Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Can conversational AI reliably assess depressive symptoms? A feasibility, acceptability and bias-aware validation study of MADRS-Chat

Samantha Weber
samantha.weber@bli.uzh.ch
Psychiatrische Universitätsklinik Zürich
Barbora Provaznikova
Psychiatrische Universitätsklinik Zürich
Nikolas Psathakis
Psychiatrische Universitätsklinik Zürich
Andrea Casanova
Psychiatrische Universitätsklinik Zürich
Lucian Amthor
Psychiatrische Universitätsklinik Zürich
Lara Kirchhofer
Psychiatrische Universitätsklinik Zürich
José Antonio García Macías
Psychiatrische Universitätsklinik Zürich
Tobias Welt
Psychiatrische Universitätsklinik Zürich
Golo Kronenberg
Psychiatrische Universitätsklinik Zürich
Tobias Kowatsch
University of Zurich
Birgit Kleim
University of Zurich
Sebastian Olbrich
Psychiatrische Universitätsklinik Zürich

Article

Keywords: Large Language Models, MADRS, feasibility, acceptability, evaluation, responsible AI

Posted Date: November 3rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10848129/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: Competing interest reported. SW and TK are affiliated with the Centre for Digital Health Interventions, a joint initiative of the Institute for Implementation Science in Health Care, University of Zurich; the Department of Management, Technology, and Economics at ETH Zurich; and the Institute of Technology Management and School of Medicine at the University of St.Gallen. CDHI is funded in part by CSS, a Swiss health insurer; Mavie Next, an Austrian care provider; and MTIP, a Swiss digital health investor. TK is also co-founder of Pathmate Technologies, a university spin-off company that creates and delivers digital clinical pathways. However, CSS, Mavie Next, MTIP, and Pathmate Technologies were not involved in this research. TK has neither shares nor any formal role with Pathmate Technologies. SO hs co-founder of DeepPsy (deeppsy.io) a start-up that enables biomarker- and data-driven psychiatry. All other authors report no competing interests.

Abstract

Structured clinical interviews such as the Montgomery-Asberg Depression Rating Scale (MADRS) are central to depression assessment, yet their routine use is limited by time constraints and variability in administration. Conversational large language models may support standardized symptom assessment, but their feasibility, validity, and interaction-related biases in clinical settings remain unclear. We evaluated *MADRS-Chat*, a text-based chatbot designed to administer the MADRS interview and generate symptom-level scores using a clinician-trained scoring model. In a prospective observational study of psychiatric in- and outpatients, we assessed patient and clinician acceptability, feasibility of chatbot-administered interviews, and agreement between chatbot-derived MADRS scores, clinician-rated MADRS scores, and patient self-rated MADRS-S scores. *MADRS-Chat* was well accepted by both patients and clinicians, and chatbot-derived scores showed meaningful agreement with human ratings. Qualitative feedback suggested that the anonymous, written format facilitated more open and detailed disclosure of depressive symptoms. However, chatbot-administered interviews systematically yielded higher severity scores than clinician ratings, indicating that conversational format and perceived anonymity may shift patient response style. This bias was substantially reduced through lightweight post-hoc calibration. These findings suggest that conversational AI can support structured psychiatric assessment in real-world clinical care when deployed within clearly defined clinical boundaries. They also show that the mode of symptom elicitation can itself influence reported severity, underscoring the need for explicit validation, calibration, and bias assessment before clinical implementation.

Introduction

Depression is frequently underdiagnosed due to non-specific presentations and underreporting, and insufficiently monitored because of limited consultation time, diagnostic complexity, and lack of routine measurement-based follow-up[1]. Structured clinical questionnaires, such as the Montgomery-Åsberg Depression Rating Scale (MADRS[2]), are well-established tools to assess depressive symptom severity[3]. However, they are time-intensive, require trained personnel, and are not routinely conducted across psychiatric settings, leading to variability in symptom documentation and treatment monitoring[4,5]. Recent advances in artificial intelligence (AI), particularly large language models (LLMs), offer new possibilities for improving access to standardized assessments through conversational agents ("chatbots"). These AI-assisted systems may support the collection of symptom-related information in a scalable and resource-efficient manner[6,7]. While such tools have shown promise in simulating human-like interaction[8], their application in clinical settings raises important methodological questions regarding validity and acceptabiliy[9], underscoring the need for clinically grounded implementations.

