Family Medicine and Community Health

# Assessing prognosis in depression: comparing perspectives of AI models, mental health professionals and the general public

Zohar Elyoseph,[1,2] Inbar Levkovich ,[3] Shiri Shinan-Altman[4]

**To cite:** Elyoseph Z, Levkovich I, Shinan-Altman S. Assessing prognosis in depression: comparing perspectives of AI models, mental health professionals and the general public. *Fam Med Com Health* 2024;**12**:e002583. doi:10.1136/fmch-2023-002583



[1]Department of Psychology and Educational Counseling, The Center for Psychobiological Research, Max Stern Yezreel Valley College, Yezreel Valley, Israel
[2]Department of Brain Sciences, Imperial College London, London, UK
[3]Faculty of Graduate Studies, Oranim Academic College, Tivon, Israel
[4]The Louis and Gabi Weisfeld School of Social Work, Bar-Ilan University, Ramat Gan, Tel Aviv, Israel

**Correspondence to**
Dr Zohar Elyoseph;
zohare@yvc.ac.il

## ABSTRACT

**Background** Artificial intelligence (AI) has rapidly permeated various sectors, including healthcare, highlighting its potential to facilitate mental health assessments. This study explores the underexplored domain of AI's role in evaluating prognosis and long-term outcomes in depressive disorders, offering insights into how AI large language models (LLMs) compare with human perspectives.

**Methods** Using case vignettes, we conducted a comparative analysis involving different LLMs (ChatGPT-3.5, ChatGPT-4, Claude and Bard), mental health professionals (general practitioners, psychiatrists, clinical psychologists and mental health nurses), and the general public that reported previously. We evaluate the LLMs ability to generate prognosis, anticipated outcomes with and without professional intervention, and envisioned long-term positive and negative consequences for individuals with depression.

**Results** In most of the examined cases, the four LLMs consistently identified depression as the primary diagnosis and recommended a combined treatment of psychotherapy and antidepressant medication. ChatGPT-3.5 exhibited a significantly pessimistic prognosis distinct from other LLMs, professionals and the public. ChatGPT-4, Claude and Bard aligned closely with mental health professionals and the general public perspectives, all of whom anticipated no improvement or worsening without professional help. Regarding long-term outcomes, ChatGPT 3.5, Claude and Bard consistently projected significantly fewer negative long-term consequences of treatment than ChatGPT-4.

**Conclusions** This study underscores the potential of AI to complement the expertise of mental health professionals and promote a collaborative paradigm in mental healthcare. The observation that three of the four LLMs closely mirrored the anticipations of mental health experts in scenarios involving treatment underscores the technology's prospective value in offering professional clinical forecasts. The pessimistic outlook presented by ChatGPT 3.5 is concerning, as it could potentially diminish patients' drive to initiate or continue depression therapy. In summary, although LLMs show potential in enhancing healthcare services, their utilisation requires thorough verification and a seamless integration with human judgement and skills.

### WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Large language models (LLMs) such as ChatGPT, Claude and Bard, pretrained on vast datasets, use transformers to offer human-like chatbot interactions and generate diverse content. While LLMs hint at potential in mental health applications, research is scant on their efficacy in evaluating depression prognosis and long-term outcomes.

### WHAT THIS STUDY ADDS

⇒ This is the first analysis to compare the ability of four leading LLMs (ChatGPT-3.5, ChatGPT-4, Claude and Bard) to evaluate prognosis and outcomes for depression cases with that of mental health professionals and the general public. It reveals variability between the LLMs: some align closely with professional human perspectives, while others differ, underscoring the need for ongoing LLM assessment and refinement.

### HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ The findings demonstrate LLMs' potential value in complementing professionals' expertise but also highlight the need for further research to optimise LLM integration and ensure equitable application in mental healthcare. This study lays the groundwork for expanded investigations into LLM capabilities and biases across diverse clinical scenarios and populations. It highlights the importance of developing responsible policies and practices as LLMs assume greater roles in mental health assessment and care.

## INTRODUCTION

Artificial intelligence (AI) has found applications in a myriad of fields from medicine to mental health.[1 2] While numerous studies have examined these diverse applications, our research is, to the best of our knowledge, the first to focus specifically on the use of AI to assess the prognosis of depressive disorders. This focus is crucial for helping patients to make informed decisions about their treatment and enhancing the transparency of the

therapeutic process.[3 4] Ultimately, it fosters a collaborative approach to healthcare, empowering patients and medical professionals to make joint informed decisions.[5–8]

Major depressive disorders (MDDs) are characterised as severe affective disorders manifesting in symptoms such as persistent low mood, anhedonia, emotional void, disruptions in sleep patterns and diminished appetite.[9] The disorder has substantial ramifications on multiple facets of an individual's life, including emotional well-being, social interactions, academic achievement and overall developmental trajectory.[10] Epidemiologically, depressive disorders exhibit high prevalence rates and are associated with significant economic burden, compromised quality of life, medical comorbidities and increased mortality rates.[11 12] Meta-analytical data incorporating 90 studies, with a cumulative sample size of 1 112 573 adults, indicated gender-specific prevalence rates of 14.4% for women and 11.5% for men.[13]

The age of onset for MDD spans from mid-adolescence to mid-adulthood, although nearly 40% of affected individuals report experiencing their inaugural episode prior to the age of 20.[14] Risk factors implicated in the aetiology of MDD include genetic predispositions, personality traits, psychopathological elements, comorbid psychiatric and physiological conditions, and specific life events such as elevated stress levels, historical trauma and a history of MDD among first-degree relatives.[15–17]

