immediately wash her hands to avoid contracting a disease. She knows that these thoughts are excessive in nature and come from within her own mind. Yet despite knowing this, she remains upset by these thoughts and is not able to stop them.

2.4. Measures

To ensure consistency across conditions, the LLMs were presented with the same vignette texts and item wording used in the clinicians' study. The survey included six questions adapted from Canavan (2022). The first questions assessed diagnostic ability. The prompt asked: *"Please indicate the most likely diagnosis for this client (choose one of 21 options)."* This statement was followed by a confidence rating: *"How confident are you in your diagnostic judgment?"* Respondents answered on a five-point Likert scale (1 = not confident at all, 5 = very confident).

Next, the participants and the LLMs were asked to provide a primary treatment recommendation. The item read: *"Please select your primary treatment recommendation from the list below."* Options included ERP (exposure and response prevention), CBT (cognitive-behavioral therapy), SSRIs, supportive therapy, psychodynamic therapy, family therapy, and "other." Participants chose their responses from this structured list rather than entering free text. A second question asked: *"How confident are you in this treatment recommendation?"* The question was also answered on a five-point Likert scale. In line with Canavan (2022), evidence-based treatments (EBTs) were operationalized as ERP/EX-RP and SSRIs, and CBT was coded as an EBT only when explicitly described as incorporating ERP principles. All other options, including supportive therapy, psychodynamic therapy, family therapy, and responses classified as "other", were coded as non-EBTs.

Finally, stigma-related items adapted from Brown (2008) were included in the questionnaire. The first item assessed perceived risk: *"To what extent do you believe this client poses a risk of harm to others?"* The item was rated on a seven-point Likert scale (1 = no risk, 7 = very high risk). This item did not represent an objectively verifiable danger level but rather served as a proxy of perceived stigma toward clients with OCD. The second item measured willingness to engage: *"How willing would you be to engage professionally with this client?"* Participants answered on a seven-point Likert scale (1 = not willing at all, 7 = very willing).

For the LLM condition, each vignette and its six questions were entered verbatim and in the same order, with each run initiated in a new chat window to avoid carryover effects. The items assessing "confidence" and "willingness to engage" were retained because they replicate validated measures used in previous clinician vignette studies (Canavan, 2022; Brown, 2008). When applied to LLMs, however, these items do not reflect internal states but rather serve only as textual proxies simulating a clinical stance. Their inclusion enables comparability with prior human data while acknowledging that such responses reflect language production patterns rather than genuine clinician attitudes.

2.5. Statistical analysis

All analyses were conducted using R software version 4.4.0 (Core Team, 2023) and R-Studio version 2023.06.1 (RStudio Team, 2020). All tests were conducted using a two-tailed approach, with an initial significance threshold set at $\alpha = .05$. However, given the large number of chi-square analyses conducted to compare OCD recognition rates across multiple vignette types and between AI models and clinicians, we applied the Bonferroni correction to control for inflated Type I error. Therefore, the reported p-values reflect the adjusted significance level after correction.

The chi-square test of independence was conducted to analyze the rates of OCD identification relative to client gender. Furthermore, a chi-square analysis was used to evaluate the overall rates of OCD identification across the six distinct OCD presentations. The statistical methodology employed in this study was designed to provide a rigorous assessment of the effectiveness of various AI tools and human professionals in recognizing OCD. There were no statistically significant discrepancies in recognition rates based on client gender across any of the vignette subtypes. Consequently, the data for the male and female versions of each vignette type were combined for the remainder of the analysis, as reported by Canavan (2022).

3. Results

3.1. RQ1: Comparison of effectiveness of AI tools (ChatGPT-3.5, ChatGPT-4, Claude, and Gemini) and of mental health professionals in recognizing OCD

As defined in the Methods section, the recognition rate refers to whether OCD was correctly identified as the most likely diagnosis from a structured list of 21 options. To evaluate the differences in performance across these entities, we conducted chi-square tests to assess the statistical significance of the observed frequencies in each category. Given the multiple comparisons across entities, post-hoc pairwise chi-square tests with Bonferroni correction were employed to identify the specific pairs of entities that exhibited significantly different recognition rates.

Chi-square tests revealed significant differences in total recognition rates across entities ($\chi^2(4) = 1906.6$, $p < .001$). Post hoc pairwise comparisons with Bonferroni correction revealed significant differences between mental health professionals and all AI entities. Specifically, ChatGPT-3.5 ($\chi^2(1) = 421.61$, $p < .001$), ChatGPT-4 ($\chi^2(1) = 712.83$, $p < .001$), Claude ($\chi^2(1) = 712.83$, $p < .001$), and Gemini ($\chi^2(1) = 712.83$, $p < .001$) all exhibited significantly higher OCD recognition rates than the mental health professionals. These findings suggest that AI tools outperform human professionals in recognizing OCD. Table 1 and Fig. 2 illustrate the OCD recognition rates for the different entities.

The analyses were conducted independently for each vignette and revealed similar results.

Pedophilia Vignette: Chi-square tests showed significant differences in recognition rates across entities ($\chi^2(4) = 2930.5$, $p < .001$). Post hoc pairwise comparisons utilizing the Bonferroni correction revealed significant differences between mental health professionals and all AI entities. Specifically, ChatGPT-3.5 ($\chi^2(1) = 621.21$, $p < .001$), ChatGPT-4, Claude, and Gemini (all $\chi^2(1) = 936.08$, $p < .001$) all demonstrated significantly higher OCD recognition rates compared to mental health professionals.