In a recent study, we demonstrated that a fine-tuned BERT (*MADRS-BERT*) was able to predict individual MADRS subscores with accuracies of up to 88% and a mean absolute error (MAE) as low as 0.7 across items when compared to expert clinical ratings[10]. This symptom-level modeling offered an interpretable and scalable framework to assess depressive severity from structured dialogue. However, *MADRS-BERT* operated *post-hoc* on existing transcripts and did not participate in real-time assessment. Building on this foundation, we now evaluate whether a text-based LLM conversational agent (CA) can administer the MADRS interview directly to psychiatric patients[7]. Drawing on established technology acceptance frameworks[11,12], emphasizing perceived usefulness, ease of use, and trust as key determinants of adoption, we examined the feasibility, acceptability, and clinical agreement of this system[13]. This system engages directly with users, enabling interactive, real-time collection of depressive symptomatology, and investigation of how conversational formats may influence symptom expression.

Methods

This study evaluated the feasibility, acceptability, and clinical validity of *MADRS-Chat*, a structured LLM-assisted conversational agent (CA) designed to guide patients through the nine MADRS items while generating item-level scores using the previously developed *MADRS-BERT*[10] model. As *MADRS-Chat* directly builds on *MADRS-BERT*, we first externally validated the scoring model by assessing agreement between model-generated and clinician-assigned scores. We then examined the feasibility and acceptability of *MADRS-Chat* and its agreement with clinician ratings based on the chatbot interview transcripts (Fig. 1A). The study was conducted as a single-arm, exploratory, prospective observational study and was conducted in a real-world psychiatric clinical setting.

Ethical considerations and participants

The study was conducted at the Psychiatric University Hospital Zurich, Switzerland, recruiting psychiatric inpatients and outpatients. The study was reviewed and approved by the Kantonale Ethikkommission Zürich (Cantonal Committee Zurich), Gesundheitsdirektion des Kantons Zürich, Switzerland (BASEC: 2025 – 01252), and was conducted according to the Declaration of Helsinki. All participants provided written informed consent. Inclusion criteria were age 18–65 years, fluency in (Swiss-) German, current psychiatric treatment (inpatient, outpatient, or day clinic), ability to provide informed consent, and clinically reported depressive symptoms or a diagnosis of depressive episode (ICD-10: F32.0–F32.2) or recurrent depressive disorder (F33.0–F33.2). Exclusion criteria included organic mental disorders (F01–F09), schizophrenia spectrum and other non-mood psychotic disorders (F20–F29), depressive episodes with psychotic symptoms (F32.3/F33.3), intellectual disability (F70–F79), critical illness, severe head injury, or any condition compromising safety or study compliance.

MADRS-BERT external validation

Structured MADRS interview data was collected from $N = 30$ psychiatric inpatients and pre-processed as described previously[10] (see Supplementary Material S1 and Supplementary Table 1). *MADRS-BERT*[10] was used to predict item-level scores (0–6). Eight clinicians (AC, NP, LK, BP, JGM, LA, TW, GK), blinded to the model predictions and to each other's ratings, assigned MADRS item scores based solely on the interview transcript. The reference score per item was defined as the mean clinician rating (Supplementary Table 2). Interrater reliability was calculated using intraclass correlation coefficients (ICC) and Krippendorff's alpha, evaluating both clinician-clinician agreement and model-clinician agreement.