The multifaceted nature of recovery for individuals grappling with enduring mental health difficulties is subject to heterogeneous interpretations. Within the clinical framework, recovery is predominantly conceptualised in terms of the alleviation of symptoms and the remediation of functional impairments.[18 19] The complete lack of psychological indicators is rarely characteristic of the typical healthy demographic. Thus, the definition of recovery is influenced by the established severity threshold of symptoms and is reliant on the categorisation and properties of the assessment tools employed. However, from the vantage point of lived experience, recovery assumes an individualised and potentially ongoing journey towards the reclamation of a meaningful life characterised by purpose and active societal participation regardless of persistent symptoms.[19 20] Literature on pharmacological interventions for MDD has evidenced a cumulative remission rate of 67% following antidepressant therapy.[21] Additional empirical studies have indicated that, following a 3-month course of antidepressant treatment, 66% of patients achieved remission while 59.5% regained normative levels of functionality.[22] Notably, incomplete remission in the context of MDD is prevalent; approximately one-third of individuals diagnosed continue to exhibit residual symptoms even during periods identified as remission.[23]

In recent decades, there have been significant advancements in the research and clinical management of depression, particularly in primary healthcare settings. A plethora of pharmacological and psychotherapeutic interventions have been validated through rigorous randomised controlled trials, thereby gaining inclusion in established treatment guidelines.[24 25] These interventions have subsequently been extensively adopted in clinical practice.[26] Notably, primary care serves as the predominant healthcare setting for the treatment of depressive disorders, accommodating the majority of affected individuals. Statistical data indicate that 73% of patients receive treatment for depression exclusively in primary care, while a substantially smaller proportion—24% and 13%, respectively—are managed by psychiatrists or other specialised mental health practitioners.[27 28]

The clinician's stance on a patient's recuperative potential is complex and has many dimensions.[29] From a utilitarian perspective, the medical professional's acumen in prognosticating a patient's likely therapeutic course—commonly referred to as 'prognosis'—is indispensable clinical competency.[28] Ethical considerations compel healthcare providers to thoroughly explain both the attendant risks and merits of prospective treatments to patients, thereby enabling the exercise of informed consent and fostering a collaborative model of decision-making.[30] Providing a nuanced and forthright prognosis serves to bolster patient morale and cultivate optimism in instances where full recovery is plausible while tempering expectations in more adverse clinical scenarios.[6 7 31] However, it should be acknowledged that clinicians' prognostic judgements are inevitably influenced by their own foundational beliefs and assumptions.[32 33]

Extensive empirical research has corroborated the efficacy of psychotherapeutic interventions, underscoring the positive correlation between a robust therapeutic alliance and favourable treatment outcomes.[34–36] These findings precipitated a growing emphasis on recovery-oriented practices which are linked to an array of beneficial patient outcomes, including enhanced functional capabilities and reduced hospitalisation rates.[37 38] Despite these advancements, it is important to acknowledge that the prevailing mental healthcare paradigm, rooted largely in the biomedical model not only foregrounds clinical recovery and symptom remission but is also influenced by clinicians' attitudes, including potential stigmatisation towards patients exhibiting delayed treatment engagement.[25 28 29] As such, practitioners' beliefs about patients' recovery potential and the depth of the therapeutic relationship play a pivotal role in the overall efficacy of the treatment regimen.[6–8]

AI has become ubiquitous across multiple domains, including but not limited to political science, economics, healthcare and biological sciences.[1 2] Previous scholarly investigations have explored the application of AI in the realm of applied psychology, either examining rudimentary clinical capabilities.[3 4] or focusing on decision-making processes in intricate clinical scenarios, such as those involving depressive disorders and suicidal ideation.[5] To date, there is a literature gap concerning the capability of AI to facilitate the process of recovery or healing in the context of mental health disorders. However, a burgeoning

body of literature has underscored the significant therapeutic implications of a clinician's belief in the patient's potential for recovery[6 7 25] as well as the deleterious consequences arising when such beliefs are absent.[34]

In light of the growing integration of AI technologies in healthcare sectors—particularly given the nascent advancements in emotion recognition and mental health risk stratification[3 4]—it has become critical to rigorously scrutinise how various AI systems conceptualise and interpret human resilience and prospects for recovery.[5] This line of inquiry assumes paramount importance as both healthcare providers and patients increasingly rely on AI for diagnostic consultations and therapeutic interventions. These understandings not only shape the future direction of patient care but also serve as a cornerstone for psychoeducational initiatives, clinical guidance and targeted interventions.

The present study was predicated on an evaluation of perspectives among mental health professionals in Australia, comprising 328 mental health nurses, 535 psychiatrists, 434 general practitioners (GPs), 211 clinical psychologists and 952 laypeople as reported by Caldwell and Jorm.[39] Respondents were surveyed concerning their beliefs about prognosis, long-term outcomes and potential discriminatory practices in the context of case vignettes featuring individuals diagnosed with depression.

The objectives of this study are to:

1. Compare the assessment of the prognosis for individuals with depression between four large language models (LLMs) (ChatGPT-3.5, ChatGPT-4, Claude and Bard), mental health professionals (including GPs, psychiatrists, clinical psychologists and mental health nurses) and the general public. Furthermore, this comparison will also consider evaluations of prognoses with and without treatment.
2. Analyse the evaluations by the four LLMs, mental health professionals and the general public regarding the positive and negative long-term outcomes for individuals with depression.