Aggressive Vignette: Chi-square tests showed significant differences in recognition rates across entities ($\chi^2(4) = 3435.7$, $p < .001$). Post-hoc pairwise comparisons involving Bonferroni correction revealed significant differences between mental health professionals and all AI entities. ChatGPT-3.5, ChatGPT-4, Claude, and Gemini (all $\chi^2(1) = 1153.9$, $p < .001$) all exhibited significantly better OCD recognition rates than those of mental health professionals.

Religious Vignette: Chi-square tests showed significant differences in recognition rates across entities ($\chi^2(4) = 2423.7$, $p < .001$). Post-hoc pairwise comparisons with Bonferroni correction indicated significant differences between mental health professionals and all AI entities. ChatGPT-3.5 ($\chi^2(1) = 609.5$, $p < .001$), ChatGPT-4, Claude, and Gemini ($\chi^2(1) = 923.19$, $p < .001$) all demonstrated significantly higher OCD recognition rates than mental health professionals.

Homosexual Vignette: Chi-square tests revealed significant differences in recognition rates across entities ($\chi^2(4) = 2679.3$, $p < .001$). The post hoc pairwise comparisons, along with the Bonferroni correction, revealed significant differences between mental health professionals and all AI entities. Claude ($\chi^2(1) = 710.92$, $p < .001$), ChatGPT-3.5, ChatGPT-4, and Gemini (all $\chi^2(1) = 1034.2$, $p < .001$) demonstrated significantly higher OCD recognition rates than mental health professionals.

Contamination/Symmetry Vignette: Chi-square tests indicated

Table 1
Comparison of OCD recognition rates across different entities.

Vignette	Mental health professionals	ChatGPT-3.5	ChatGPT-4	Claude	Gemini
Total Recognition	47.3 %	90.0 %	100.0 %	100.0 %	100.0 %
Pedophilia	36.1 %	100.0 %	100.0 %	100.0 %	100.0 %
Aggressive	26.7 %	100.0 %	100.0 %	100.0 %	100.0 %
Religious	36.7 %	90.0 %	100.0 %	100.0 %	100.0 %
Homosexuality	31.7 %	100.0 %	100.0 %	90.0 %	100.0 %
Contamination/Symmetry	77.3 %	30.0 %	100.0 %	100.0 %	100.0 %

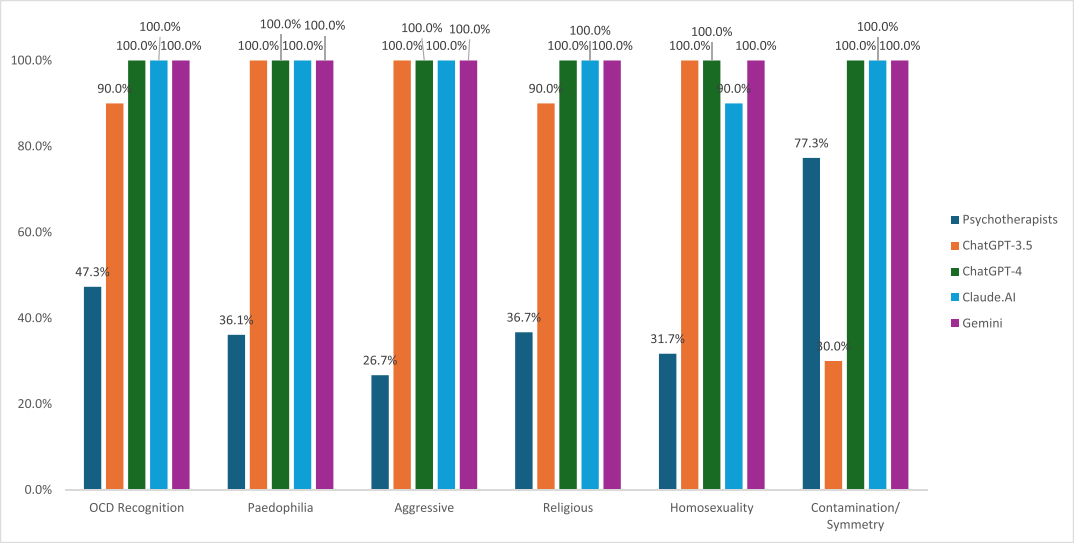


Fig. 2. Comparison of OCD recognition rates across different entities.

significant differences in recognition rates across entities ($\chi^2(4) = 2447.7$, $p < .001$). The results of post-hoc pairwise comparisons with Bonferroni correction demonstrated notable disparities between mental health professionals and all AI entities. ChatGPT-4, Claude, and Gemini (all $\chi^2(1) = 253.81$, $p < .001$) demonstrated significantly higher OCD recognition rates than did mental health professionals. ChatGPT-3.5 ($\chi^2(1) = 447.96$, $p < .001$) was the only AI that performed worse than the mental health professionals.

With the minor exception of ChatGPT-3.5's performance in the contamination/symmetry vignette, in all cases AI exhibited significantly better OCD recognition rates than mental health professionals, indicating that AI tools outperformed human professionals in recognizing OCD across various vignettes.