MADRS-Chat Technical Details

MADRS-Chat is a structured LLM-assisted CA designed to guide psychiatric patients through a standardized MADRS-based assessment of depressive symptoms. The system follows a predefined rule-based dialogue structure with an architecturally separated design in which a generative model (https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct) guides the interaction, while symptom scoring were handled by *MADRS-BERT*[10]. Interviews were conducted on a laptop under human supervision in clinical settings (Fig. 1B; Supplementary Video). Further technical details are provided in Supplementary Material S2.

Acceptability and Feasibility

At the study visit, patients first completed a self-assessment using MADRS-S[3] and subsequently interacted with *MADRS-Chat*. Acceptability was assessed using structured self-report questionnaires completed by both patients and clinicians. Patients rated their experience using a six-items usability and satisfaction questionnaire adapted from established digital health frameworks and prior CA evaluations[6,11,12]. In addition, patients completed an adapted version of the six-item Working Alliance Inventory (WAI-6[14]). Items related to therapeutic progress were omitted. Optional qualitative feedback was collected from patients, focusing on positive aspects of the interaction and potential areas for improvement. Patients received CHF50 (~ 65US-Dollar) compensation.

Clinicians completed a six-item satisfaction questionnaire adapted from Heinz[6], focused on usability and perceived value of the CA in a psychiatric setting. Furthermore, clinicians evaluated transcript quality (appropriateness, factuality, empathy) using a seven-item Likert scale. Lastly, clinicians assigned MADRS item scores from the CA interview transcript using MADRS rating criteria based solely on the verbal content.

Feasibility was assessed through completion rates, interview duration, and total word count. All questionnaires are provided in Supplementary Tables 3–5.

Clinician Evaluation

To assess clinical agreement, *MADRS-Chat* derived scores were compared with clinician-rated MADRS and patient self-rated MADRS-S scores at both item and total levels using Pearson correlations, paired statistical tests, and Bland-Altman analyses.

Post-hoc calibration on conversational agent interview domain

Although *MADRS-Chat* demonstrated strong agreement with clinician ratings, systematic differences in absolute score levels were observed (see Results), with the CA tending to assign higher severity scores. This pattern was not observed during prior external validation on clinician-led interview transcripts, indicating a context-dependent shift associated with interview modality. The written conversational format elicited more detailed and expressive responses, likely altering the input distribution presented to the scoring model and reflecting a modality-specific covariate shift between clinician-led (oral) and CA-led (written) interviews. To correct for this systematic shift while preserving the underlying scoring model, a *post-hoc* calibration procedure was applied using clinician ratings as the reference standard. Calibration consisted of a global bias-correction adjustment to account for score inflation while minimally affecting relative differences across participants and items. This approach preserved rank-order information and reduced the risk of overfitting to the clinician-rated calibration sample.

Results

Study population

Thirty-two (32) patients were recruited. Two did not complete participation (one early discharge, one exclusion due to medical burden), resulting in a final sample of 30 patients. Of these, 26 were inpatients and four outpatients. The sample included $N = 21$ females and mean age was 31 ± 12 years (range 18–61). Detailed demographics are provided in Supplementary Table 6.

MADRS-BERT external validation

Eight clinicians independently rated the 30 MADRS interview transcripts. Mean total clinician-rated MADRS scores (mean = 16.3, SD = 8.6) and *MADRS-BERT* predictions (mean = 18.9, SD = 5.5) were comparable across items, with the model assigning slightly higher scores on the topic *sleep disturbances*, *loss of appetite*, *lassitude*, and *suicidal ideations*. Inter-clinician agreement across the nine MADRS items was high (mean ICC = 0.92). Agreement between *MADRS-BERT* and mean clinician ratings was slightly lower (mean ICC = 0.78), indicating that the model aligns closely but not identical with the clinical reference. Krippendorff's alpha mirrored this pattern, confirming good overall reliability between clinicians and the model. Reliability and agreement metrics are summarized in Supplementary Table 2. Mean item-level scores are shown in Fig. 1C; topic distributions are provided in Supplementary Figure S2.