## METHODS

### AI procedure

During August 2023, we used ChatGPT-3.5 and 4 (by OpenAI; 3 August version), Claude.AI (by Anthropic) and Bard (by Google) to evaluate the prognosis of depression compared with the results of mental health professionals[40] and of the general public reported by Caldwell and Jorm.[39]

### Input source

In the present study, we adopted the text vignette methodology, as elucidated by Caldwell and Jorm[39] and Jorm *et al*[40] and to examine mental health professionals (GPs, psychiatrists, clinical psychologists and mental health nurses) and general public beliefs about the helpfulness of interventions for depression. Demographic data of the mental health professionals are present in table 1. The study contains two vignettes that differed only in the gender of the main character and were entered into the LLMs' interface. Each of the vignettes was evaluated 10 times in ChatGPT-3.5, ChatGPT-4, Claude and Bard in a new tab, with 80 evaluations in total. Following is the male version of the vignette:

John is 30 years old. He has been feeling unusually sad and miserable for the last few weeks. Even though he is tired all the time, he has trouble sleeping nearly every night. John doesn't feel like eating and has lost weight. He can't keep his mind on his work and puts off making decisions. Even day-to-day tasks seem too much for him. This has come to the attention of John's boss who is concerned about his lowered productivity.

### Measures

After each of the two vignettes, we submitted six questions to the LLMs taken from Jorm *et al*[40] :

Q1: '*Describe what, if anything, is wrong with John.*'

Q2: '*What will be helpful for John?*'

Q3: *'hat is most likely to be the result for John if he receives the sort of professional help that you think most appropriate?' (A. Full recovery with no further problems; B. Full recovery but problems will probably reoccur; C. Partial recovery; D. Partial recovery, but problems will probably reoccur; E. No improvement; F. Get worse*)

**Table 1** Demographic data of the mental health professionals (psychiatrists, clinical psychologists, GPs and. nurses, as reported by Caldwell and Jorm)[39]

| | Psychiatrists | Clinical psychologists | GPs | Nurses |
|---|---|---|---|---|
| N | 535 | 211 | 438 | 328 |
| ≥39 | 15% | 20% | 32% | |
| 40–49 | 37% | 46% | 36% | |
| 50–59 | 30% | 21% | 17% | |
| 60+ | 18% | 13% | 15% | |
| Response rate | 75% | 65% | 61% | 70% |
| F/M ratio | ~37% male | | | |

GPs, general practitioners.

Q4: '*What is most likely to be the result for John if he does not receive any professional help?' (A. Full recovery with no further problems; B. Full recovery but problems would probably reoccur; C. Partial recovery; D. Partial recovery but problems would probably reoccur; E. No improvement; F. Get worse.*).

Q5: 'How do you think he will be in the long term compared to other people in the community?' (Rate each of the following as: more likely, just as likely, or less likely.) Negative long-term outcomes: (a) to be violent, (b) to drink too much, (c) to take illegal drugs, (d) to have poor friendships, (e) to attempt suicide; Positive long-term outcomes: (f) to be understanding of other people's feelings, (g) to have a good marriage, (h) to be a caring parent, (i) to be a productive worker and (j) to be creative or artistic.

Q6. 'Do you think that John will be discriminated against by others in the community if they know about his problems?' (Yes/No).

### Scoring

The performances of all the LLMs were scored according to Caldwell and Jorm[39] and Jorm *et al*[40] We then compared the LLMs' performance to the results of previous 535 psychiatrists, 211 clinical psychologists, 434 GPs, 328 mental health nurses and 952 people from the general public collected in Australia and reported by Caldwell and Jorm[39] and Jorm *et al*.[40] Q5, evaluating the positive and negative long-term outcomes, was calculated according to Caldwell and Jorm.[39] Each of the 10 statements received a score of 1, 0 or −1 (more likely, just as likely or less likely, respectively). The answers were then summed up, with each of the positive and negative long-term outcome scales ranging from −5 to 5.

### Patient and public involvement

This study did not involve patient participation; it focused on AI insights. Results will be shared via social media in simplified language for patient communities.

### Statistical analysis

The likely outcome with and without professional help for the female and male vignettes, evaluated by the LLMs, mental help professionals and the general public (as reported by Caldwell and Jorm[39] and Jorm *et al*[40]), was analysed using a one-way ANOVA with Fisher's least significant difference as post hoc. A comparison of the differences between LLMs in the positive and negative long-term outcome was analysed using a one-way ANOVA with Bonferroni as post hoc.

## RESULTS

All four LLMs recognised depression as the primary diagnosis in all the vignette cases and suggested a combination of psychotherapy and antidepressant drugs as the preferred treatment.

### Likely outcome with professional help

Table 2 delineates the distribution of selected outcomes according to LLMs, mental health professionals groups, and the general public for a case vignette of an individual diagnosed with depression after treatment. On conducting an analysis of variance (ANOVA), we identified a significant difference in the outcomes selected across the eight groups ($F(8, 2540)=13.56$, $p<0.001$) (see figure 1). Post hoc analysis yielded the following insights: ChatGPT-3.5 displayed a distinctively pessimistic prognosis, differing significantly from the outcomes chosen by the three other LLMs, professional groups and the general public; ChatGPT-4, Claude and Bard were congruent with the psychiatrists, GPs, clinical psychologists, mental health nurses, but was more pessimistic than the general public; and no significant differences were found when comparing the projections of ChatGPT-4, Claude and Bard directly (see table 3).

### Likely outcome without professional help

Table 2 delineates the distribution of selected outcomes according to LLMs, mental health professional groups and the general public for a case vignette of an individual diagnosed with depression without receiving professional help. All asserted that the person with depression without treatment will show no improvement or get worse. On conducting an ANOVA, we identified a significant difference in the prognoses selected across the eight groups ($F(8, 2540)=21.06$, $p<0.001$) (figure 1). A post hoc analysis yielded the following insight: all four LLMs displayed a distinctively pessimistic prognosis than the prognosis chosen by mental health professionals but were not different from each other or from the general public (see table 3).