Moreover, when examining the differences between mental health professionals and AI entities regarding their confidence in identifying the correct psychological diagnosis based on the provided vignettes, chi-square tests revealed significant differences across the entities ($\chi^2(4) = 533.88$, $p < .001$). Post-hoc pairwise comparisons incorporating the Bonferroni correction revealed statistically significant differences between mental health professionals and all AI entities: ChatGPT-3.5, ChatGPT-4, Claude, and Gemini (all $\chi^2(1) = 136.91$, $p < .001$), with the AI entities displaying significantly higher confidence rates than the mental health professionals (100 % vs. 87 %).

3.2. RQ2: Comparison between mental health professionals and AIs in their treatment decisions (whether or not they recommended evidence-based intervention)

Based on the provided vignettes, the chi-square test results indicated a significant difference between the entities ($\chi^2(4) = 1568.4$, $p < .001$). Further analysis using post-hoc pairwise comparisons with Bonferroni correction revealed a significant difference between mental health professionals and all AI entities. ChatGPT-3.5 ($\chi^2(1) = 786.61$, $p < .001$), ChatGPT-4 ($\chi^2(1) = 486.75$, $p < .001$), Claude ($\chi^2(1) = 786.61$, $p < .001$), and Gemini ($\chi^2(1) = 54.513$, $p < .001$) all demonstrated significantly higher evidence-based treatment (EBT) recommendation rates than mental health professionals (Fig. 3).

Moreover, chi-square tests that compared the EBT recommendation rates between mental health professionals and AI entities only for those that recognized OCD pointed to significant differences across the entities ($\chi^2(4) = 1101.8$, $p < .001$). Additionally, post hoc pairwise comparisons with Bonferroni correction revealed significant differences between mental health professionals and most AI entities. ChatGPT-3.5 ($\chi^2(1) =$

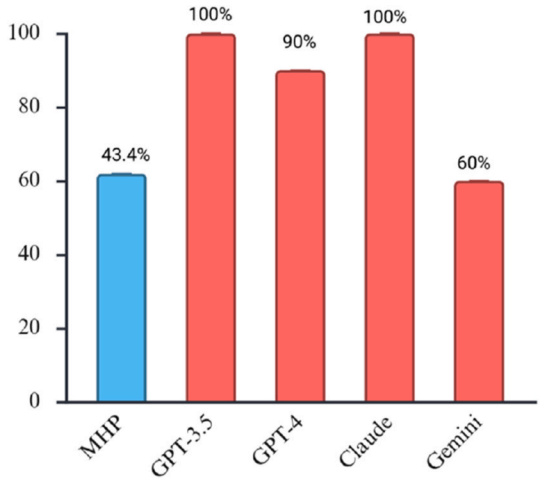


Fig. 3. *Comparison of EBT recommendation rates across different entities.*
Note: EBT = Evidence-Based Treatment; MHPs = Mental Health Professionals.

468.19, p < .001), ChatGPT-4 ($\chi^2(1) = 214.61$, p < .001), and Claude ($\chi^2(1) = 468.19$, p < .001) all demonstrated significantly higher EBT recommendation rates than mental health professionals, whereas Gemini ($\chi^2(1) = 0.68$, p = .41) showed rates similar to those of mental health professionals.

Chi-square tests examining differences between mental health professionals and AIs regarding confidence in their EBT recommendation pointed to significant differences across entities ($\chi^2(4) = 807.15$, p < .001). The results of post-hoc pairwise comparisons with Bonferroni correction pointed to significant differences between mental health professionals and all AI entities. ChatGPT-3.5 ($\chi^2(1) = 189.39$, p < .001), ChatGPT-4 ($\chi^2(1) = 436.59$, p < .001), Claude ($\chi^2(1) = 62.701$, p < .001), and Gemini ($\chi^2(1) = 436.59$, p < .001) all demonstrated significantly higher confidence rates than mental health professionals (Fig. 4).

3.3. RQ3: Comparison between AIs and mental health professionals in their stigma and danger estimations

Stigma estimations were based on two validated items (Brown, 2008) that assessed perceived risk of harm and willingness to engage professionally. As the danger item reflects perceived stigma rather than objective clinical risk, the observed differences should be interpreted as variations in stigma attribution, not as accuracy in risk assessment.

Chi-square tests indicated significant differences across entities ($\chi^2(4) = 673.97$, p < .001) in stigma and danger estimations. Post-hoc pairwise comparisons with Bonferroni correction also revealed significant differences between mental health professionals and all AI entities. ChatGPT-3.5, ChatGPT-4, Claude, and Gemini (all $\chi^2(1) = 175.29$, p < .001) all exhibited significantly lower estimations (all 0 %) of danger than mental health professionals (16.3 %).

Finally, in answer to the question regarding willingness to treat the person described in the vignettes, chi-square tests showed significant differences across the entities ($\chi^2(4) = 251.11$, p < .001). Post-hoc pairwise comparisons with Bonferroni correction revealed significant differences between mental health professionals and all AI entities. ChatGPT-3.5, ChatGPT-4, Claude, and Gemini (all $\chi^2(1) = 61.936$, p < .001) all demonstrated significantly higher willingness to treat the person than did the mental health professionals (100 % vs. 93.8 %).

4. Discussion

This study aimed to assess the ability of four advanced AI models to accurately recognize obsessive-compulsive disorder compared to human professionals. Additionally, the study assessed the recommended therapies and stigma attributions provided by these AI models. The study examined the effectiveness of AI tools in recognizing OCD. Significant differences were found between the mental health professionals and all AI models. All AI tools demonstrated significantly higher OCD recognition rates than mental health professionals.