Feasibility

Two participants did not complete all topics (7/9 and 5/9) due to technical issues. Mean interview duration was 48.2 ± 19.8 minutes (excluding the first three participants affected by a technical timing artifact). Topic-level analyses showed the longest discussions for *suicidal ideation* (mean = 8.4 ± 4.8min), and the shortest for *emotional numbness* (4.2 ± 2.0min). Patients engaged in 45.1 ± 4.7 conversational turns and produced a mean of 594.9 ± 364.0 words per

interview. Higher word counts were observed for emotionally salient topics such as *suicidal ideation* and *pessimistic thoughts* (84.4 ± 64.6 words), whereas more concrete symptom domains, such as *loss of appetite* (53.8 ± 44.4 words) elicited shorter responses (Table 1).

Table 1

Feasibility Measures. Feasibility outcomes include overall interview duration, and mean time spent per MADRS topic. User engagement is summarized by total word count per interview and mean word count per topic. All values are reported as mean ± standard deviation.

Topics	**Interview duration (min)**	**Word count**
	Mean time per topic (minutes)	***Per topic word count***
Reported sadness (*N* = 27)	4.31 ± 2.12	69.70 ± 48.77
Inner tension (*N* = 26)	5.09 ± 2.27	65.27 ± 49.70
Sleep disturbances (*N* = 27)	4.56 ± 3.11	59.63 ± 51.74
Loss of appetite (*N* = 27)	4.32 ± 2.65	54.70 ± 43.91
Difficulties concentrating (*N* = 26)	4.42 ± 2.36	55.14 ± 36.37
Lassitude (*N* = 26)	5.87 ± 3.04	70.29 ± 56.59
Emotional numbness (*N* = 26)	4.15 ± 1.95	48.14 ± 30.83
Pessimistic thoughts (*N* = 27)	6.32 ± 2.80	84.83 ± 63.54
Suicidal ideations (*N* = 27)	8.42 ± 4.75	104.31 ± 98.37
Total (N = 30)	**48.21 ± 19.79**	**594.93 ± 364.01**

Acceptability

User Acceptance and Session Alliance

Patients reported overall positive acceptance and a strong perceived alliance with MADRS-Chat (Fig. 2A). Ease of use was rated highly (mean = 5.90, SD = 1.2; 83% agreement). Ratings of usability (mean = 4.3, SD = 1.58; 47% agreement) and enjoyment (mean = 4.23, SD = 1.79; 43% agreement) were more mixed. Alliance-related dimensions were strongly endorsed, particularly respect (mean = 6.67, SD = 0.84; 97% agreement) and appreciation (mean = 5.33, SD = 1.69; 80% agreement). Overall satisfaction was moderate (mean = 4.53, SD = 1.59; 50% agreement).

User Satisfaction

Patient satisfaction showed a differentiated pattern across interaction features (Fig. 2B). Usability was rated highly, including ease of learning (mean = 6.4, SD = 0.77; 97% agreement) and intuitiveness (mean = 5.83, SD = 1.21; 83% agreement). Interaction-related aspects were rated lower. Perceived helpfulness was mixed (mean = 4.1, SD = 1.56; 37% agreement), and conversational naturalness received the lowest ratings, particularly regarding repetition (mean = 3.33, SD = 1.84; 63% disagreement) and resemblance to a therapist interaction (mean = 3.03, SD = 1.88; 67% disagreement). Willingness for personal use was moderate (mean = 4.4, SD = 1.83; 50% agreement).

Clinicians reported consistently positive evaluations (Fig. 2C). Ease of learning (mean = 6.33, SD = 0.66) and intuitiveness (mean = 5.8, SD = 0.96) received unanimous agreement. Clinical usefulness was rated highly (mean = 5.77, SD = 1.38; 80% agreement). Most clinicians agreed that the system retrieved relevant information (mean = 4.83, SD = 1.09; 63% agreement) and would consider using it in practice (mean = 5.73, SD = 1.62; 77% agreement). Overall satisfaction was high (mean = 5.23, SD = 1.19; 70% agreement).