### Long-term outcomes

Figure 2 illustrates the LLMs' output concerning the positive and negative long-term outcomes. On conducting an ANOVA, we identified a significant difference in the negative outcomes selected across the four LLMs groups ($F(3, 80)=9.34$, $p<0.001$). ChatGPT 3.5, Claude and Bard generated a significantly lower negative long-term outcome after receiving treatment than ChatGPT-4.

On conducting an ANOVA, we identified significant differences also in the positive outcomes selected across the four LLMs groups ($F(3, 80)=14.43$, $p<0.001$): Bard generated a more optimistic evaluation of a positive long-term outcome after receiving treatment than ChatGPT 3.5, ChatGPT-4 and Claude; ChatGPT-3.5 generated a more pessimistic evaluation than ChatGPT-4 and Claude; and no significant differences were found between ChatGPT-4 and Claude.

## DISCUSSION

This study investigated how different LLMs, including ChatGPT-3.5, ChatGPT-4, Claude, and Bard, evaluate the prognosis, long-term outcomes and potential discriminatory practices associated with depression. The results were analysed according to the responses from LLMs, mental health professionals (GPs, psychiatrists, clinical

**Table 2** The likely outcome evaluations of LLMs, mental health professionals and the general public reported by Caldwell and Jorm and Jorm *et al*[39 40] and with and without professional help (%)

| With professional help | ChatGPT 3.5 | ChatGPT 4 | Bard | Claude | General public | Nurses | Clinical psychologists | Psychiatrists | GPs |
|---|---|---|---|---|---|---|---|---|---|
| Full recovery with no further problems | 0.0 | 0.0 | 75.0 | 35.0 | 46.5 | 42.1 | 45.3 | 37.0 | 24.4 |
| Full recovery but problems will probably reoccur | 0.0 | 100.0 | 0.0 | 45.0 | 39.1 | 54.5 | 48.1 | 61.3 | 70.8 |
| Partial recovery | 75.0 | 0.0 | 25.0 | 20.0 | 6.8 | 1.2 | 2.4 | 0.6 | 0.7 |
| Partial recovery but problems will probably reoccur | 25.0 | 0.0 | 0.0 | 0.0 | 7.0 | 2.1 | 4.2 | 0.9 | 3.9 |
| No improvement | 0.0 | 0.0 | 0.0 | 0.0 | 0.2 | 0.0 | 0.0 | 0.2 | 0.0 |
| Get worse | 0.0 | 0.0 | 0.0 | 0.0 | 0.3 | 0.0 | 0.0 | 0.0 | 0.2 |
| **Without professional help** | | | | | | | | | |
| Full recovery with no further problems | 0.0 | 0.0 | 0.0 | 0.0 | 1.2 | 1.5 | 1.9 | 2.2 | 1.2 |
| Full recovery, but problems will probably reoccur | 0.0 | 0.0 | 0.0 | 0.0 | 4.7 | 5.2 | 12.4 | 9.6 | 6.0 |
| Partial recovery | 0.0 | 0.0 | 0.0 | 0.0 | 4.0 | 5.2 | 4.8 | 4.9 | 2.5 |
| Partial recovery but problems will probably reoccur | 0.0 | 20.0 | 0.0 | 0.0 | 12.2 | 27.1 | 43.1 | 32.2 | 27.3 |
| No improvement | 45.0 | 10.0 | 0.0 | 0.0 | 18.5 | 8.6 | 12.0 | 10.5 | 13.7 |
| Get worse | 55.0 | 70.0 | 100.0 | 100.0 | 59.5 | 52.3 | 25.8 | 40.6 | 49.3 |

GPs, general practitioners; LLMs, large language models.

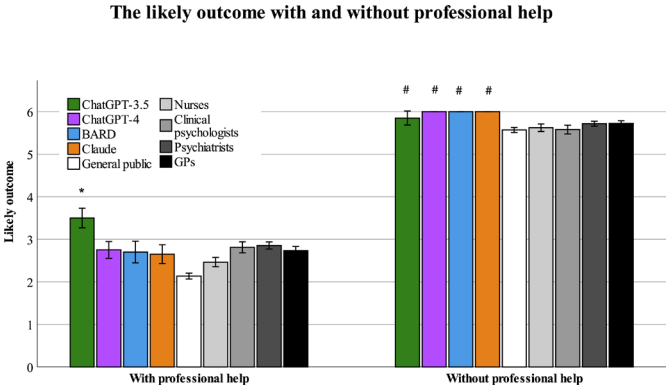


**Figure 1** The likely outcome evaluations of LLMs, mental health professionals and the general public (mean±SEM), with and without professional help. *Significant compared with ChatGPT-4, Bard Claude and all Human groups, #significant compared with all human groups instead of Chatgpt-3 that was not significantly different from the general public. GPs, general practitioners; LLMs, large language models.

psychologists and mental health nurses) and the general public. Several key findings emerged which shed light on the perspectives and knowledge on depression and its treatment.

The consistent recognition of depression cases and recommendation of a combination of psychotherapy and antidepressant drugs by all four LLMs in all of the cases is a noteworthy finding. It underscores the proficiency of these AI models in accurately identifying mental health conditions, aligning their diagnoses with established clinical practices, and suggesting evidence-based treatment approaches.[41] This not only demonstrates LLMs' reliability in mental health assessments but also highlights their potential to provide valuable support and guidance to individuals seeking information or coping with depression. This finding also has implications for clinical decision support, whereby LLMs can assist healthcare professionals in their initial assessments and treatment planning. This appears to be consistent with the views of healthcare experts concerning the utilisation of ChatGPT. In a survey that evaluated healthcare professionals' experiences with ChatGPT, a substantial majority (75.1%) indicated that they felt at ease with the prospect of incorporating ChatGPT into their healthcare routines, with 39.5% favouring using ChatGPT to assist in making medical decisions.[42] However, it must be emphasised that while LLMs can serve as valuable tools, they should complement rather than replace the expertise of qualified mental health professionals to ensure comprehensive and personalised care.