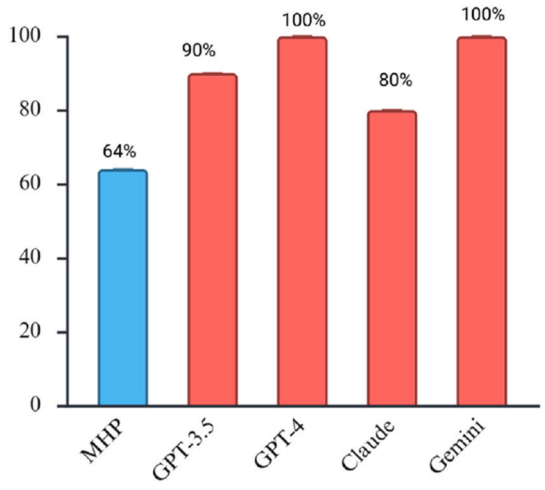


Fig. 4. Comparison of confidence in EBT recommendation across different entities
Note: EBT = Evidence-Based Treatment; MHPs = Mental Health Professionals.

These findings point to the potential advantages of AI tools in vignette-based recognition of OCD. These results, however, do not imply superiority in clinical practice, where patients often present with atypical symptoms and comorbid conditions that were not captured in our vignettes. Further research using real-world patient data is required. All the AI models also demonstrated higher confidence in their identification than the mental health professionals. Whereas the mental health professionals reported an 87 % confidence rate in their recognition, ChatGPT-3.5, Claude, and Gemini all reported a 100 % confidence rate.

These findings should be interpreted with caution, as confidence ratings in simulated vignettes may not directly reflect clinical accuracy in complex, real-world cases. Consistent with prior work, our findings suggest that clinicians frequently misidentify OCD presentations, especially when obsessions involve taboo content, compared to more stereotypical subtypes such as contamination or symmetry (Glazier et al., 2013, 2015; Perez et al., 2022). In contrast, the AI models in this study demonstrated higher diagnostic accuracy across vignette-based cases, echoing the results reported by Kim et al. (2024).

In the current study, the AI models demonstrated significantly higher evidence-based intervention recommendations than mental health professionals. Whereas mental health professionals reported confidence rates of 61.9 %, ChatGPT-3.5 and Claude reported 100 % confidence, ChatGPT-4 reported approximately 90 %, and Gemini reported 60 %. In comparison, confidence in evidence-based treatment recommendations among the AI models revealed levels of cultural intelligence ranging from 80 to 100 %, whereas mental health professionals reported 64 % confidence in evidence-based treatment. This finding supports previous research on AI languages in mental health care showing that AI exhibited better results than human samples (Levkovich & Elyoseph, 2023a, Levkovich & Elyoseph, 2023b, Elyoseph & Levkovich, 2024). Part of the delay between symptom onset and receipt of evidence-based interventions may be attributed to a lack of healthcare provider awareness of the diverse symptom presentations of OCD, particularly those related to taboo thoughts (Thornicroft et al., 2019; Leichsenring et al., 2019; Dell'Osso et al., 2015). A study examining the diagnostic impressions and treatment recommendations of primary care physicians revealed that those who misidentified OCD vignettes were less likely to recommend empirically supported treatment (Glazier et al., 2015).

In this study, we compared AI models with mental health professionals. Our findings revealed significant differences in stigma and danger estimations regarding the person described in the vignette. All the LLM models demonstrated significantly lower stigma and danger estimations than did the mental health professionals. When asked if they would be willing to treat the person described in the vignettes, both the AI models and the mental health professionals demonstrated high willingness (AI models 100 % and mental health professionals 93.8 %). Stigma among the general population can discourage people from seeking help, and stigma among mental health professionals can be even more detrimental (Chaves et al., 2022). The behavior and attitude of mental health providers toward their clients can significantly affect treatment outcomes (Lauber et al., 2006; Ponzini et al., 2022). Nevertheless, the present study and Canavan's (2022) study provide encouraging data showing low levels of stigma among mental health professionals. This finding contradicts several studies that reported higher stigma levels among mental health professionals. Despite the expectation that clinicians are trained to be free of bias, research indicates that they may hold negative views of individuals with mental illnesses, mirroring the general public's views (Kassam et al., 2010; Lauber et al., 2006).

Previous research has explored the impact of clinicians' attitudes and

biases on their professional practice. This research found that stigma regarding sexual obsessions is significantly greater than regarding contamination obsessions among adults (Cathey & Wetterneck, 2013). Additionally, Steinberg and Wetterneck (2017) noted that clinicians hold stronger stigmas toward clients with contamination, harm, and sexual obsessions than toward those with scrupulous obsessions, discouraging clients from disclosing these thoughts.

This study represents an initial effort to evaluate and contrast the utility of various AI models in recognizing and addressing OCD (Kim et al., 2025; Masoumi, 2025). The empirical findings demonstrate that AI models exhibited higher accuracy in this vignette-based study than did human participants. Nevertheless, this advantage requires further validation with real-world patients presenting atypical symptoms and comorbid conditions. Inherent biases or insufficient demographic diversity within these datasets may result in erroneous predictions (Omar et al., 2024). Furthermore, the inherently opaque nature of LLM algorithms often obfuscates the underlying mechanisms that drive their inferential processes. These variations have profound implications for the identification and reporting of OCD in LLMs. Acknowledging these disparities and their potential impact on clinical outcomes is paramount for the ongoing refinement and integration of LLMs into mental healthcare paradigms. These results underscore the importance of utilizing multiple LLM iterations in a complementary manner rather than relying exclusively on a single model (Stade et al., 2024).