## Likely outcome with/without professional help

The study's findings offer valuable insights into how different groups perceive the likely outcomes for individuals with depression when they receive professional help compared with when they do not. When individuals with depression receive professional help, there is a shared sense of optimism among all groups (LLMs, mental health

**Table 3** Post hoc analyses for differences between LLMs, professionals and the general public in assessing the outcome of depression with and without treatment

| | | ChatGPT-4 | Bard | Claude | General public | Nurses | Clinical psychologists | Psychiatrists | GPs |
|---|---|---|---|---|---|---|---|---|---|
| With professional help | ChatGPT-3.5 | *** | *** | *** | *** | *** | *** | *** | *** |
| | ChatGPT-4 | | ns | ns | ** | * | ns | ns | ns |
| | Bard | | | ns | ns | ns | ns | ns | ns |
| | Claude | | | | ns | ns | ns | ns | ns |
| | General Public | | | | | * | *** | *** | *** |
| | Nurses | | | | | | *** | *** | *** |
| | Clinical Psychologists | | | | | | | ns | ns |
| | Psychiatrists | | | | | | | | ns |
| Without professional help | ChatGPT-3.5 | ns | ns | ns | ns | * | *** | * | * |
| | ChatGPT-4 | | ns | ns | * | 0.53 | *** | ** | 0.057 |
| | Bard | | | ns | *** | *** | *** | *** | *** |
| | Claude | | | | *** | *** | *** | *** | *** |
| | General public | | | | | ns | *** | ** | ns |
| | Nurses | | | | | | *** | ns | ns |
| | Clinical psychologists | | | | | | | *** | ** |
| | Psychiatrists | | | | | | | | *** |

*p<0.05; **p<0.01; ***p<0.001.
GPs, general practitioners; LLMs, large language models; ns, not stated.

professionals and the general public), indicating the significance of receiving professional help when coping with depression.[43] Professional help is known to be effective for preventing and managing mental health issues, improving understanding, coping skills, and interpersonal support, and thereby reducing suicidal thoughts.[44]

However, the findings paint a pessimistic picture when individuals with depression do not receive professional assistance regardless of the group: all share the belief that the likelihood of positive outcomes diminishes while the probability of negative outcomes increases. In other words, AI models consistently align with established clinical practices, recommending evidence-based treatments and emphasising the significance of professional intervention. Conversely, without professional assistance, LLMs, along with other groups, converge on pessimistic prognoses, underscoring their adherence to a deterministic view of mental health outcomes. Variability in assessments, particularly ChatGPT-3.5's pessimistic outlook on recovery, mirrors the multifaceted nature of mental health recovery and highlights that LLMs may not fully capture phenomenological and social models which take into account personal choice and decision as well as social acceptance and lack of exclusion.[19 20] These tendencies, rooted in LLMs' reliance on clinical data and established medical knowledge, emphasise the need for their further development to incorporate a more holistic understanding of mental health experiences, bridging the gap between medical and phenomenological social models in mental health assessment and support.

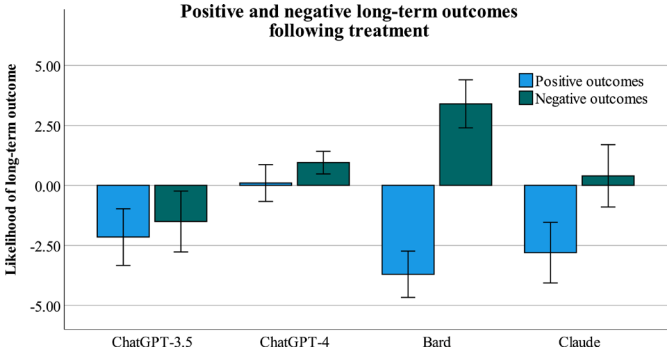


**Figure 2** The positive and negative long-term outcomes evaluated by LLMs (ChatGPT-3.5, ChatGPT-4, Claude and Bard; mean±SEM). LLMs, large language models.

The study's findings indicated that AI models can vary in their assessments of recovery outcomes. ChatGPT-3.5 had a consistently pessimistic outlook, while ChatGPT-4, Claude and Bard aligned more closely with human perspectives, namely, mental health professionals and the general public. ChatGPT-3.5's pessimism in evaluating recovery outcomes could, worryingly, deter individuals from seeking professional assistance; its widespread use worldwide amplifies this potential risk.[45]

## Long-term outcomes

This study explores long-term outcomes after depression treatment, shedding light on how different LLMs evaluate prospects for individuals with this condition. Notably, there are distinctions between LLMs regarding negative long-term consequences post-treatment: ChatGPT-3.5, Claude and Bard tend to offer more optimistic assessments

than ChatGPT-4, indicating a less pessimistic outlook on potential negative effects. Bard stands out as the most optimistic regarding positive long-term consequences, while ChatGPT-3.5 leans towards a more pessimistic view than ChatGPT-4 and Claude. These findings highlight the importance of acknowledging LLM variability in evaluating long-term outcomes for those dealing with mental health issues.[46] They suggest that the choice of LLM can impact the level of optimism or pessimism conveyed in their assessments. It should be noted that the accuracy of LLMs' predictions is closely tied to the quality and inclusivity of the data used for their training. Biases in the data or a lack of representation from diverse demographics can result in incorrect predictions.[47] Additionally, LLMs algorithms often operate as opaque systems, making it difficult to understand how each AI arrived at specific conclusions.[48] These variations have implications for how LLMs report on long-term outcomes of depression. Recognising these differences and their potential influence on patient outcomes is pivotal for the ongoing advancement and incorporation of LLMs in mental healthcare. Moreover, the results indicate the importance of using various ChatGPT versions in a complementary manner rather than relying exclusively on one. Previous studies on AI have shown its potential for predicting clinical outcomes. Brendese[49] illuminates the AI's potential in neurodegenerative diseases, highlighting its role in diagnosis, treatment and prediction of disease progression and response to intervention. Additionally, Liao *et al*[50] demonstrated the AI's potential to predict clinical prognosis in tuberculosis. In the realm of mental health, AI's applicability extends to both diagnostic and prognostic evaluations and potential contributions to personal medicine.[7 51–56] Our research contributes uniquely to this body of work by evaluating open-to-the-public tools for mental health prognosis, making them accessible and operational across languages and regions, thus opening new avenues for global mental healthcare enhancements.[54]

### Transparency, data integrity and integration with conventional methods

The integration of LLMs in evaluating mental health prognoses warrants a discussion on their decision-making transparency, data input considerations and comparison with traditional depression scales such as the Patient Health Questionnaire-9 or Beck Depression Inventory. While AI can analyse a broader range of textual data, offering potentially nuanced insights, it's crucial to acknowledge how the omission of key data can impact their accuracy. For instance, neglecting to input comprehensive details about symptom duration or severity could lead to less precise prognoses, emphasising the need for complete and relevant data. Furthermore, the limitations of AI in understanding and interpreting mental health conditions must be recognised. The decisions of AI models, shaped by their training datasets, lack the depth of human comprehension. This raises questions about the reliability of AI-generated prognoses and their dependence on the quality and diversity of training data.[52] Enhancing AI transparency is vital, especially in explaining the rationale behind AI-generated decisions in a way that is comprehensible to healthcare professionals. This is crucial for fostering trust and ensuring that AI tools are used appropriately as a complement to, rather than a replacement for, traditional assessment methods. Our findings add to the expanding understanding of AI's role in mental health,[3–5 7 46 51–55] highlighting the importance of combining AI tools with traditional techniques and clinical expertise for a comprehensive approach to mental health diagnosis and treatment.

While the findings of this study offer valuable insights, several limitations warrant consideration. First, the utilisation of case vignettes, though a useful tool, simplifies the complexities often seen in real-life clinical scenarios. Future investigations could benefit from incorporating real patient data or using clinical simulations to offer a more nuanced understanding of prognoses. Second, it should be acknowledged that this study centred on specific LLMs, potentially limiting the generalisability of the findings to other AI models. To enhance the comprehensiveness of future studies, a more diverse range of LLMs should be examined. Third, this study focused primarily on comparing LLM assessments with human perspectives without directly evaluating their clinical accuracy. Future research endeavours should incorporate validation studies to assess the clinical utility and reliability of LLM predictions. Fourth, Since the results of professionals and the general public were based on findings from previous studies,[39 40] it was not possible to test the effect of demographic variables (different career stages, gender, etc) on the clinical assessment.

Lastly, the study did not explore potential cultural or demographic biases within the LLMs. Addressing these aspects in future research is crucial to ensure the equitable application of AI in mental health across diverse populations.

In addressing the divergent prognoses of ChatGPT-3.5 compared with ChatGPT-4, Claude and Bard in our study, several factors may contribute to these results. ChatGPT-4's larger number of parameters might contribute to its more accurate processing capabilities. Differences in data handling and retrieval methods, as seen with Claude's updated processing or Bard's access to Google's search index, are also potential explanations. The specific reason for ChatGPT-3.5's divergent prognosis remains unclear but could involve different training data, alignment processes and human feedback. This is consistent with previous findings[46] where ChatGPT was found to be less accurate than ChatGPT-4, further emphasising the advancements in AI model development.

## CONCLUSION

This study offers valuable insights into the integration of LLMs within the context of mental health assessment and support. These LLMs demonstrated a consistent ability

to recognise depression cases and recommend evidence-based treatment, aligning with established clinical practices. Moreover, the study highlights the optimism shared by LLMs, mental health professionals and the general public regarding the positive outcomes associated with professional help in depression management. The importance of seeking professional assistance is thus underscored, emphasising LLMs' potential to encourage individuals to access qualified healthcare support. The influence of the assessment of recovery over the therapeutic relationship should be noted here. It directly affects the therapist's commitment to dedicating time and effort to the patient as well as influencing the patient's motivation to initiate or sustain treatment. Indeed, the concept of recovery and hope is a pivotal theme in psychotherapy.[56] The integration of LLMs into this domain can significantly impact the therapeutic relationship.[55] We emphasise the necessity for ongoing validation and rigorous assessment of AI models like ChatGPT in clinical contexts. This process should include meticulous evaluations to mitigate biases, errors and hallucinations, and the utilisation of extensive retrospective medical data to enhance the models' precision. Future integrations could see AI functioning as a supportive copilot in clinical decision-making, complemented by a 'human in the loop' system, which also serves to alert clinicians to the high probability of inaccurate assessments. It is essential for mental health professionals to receive education on the complexities of AI in clinical practice, encouraging a critical yet objective stance in interpreting AI-generated outcomes.

**Contributors** Conceptualisation: ZE, IL and SA; Methodology: ZE; Formal analysis, ZE; Writing—original draft preparation, ZE, IL and SA; Writing—review and editing: IL, ZE and SA; Guarantor: ZE assumes full responsibility for the overall content of the work, had access to the data, and controlled the decision to publish.