While intelligent language models offer significant potential as adjunctive tools for training, support, and expert consultation, it is imperative to recognize their limitations in clinical practice. These technologies should not be construed as replacements for skilled clinicians who possess the capacity for holistic assessments and nuanced interpretations of complex clinical presentations. Previous studies have stressed that the opaque nature of LLMs and their inherent biases may obscure diagnostic processes and contribute to erroneous predictions (Omar et al., 2024; Stade et al., 2024; Tornero-Costa et al., 2023). Accordingly, the integration of LLMs into mental health practices necessitates a judicious approach that leverages their strengths while remaining cognizant of their constraints (Gómez-Cabello, 2024; Haber et al., 2024; Motlagh et al., 2023; Rudolph et al., 2023; Tal et al., 2023). Ultimately, LLMs should complement rather than replace human clinicians, whose expertise, empathy, and ability to provide holistic care remain irreplaceable (Kassam et al., 2010; Lauber et al., 2006; Thornicroft et al., 2019).

Note that the vignettes used in this study focused on common OCD subtypes and did not include patients with comorbid psychiatric disorders or more unusual idiosyncratic symptom presentations. In clinical practice, many individuals present with atypical or complex OCD, often alongside other conditions (Shavitt et al., 2022; Dell'Osso et al., 2015). Consequently, the present findings should be interpreted with caution and cannot be directly generalized to such cases (Leichsenring et al., 2019; Thornicroft et al., 2019). Future research should explore the performance of AI models when applied to real-world patient data, including comorbidities and atypical symptomatology.

4.1. Limitations

The current study has some limitations that warrant acknowledgment. First, this study was confined to specific iterations of LLMs at a particular point in time, without considering subsequent versions. Future investigations should address this limitation by examining updates and improvements in future versions of these AI models. Second, the data were compared with those of a representative sample of human professionals (mental health professionals). Although the sample was intended to be representative, it did not encompass the full spectrum of global psychotherapeutic practices. Therefore, the findings may not be generalizable.

Third, this study utilized case vignettes, which, although valuable methodologically, have the potential to oversimplify the intricacies of authentic clinical situations. Although these vignettes were validated and used in several articles, they do not encompass the full range of symptoms and background details present in actual patient cases. Hence, this methodology has inherent limitations, especially considering the multifaceted nature of OCD. Moreover, the vignettes reflected common OCD subtypes and did not include comorbidities or atypical symptom expressions. Since many patients present with complex and comorbid forms of OCD (Shavitt et al., 2022; Dell'Osso et al., 2015), the findings should be interpreted with caution and not generalized directly to such cases (Leichsenring et al., 2019; Thornicroft et al., 2019).

Another limitation is the possibility of data leakage. LLMs may have been exposed to vignette materials or similar descriptions during the pretraining phase and this exposure could partially explain the high recognition rates. Future research should replicate these findings using novel and unpublished vignettes. Additionally, note that the measures of "confidence" and "willingness" do not reflect genuine clinical attitudes, as LLMs lack the internal states that humans have. These items were retained because they replicated validated measures used in previous vignette studies with clinicians, allowing comparability across samples. In the context of LLMs, however, they serve only as textual proxies for clinical stance and should therefore be interpreted with caution. A further limitation concerns the repeated runs of each vignette through the LLMs, which represent non-independent outputs from the same models. While such repetitions were included to assess within-model consistency, they may underestimate variability relative to the fully independent responses of human clinicians. Finally, the "danger" item reflects perceived stigma rather than objective clinical risk, as the vignettes were not designed to depict verifiable levels of potential harm.

Finally, this study did not directly evaluate the clinical accuracy of large language models (LLMs) compared to human professionals. While the analysis provides significant insights into the competencies of AI models, future research should incorporate validation studies to determine the clinical applicability and dependability of LLM predictions in practical environments. Moreover, this investigation did not examine the possible biases within the LLMs.

5. Conclusion

This study evaluated the efficacy of four AI models (ChatGPT-3.5, ChatGPT-4, Claude, and Gemini) in recognizing OCD and recommending evidence-based treatments and compared their recognition and recommendations with those of human professionals. Utilizing six vignettes depicting clients with OCD, the study found that AI models significantly outperformed mental health professionals in terms of recognition rates, confidence levels, and recommendation of appropriate interventions. Additionally, the AI models exhibited lower stigma and danger estimates in OCD cases.

These results underscore the capacity of AI tools to improve OCD diagnosis and treatment in clinical settings. AI models demonstrated higher confidence and more consistent recognition of OCD across various subtypes, including those related to taboo-intrusive thoughts, which often pose a challenge for human professionals. The results indicate that AI tools have the potential to enhance the precision and promptness of OCD diagnosis and treatment in clinical practice.

Nevertheless, it is crucial to acknowledge that AI should complement, not replace, human professionals. While AI can provide diagnostic support and treatment recommendations, the nuanced understanding, empathy, and knowledge that trained psychologists offer are irreplaceable. Future research should explore the integration of AI into clinical workflows and address ethical, privacy, and cultural considerations to maximize AI's benefits in mental healthcare.

Ethics approval

This study was approved by the Oranim University Review Board (No. Applicable 1852023).