**Funding** The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

**Competing interests** None declared.

**Patient and public involvement** Patients and/or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

**Patient consent for publication** Not applicable.

**Provenance and peer review** Not commissioned; externally peer reviewed.

**Data availability statement** Data are available on reasonable request.



**ORCID iD**

Inbar Levkovich http://orcid.org/0000-0002-5717-4074

## REFERENCES

1 Ali O, Abdelbaki W, Shrestha A, *et al*. A systematic literature review of artificial intelligence in the healthcare sector: benefits, challenges, methodologies, and functionalities. *Journal of Innovation & Knowledge* 2023;8:100333.
2 Mariani MM, Machado I, Nambisan S. Types of innovation and artificial intelligence: a systematic quantitative literature review and research agenda. *Journal of Business Research* 2023;155:113364.
3 Elyoseph Z, Hadar-Shoval D, Asraf K, *et al*. Chatgpt outperforms humans in emotional awareness evaluations. *Front Psychol* 2023;14:1199058:1199058.:.
4 Hadar-Shoval D, Elyoseph Z, Lvovsky M. The plasticity of ChatGPT's mentalizing abilities: personalization for personality structures. *Front Psychiatry* 2023;14:1234397.
5 Elyoseph Z, Levkovich I. Beyond human expertise: the promise and limitations of ChatGPT in suicide risk assessment. *Front Psychiatry* 2023;14:1213141.
6 Patterson JE, Edwards TM, Vakili S. Global mental health: a call for increased awareness and action for family therapists. *Fam Process* 2018;57:70–82.
7 Wampold BE, Flückiger C. The alliance in mental health care: conceptualization, evidence and clinical applications. *World Psychiatry* 2023;22:25–41.
8 Zilcha-Mano S. Toward personalized psychotherapy: the importance of the trait-like/state-like distinction for understanding therapeutic change. *Am Psychol* 2021;76:516–28.
9 American Psychiatric Association, A. P., & American Psychiatric Association. *Diagnostic and statistical manual of mental disorders: DSM-IV*. Washington, DC: American psychiatric association, 1994.
10 Taylor CB, Graham AK, Flatt RE, *et al*. Current state of scientific evidence on Internet-based interventions for the treatment of depression, anxiety, eating disorders and substance abuse: an overview of systematic reviews and meta-analyses. *Eur J Public Health* 2021;31:i3–10.
11 GBD 2019 Mental Disorders Collaborators. Global, regional, and national burden of 12 mental disorders in 204 countries and territories, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019. *Lancet Psychiatry* 2022;9:137–50.
12 Jorm AF, Patten SB, Brugha TS, *et al*. Has increased provision of treatment reduced the prevalence of common mental disorders? Review of the evidence from four countries. *World Psychiatry* 2017;16:90–9.
13 Lim GY, Tam WW, Lu Y, *et al*. Prevalence of depression in the community from 30 countries between 1994 and 2014. *Sci Rep* 2018;8:2861.
14 Achterbergh L, Pitman A, Birken M, *et al*. The experience of loneliness among young people with depression: a qualitative meta-synthesis of the literature. *BMC Psychiatry* 2020;20:415.
15 Davis A, McMaster P, Christie DC, *et al*. Psychiatric comorbidities of substance use disorders: does dual diagnosis predict inpatient detoxification treatment outcomes. *Int J Ment Health Addiction* 2023;21:3785–99.
16 Park LT, Zarate CA. Depression in the primary care setting. *N Engl J Med* 2019;380:559–68.
17 Wulandari P. An overlap between depression and anxiety: a literature review. *SciPsy* 2021;2:71–3.
18 Gunasekaran S, Tan GTH, Shahwan S, *et al*. The perspectives of healthcare professionals in mental health settings on stigma and recovery - a qualitative inquiry. *BMC Health Serv Res* 2022;22:888.
19 Slade M, Amering M, Farkas M, *et al*. Uses and abuses of recovery: implementing recovery-oriented practices in mental health systems. *World Psychiatry* 2014;13:12–20.
20 Andresen R, Oades LG, Caputi P. *Psychological recovery: beyond mental illness*. John Wiley & Sons, 2011.
21 Rush AJ, Trivedi MH, Wisniewski SR, *et al*. Acute and longer-term outcomes in depressed outpatients requiring one or several treatment steps: a STAR* D report. *FOC* 2008;6:128–42.
22 Romera I, Pérez V, Ciudad A, *et al*. Residual symptoms and functioning in depression, does the type of residual symptom matter? A post-hoc analysis. *BMC Psychiatry* 2013;13:51.
23 Cleare A, Pariante CM, Young AH, *et al*. Evidence-based guidelines for treating depressive disorders with antidepressants: a revision of the 2008 British Association for Psychopharmacology guidelines. *J Psychopharmacol* 2015;29:459–525.
24 Barth J, Munder T, Gerger H, *et al*. Comparative efficacy of seven psychotherapeutic interventions for patients with depression: a network meta-analysis. *Focus (Am Psychiatr Publ)* 2016;14:229–43.
25 Cipriani A, Furukawa TA, Salanti G, *et al*. Comparative efficacy and acceptability of 21 antidepressant drugs for the acute treatment of adults with major depressive disorder: a systematic review and network meta-analysis. *Lancet* 2018;391:1357–66.
26 Cuijpers P, Reijnders M, Huibers MJH. The role of common factors in psychotherapy outcomes. *Annu Rev Clin Psychol* 2019;15:207–31.
27 Olfson M, Blanco C, Marcus SC. Treatment of adult depression in the United States. *JAMA Intern Med* 2016;176:1482–91.
28 Wittchen H-U, Mühlig S, Beesdo K. Mental disorders in primary care. *Dialogues Clin Neurosci* 2003;5:115–28.