Consent for publication

All participants provided informed consent for publication of anonymized data.

Availability of data and materials

The data and materials used in this study are available upon request from the corresponding author.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The author did not receive support from any organization for the submitted work.

Abbreviations

OCD	Obsessive-Compulsive Disorder
LLM	Large Language Models
NLP	Natural Language Processing
AI	Artificial Intelligence
EBT	Evidence-based treatment

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.chbah.2025.100212.

References

Abramowitz, J. S., Deacon, B. J., Olatunji, B. O., Wheaton, M. G., Berman, N. C., Losardo, D., … Hale, L. R. (2010). Assessment of obsessive-compulsive symptom dimensions: Development and evaluation of the dimensional obsessive-compulsive scale. *Psychological Assessment, 22*(1), 180–198.

Allely, C. S., & Pickard, M. (2024). A systematic scoping review of the literature on sexual orientation obsessive compulsive disorder (SOOCD): Important clinical considerations and recommendations. *Psychiatry Research, 342*, Article 116198. https://doi.org/10.1016/j.psychres.2024.116198

Barnhill, J. (2022). Obsessive-compulsive and related disorders. In *Textbook of psychiatry for intellectual disability and autism spectrum disorder* (pp. 625–654). Springer.

Borna, S., Barry, B. A., Makarova, S., Parte, Y., Haider, C. R., Sehgal, A., Leibovich, B. C., & Forte, A. J. (2024). Artificial intelligence algorithms for expert identification in medical domains: A scoping review. *European Journal of Investigation in Health Psychology and Education, 14*(5), 1182–1196. https://doi.org/10.3390/ejihpe14050078

Brown, S. A. (2008). Factors and measurement of mental illness stigma: A psychometric examination of the attribution questionnaire. *Psychiatric Rehabilitation Journal, 32*(2), 89–94. https://doi.org/10.2975/32.2.2008.89.94

Bruce, S. L., Ching, T. H., & Williams, M. T. (2018). Pedophilia-themed obsessive–compulsive disorder: Assessment, differential diagnosis, and treatment with exposure and response prevention. *Archives of Sexual Behavior, 47*, 389–402. https://doi.org/10.1007/s10508-017-1031-4

Canavan, R. (2024). Recognition rates, treatment recommendations and stigma attributions for clients presenting with taboo intrusive thoughts: A vignette-based survey of psychotherapists. *Counselling and Psychotherapy Research, 24*(1), 51–62. https://doi.org/10.1002/capr.12557

Cathey, A. J., & Wetterneck, C. T. (2013). Stigma and disclosure of intrusive thoughts about sexual themes. *Journal of Obsessive-Compulsive and Related Disorders, 2*(4), 439–443. https://doi.org/10.1016/j.jocrd.2013.09.001

Cervin, M. (2023). Obsessive-compulsive disorder: Diagnosis, clinical features, nosology, and epidemiology. *Psychiatria Clinica, 46*(1), 1–16. https://doi.org/10.1016/j.psc.2022.10.006

Chaves, A., Arnáez, S., Castilla, D., Roncero, M., & García-Soriano, G. (2022). Enhancing mental health literacy in obsessive-compulsive disorder and reducing stigma via smartphone: A randomized controlled trial protocol. *Internet Interventions, 29*, Article 100560. https://doi.org/10.1016/j.invent.2022.100560

Cvetkovic, I., Grashoff, I., Jovancevic, A., & Bittner, E. (2025). Quid pro quo: Information disclosure for AI feedback in human-AI collaboration. *Computers in Human Behavior: Artificial Humans, 4*, Article 100137. https://doi.org/10.1016/j.chbah.2025.100137

Dell'Osso, B., Benatti, B., Oldani, L., Spagnolin, G., & Altamura, A. C. (2015). Differences in duration of untreated illness, duration, and severity of illness among clinical phenotypes of obsessive-compulsive disorder. *CNS Spectrums, 20*(5), 474–478. https://doi.org/10.1017/S1092852914000339

Eisen, J. L., Sibrava, N. J., Boisseau, C. L., Mancebo, M. C., Stout, R. L., Pinto, A., & Rasmussen, S. A. (2013). Five-year course of obsessive-compulsive disorder: Predictors of remission and relapse. *The Journal of Clinical Psychiatry, 74*(3), 7286. https://doi.org/10.4088/JCP.12m07657

Elyoseph, Z., & Levkovich, I. (2024). Comparing the perspectives of generative AI, mental health experts, and the general public on schizophrenia recovery: Case vignette study. *JMIR Mental Health, 11*(1), Article e53043.

Fineberg, N. A., Dell'Osso, B., Albert, U., Maina, G., Geller, D., Carmi, L., Sireau, N., Walitza, S., Grassi, G., & Pallanti, S. (2019). Early intervention for obsessive compulsive disorder: An expert consensus statement. *European Neuropsychopharmacology, 29*(4), 549–565. https://doi.org/10.1016/j.euroneuro.2019.02.002

Glazier, K., Calixte, R. M., Rothschild, R., & Pinto, A. (2013). High rates of OCD symptom misidentification by mental health professionals. *25*(3), 201–209.

Glazier, K., & McGinn, L. K. (2015). Non-contamination and non-symmetry OCD obsessions are commonly not recognized by clinical, counselling and school psychology doctoral students. *Journal of Depression & Anxiety, 4*(3), 10–4172. https://doi.org/10.4190/2167-1044.1000190

Glazier, K., Swing, M., & McGinn, L. K. (2015). Half of obsessive-compulsive disorder cases misdiagnosed: Vignette-based survey of primary care physicians. *The Journal of Clinical Psychiatry, 76*(6), 7995.

Gomez-Cabello, C. A., Borna, S., Pressman, S., Haider, S. A., Haider, C. R., & Forte, A. J. (2024). Artificial-intelligence-based clinical decision support systems in primary care: A scoping review of current clinical implementations. *European Journal of Investigation in Health Psychology and Education, 14*(3), 685–698. https://doi.org/10.3390/ejihpe14030045

Haber, Y., Hadar-Shoval, D., & Elyoseph, Z. (2024). The artificial third: A broad view of the effects of introducing generative artificial intelligence on psychotherapy. *JMIR Mental Health, 11*, Article e54781. https://doi.org/10.2196/54781

Horwath, E., & Weissman, M. M. (2022). The epidemiology and cross-national presentation of obsessive-compulsive disorder. *Obsessive-Compulsive Disorder and Tourette's Syndrome*, 35–49.

Jacoby, R. J., Blakey, S. M., Reuman, L., & Abramowitz, J. S. (2018). Mental contamination obsessions: An examination across the obsessive-compulsive symptom dimensions. *Journal of Obsessive-Compulsive and Related Disorders, 17*, 9–15. https://doi.org/10.1016/j.jocrd.2017.08.005

Kassam, A., Glozier, N., Leese, M., Henderson, C., & Thornicroft, G. (2010). Development and responsiveness of a scale to measure clinicians' attitudes to people with mental illness (medical student version). *Acta Psychiatrica Scandinavica, 122*(2), 153–161. https://doi.org/10.1111/j.1600-0447.2010.01562.x

Kim, J., Leonte, K. G., Chen, M. L., Torous, J. B., Linos, E., Pinto, A., & Rodriguez, C. I. (2024). Large language models outperform mental and medical health care professionals in identifying obsessive-compulsive disorder. *npj Digital Medicine, 7*(1), 193. https://doi.org/10.1038/s41746-024-01181-x

Kim, J., Pacheco, J. P. G., Golden, A., Aboujaoude, E., van Roessel, P., Gandhi, A., … Rodriguez, C. I. (2025). Artificial intelligence in obsessive-compulsive disorder: A systematic review. *Current Treatment Options in Psychiatry, 12*(1), 23. https://doi.org/10.1109/ACCESS.2025.3558845

Lauber, C., Nordt, C., Braunschweig, C., & Rössler, W. (2006). Do mental health professionals stigmatize their patients? *Acta Psychiatrica Scandinavica, 113*, 51–59. https://doi.org/10.1111/j.1600-0447.2005.00718.x

Lee, H. (2023). The rise of ChatGPT: Exploring its potential in medical education. *Anatomical Sciences Education*. https://doi.org/10.1002/ase.2270

Leichsenring, F., Sarrar, L., & Steinert, C. (2019). Drop-outs in psychotherapy: A change of perspective. *World Psychiatry, 18*(1), 32. https://doi.org/10.1002/wps.20588

Levkovich, I. (2025). Evaluating diagnostic accuracy and treatment efficacy in mental health: A comparative analysis of large language model tools and mental health professionals. *European Journal of Investigation in Health, Psychology and Education, 15*(1), 9.

Levkovich, I., & Elyoseph, Z. (2023a). Suicide risk assessments through the eyes of ChatGPT-3.5 versus ChatGPT-4: vignette study. *JMIR mental health, 10*, Article e51232.

Levkovich, I., & Elyoseph, Z. (2023b). Identifying depression and its determinants upon initiating treatment: ChatGPT versus primary care physicians. *Family Medicine and Community Health, 11*(4), Article e002391.

Levkovich, I., & Omar, M. (2025). Evaluating of bert-based and large language mod for suicide detection, prevention, and risk assessment: A systematic review. *Journal of Medical Systems, 48*(1), 113.

Liu, J., Cui, Y., Yu, L., Wen, F., Wang, F., Yan, J., Yan, C., & Li, Y. (2021). Long-term outcome of pediatric obsessive-compulsive disorder: A meta-analysis. *Journal of Child and Adolescent Psychopharmacology, 31*(2), 95–101. https://doi.org/10.1089/cap.2020.0051

Masoumi, S. J. (2025). Applications and efficacy of artificial intelligence in obsessive-compulsive disorder: A narrative review. *Journal of Nursing Reports in Clinical Practice, 3*(6), 601–608. https://doi.org/10.32598/JNRCP.2502.1230

Melkonian, M., McDonald, S., Scott, A., Karin, E., Dear, B. F., & Wootton, B. M. (2022). Symptom improvement and remission in untreated adults seeking treatment for obsessive-compulsive disorder: A systematic review and meta-analysis. *Journal of Affective Disorders, 318*, 175–184. https://doi.org/10.1016/j.jad.2022.08.037

Motlagh, N. Y., Khajavi, M., Sharifi, A., & Ahmadi, M. (2023). The impact of artificial intelligence on the evolution of digital education: A comparative study of OpenAI

text generation tools including ChatGPT, bing chat, bard, and ernie. *arXiv Preprint arXiv:2309.02029*. https://doi.org/10.48550/arXiv.2309.02029

Omar, M., & Levkovich, I. (2025). Exploring the efficacy and potential of large language models for depression: A systematic review. *Journal of Affective Disorders, 371*, 234–244.