29 Sullivan PW, Starnino VR, Raster CG. In the eye of the beholder: recovery and personal narrative. *J Psychosoc Rehabil Ment Health* 2017;4:221–9.
30 Kennedy S, Lanceley A, Whitten M, *et al*. Consent on the labour ward: a qualitative study of the views and experiences of healthcare professionals. *Eur J Obstet Gynecol Reprod Biol* 2021;264:150–4.
31 Fimiani R, Gazzillo F, Gorman B, *et al*. The therapeutic effects of the therapists' ability to pass their patients' tests in psychotherapy. *Psychother Res* 2023;33:729–42.
32 Babcock G, McShea DW. Resolving teleology's false dilemma. *Biological Journal of the Linnean Society* 2023;139:415–32.
33 Hochstetter A, Vernekar R, Austin RH, *et al*. Deterministic lateral displacement: challenges and perspectives. *ACS Nano* 2020;14:10784–95.
34 Cuijpers P, Quero S, Dowrick C, *et al*. Psychological treatment of depression in primary care: recent developments. *Curr Psychiatry Rep* 2019;21:129.
35 Flückiger C, Del Re AC, Wampold BE, *et al*. The alliance in adult psychotherapy: a meta-analytic synthesis. *Psychotherapy (Chic)* 2018;55:316–40.
36 Barkham M, Lambert MJ. The efficacy and effectiveness of psychological therapies. In: Barkham M, Lutz W, Castonguay LG, eds. *Bergin and Garfield's handbook of psychotherapy and behavior change*. 50th anniversary edition. John Wiley & Sons, 2021: 135–89.
37 White C, Frimpong E, Huz S, *et al*. Effects of the personalized recovery-oriented services (PROS) program on hospitalizations. *Psychiatr Q* 2018;89:261–71.
38 Wong DFK, Chan V, Ip P, *et al*. The effects of recovery-oriented cognitive–behavior approach for Chinese with severe mental illness. *Research on Social Work Practice* 2019;29:311–22.
39 Caldwell TM, Jorm AF. Mental health nurses' beliefs about likely outcomes for people with schizophrenia or depression: a comparison with the public and other healthcare professionals. *Aust N Z J Ment Health Nurs* 2001;10:42–54.
40 Jorm AF, Korten AE, Jacomb PA, *et al*. Beliefs about the helpfulness of interventions for mental disorders: a comparison of general practitioners, psychiatrists and clinical psychologists. *Aust N Z J Psychiatry* 1997;31:844–51.
41 Uludag K. Testing creativity of ChatGPT in psychology: interview with ChatGPT. *SSRN Journal* 2023.
42 Temsah M-H, Aljamaan F, Malki KH, *et al*. Chatgpt and the future of digital health: a study on healthcare workers' perceptions and expectations. *Healthcare (Basel)* 2023;11:1812.
43 McLaren T, Peter L-J, Tomczyk S, *et al*. The seeking mental health care model: prediction of help-seeking for depressive symptoms by stigma and mental illness representations. *BMC Public Health* 2023;23:69.
44 Li XY, Liu Q, Chen P, *et al*. Predictors of professional help-seeking intention toward depression among community-dwelling populations: a structural equation modeling analysis. *Front Psychiatry* 2022;13:801231.
45 Pan C, Banerjee JS, De D, *et al*. Chatgpt: A Openai platform for society 5.0. 2. In: Bhattacharyya S, Banerjee JS, De D, et al., eds. *Intelligent human centered computing: Proceeds of Human 2023*. Springer, 2023: 384–97.
46 Levkovich I, Elyoseph Z. Suicide risk assessments through the eyes of Chatgpt-3.5 versus ChatGPT-4: vignette study. *JMIR Ment Health* 2023;10:e51232.
47 Sallam M. Chatgpt utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. *Healthcare* 2023;11:887.
48 Dahmen J, Kayaalp ME, Ollivier M, *et al*. Artificial intelligence Bot ChatGPT in medical research: the potential game changer as a double-edged sword. *Knee Surg Sports Traumatol Arthrosc* 2023;31:1187–9.
49 Brendese PJ. Artificial intelligence and machine learning models for diagnosing neurodegenerative disorders. *Cognitive Technologies* 2023.
50 Liao K-M, Liu C-F, Chen C-J, *et al*. Using an artificial intelligence approach to predict the adverse effects and prognosis of tuberculosis. *Diagnostics* 2023;13:1075.
51 Andrew J, Rudra M, Eunice J, *et al*. Artificial intelligence in adolescents mental health disorder diagnosis, prognosis, and treatment. *Front Public Health* 2023;11:52.
52 Hadar-Shoval D, Asraf K, Mizrachi Y. The invisible embedded "values" within large language models: implications for mental health use.
53 Elyoseph Z, Refoua E, Asraf K, *et al*. Can large language models "read your mind in your eyes"? (preprint). *JMIR Mental Health* [Preprint].
54 Levkovich I, Elyoseph Z. Identifying depression and its determinants upon initiating treatment: ChatGPT versus primary care physicians. *Fam Med Community Health* 2023;11:e002391.
55 Tal A, Elyoseph Z, Haber Y, *et al*. The artificial third: utilizing ChatGPT in mental health. *Am J Bioeth* 2023;23:74–7.
56 Sekechi M, Chiesa M. From hopelessness and despair to hope and recovery: psychoanalytic psychotherapy as effective agent of change in the treatment of a psychiatric patient. *Brit J Psychotherapy* 2022;38:483–99.