Omar, M., Soffer, S., Charney, A. W., Landi, I., Nadkarni, G. N., & Klang, E. (2024). Applications of large language models in psychiatry: A systematic review. *Frontiers in Psychiatry, 15*, Article 1422807. https://doi.org/10.3389/fpsyt.2024.1422807

OpenAI. (2023a). *OpenAI. GPT3.5 turbo*.

OpenAI. (2023b). *OpenAI. GPT-4 technical report.arXiv:2303.08774*.

Perez, M. I., Limon, D. L., Candelari, A. E., Cepeda, S. L., Ramirez, A. C., Guzick, A. G., Kook, M., Ariza, V. L. B., Schneider, S. C., & Goodman, W. K. (2022). Obsessive-compulsive disorder misdiagnosis among mental healthcare providers in Latin America. *Journal of Obsessive-Compulsive and Related Disorders, 32*, Article 100693. https://doi.org/10.1016/j.jocrd.2021.100693

Perris, F., Sampogna, G., Giallonardo, V., Agnese, S., Palummo, C., Luciano, M., Fabrazzo, M., Fiorillo, A., & Catapano, F. (2021). Duration of untreated illness predicts 3-year outcome in patients with obsessive-compulsive disorder: A real-world, naturalistic, follow-up study. *Psychiatry Research, 299*, Article 113872. https://doi.org/10.1016/j.psychres.2021.113872

Ponzini, G. T., & Steinman, S. A. (2022). A systematic review of public stigma attributes and obsessive–compulsive disorder symptom subtypes. *Stigma and Health, 7*(1), 14. https://doi.org/10.1037/sah0000310

Rudolph, J., Tan, S., & Tan, S. (2023). War of the chatbots: Bard, bing chat, ChatGPT, Ernie and beyond. the new AI gold rush and its impact on higher education. *Journal of Applied Learning and Teaching, 6*(1). https://doi.org/10.37074/jalt.2023.6.1.23

Shams, G., Whittal, M. L., Past, N., por, V. M., Yousefi, Y., & Abasi, I. (2024). Intrusive thoughts in patients with obsessive–compulsive and major depressive disorders and non-clinical participants: A comparison using the international intrusive thoughts interview schedule. *Current Psychology*, 1–15. https://doi.org/10.1007/s12144-024-06063-9

Sharma, E., Sharma, L. P., Balachander, S., Lin, B., Manohar, H., Khanna, P., Lu, C., Garg, K., Thomas, T. L., & Au, A. C. L. (2021). Comorbidities in obsessive-compulsive disorder across the lifespan: A systematic review and meta-analysis. *Frontiers in Psychiatry, 12*, Article 703701. https://doi.org/10.3389/fpsyt.2021.703701

Shavitt, R. G., van den Heuvel, O. A., Lochner, C., Reddy, Y. J., Miguel, E. C., & Simpson, H. B. (2022). Obsessive-compulsive disorder (OCD) across the lifespan: Current diagnostic challenges and the search for personalized treatment. *Frontiers in Psychiatry, 13*, Article 927184. https://doi.org/10.3389/fpsyt.2022.927184

Stade, E. C., Stirman, S. W., Ungar, L. H., Boland, C. L., Schwartz, H. A., Yaden, D. B., … Eichstaedt, J. C. (2024). Large language models could change the future of behavioral healthcare: A proposal for responsible development and evaluation. *NPJ Mental Health Research, 3*(1), 12. https://doi.org/10.1038/s44184-024-00056-z

Steinberg, D. S., & Wetterneck, C. T. (2017). OCD taboo thoughts and stigmatizing attitudes in Clinicians. *Community Mental Health Journal, 53*, 275–280. https://doi.org/10.1007/s10597-016-0055-x

Tal, A., Elyoseph, Z., Haber, Y., Angert, T., Gur, T., Simon, T., & Asman, O. (2023). The artificial third: Utilizing ChatGPT in mental health. *The American Journal of Bioethics, 23*(10), 74–77. https://doi.org/10.1080/15265161.2023.2250297

Thornicroft, G., Bakolis, I., Evans-Lacko, S., Gronholm, P. C., Henderson, C., Kohrt, B. A., Koschorke, M., Milenova, M., Semrau, M., & Votruba, N. (2019). Key lessons learned from the INDIGO global network on mental health related stigma and discrimination. *World Psychiatry, 18*(2), 229–230. https://doi.org/10.1002/wps.20628

Tornero-Costa, R., Martinez-Millana, A., Azzopardi-Muscat, N., Lazeri, L., Traver, V., & Novillo-Ortiz, D. (2023). Methodological and quality flaws in the use of artificial intelligence in mental health research: Systematic review. *JMIR Mental Health, 10*(1), Article e42045. https://doi.org/10.2196/42045

Wairauch, Y., Siev, J., Hasdai, U., & Dar, R. (2024). Compulsive rituals in obsessive-compulsive disorder–A qualitative exploration of thoughts, feelings and behavioral patterns. *Journal of Behavior Therapy and Experimental Psychiatry, 84*, Article 101960. https://doi.org/10.1016/j.jbtep.2024.101960

Zhang, H., Zhang, C., & Wang, Y. (2024). Revealing the technology development of natural language processing: A scientific entity-centric perspective. *Information Processing & Management, 61*(1), Article 103574. https://doi.org/10.1016/j.ipm.2023.103574