Clinical Validity

Clinician evaluations indicated high perceived response quality across appropriateness (mean = 5.87, SD = 0.81), factual accuracy (mean = 5.74, SD = 0.79), and empathy (mean = 5.73, SD = 1.05) (Fig. 3A). Most ratings were within the agree range (90–94%). Item-level distributions are shown in Supplementary Figure S3. *MADRS-Chat* total scores (mean = 35.4, SD = 6.4) correlated with clinician-rated (mean = 23.1, SD = 8.7; $r = 0.76$, $P < 0.001$) and self-rated MADRS scores (mean = 25.2, SD = 8.2; $r = 0.37$, $P = 0.042$) (Fig. 3B). However, *MADRS-Chat* produced significantly higher total scores than both clinician and self-ratings (Wilcoxon signed-rank, $P < 0.001$), whereas clinician and self-ratings did not differ (Fig. 3C).

Bland-Altman analyses showed a systematic positive bias of *MADRS-Chat* relative to clinician ratings (mean bias = + 12.3; limits 1.16 to 23.44) and self-ratings (mean bias = + 10.17; limits − 6.10 to 26.44). In contrast, clinician and self-ratings demonstrated minimal bias (− 2.13; limits − 19.46 to 15.19). Following *post-hoc* bias correction (Methods), mean *MADRS-Chat* scores (mean = 26.6, SD = 6.1) no longer differed significantly from clinician or self-ratings. Bland-Altman analysis confirmed substantial reduction of systematic bias relative to clinician ratings (mean bias = + 3.53; limits − 7.30 to 14.36) and self-ratings (mean bias = + 1.40; limits − 14.12 to 16.92) (Bland-Altman plots in Supplementary Figures S4 and S5). Table 2 presents item-level and total MADRS comparisons between clinician ratings and *MADRS-Chat*, and bias-corrected *MADRS-Chat*, including item-level *P*-values adjusted for multiple comparisons using false discovery rate. Analyses involving patient self-ratings are provided in the Supplementary Tables 7 and 8.

Table 2

Item-level and total-score comparisons between MADRS-Chat and clinician-rated MADRS. Values are reported as mean ± standard deviation (SD) for clinici *Chat*, and bias-corrected *MADRS-Chat*. Associations between *MADRS-Chat* and clinician ratings are quantified using Pearson correlation coefficients (*r*). Dif conversational agent ratings and clinician ratigns were assessed using Wilcoxon signed-rank tests, with test statistics (W) and associated *P*-values. For iten values were adjusted for multiple comparisons using the Benjamini–Hochberg false discovery rate (FDR) procedure.

Total	Suicidal ideations	Pessimistic thoughts	Emotional numbness	Lassitude	Difficulties concentrating	Loss of appetite	Sleep disturbances	Inner tension	Reported sadess	MADF Item
30	29	30	28	28	29	30	30	30	30	**N**
23.07 (± 8.7)	2.21 (± 1.9)	2.53 (± 1.3)	2.57 (± 1.6)	2.86 (± 1.4)	2.79 (± 1.3)	2.10 (± 2.0)	2.33 (± 1.3)	3.03 (± 0.9)	3.17 (± 1.3)	**Clinic**
35.37 (± 6.4)	4.03 (± 1.4)	4.27 (± 0.9)	3.54 (± 1.2)	4.36 (± 1.2)	4.03 (± 1.2)	3.80 (± 1.4)	3.97 (± 1.2)	4.10 (± 1.0)	6.07 (± 1.1)	**MADF Chat**
26.60 (± 6.1)	3.03 (± 1.4)	3.27 (± 0.9)	2.53 (± 1.2)	3.36 (± 1.2)	3.03 (± 1.2)	2.8 (± 1.4)	2.97 (± 1.2)	3.10 (± 1.0)	3.10 (± 1.0)	**MADF Chat (bias-corr)**
0.76 (P = < 0.001)	0.49 (P_{adj}=0.007)	0.43 (P_{adj}=0.018)	0.49 (P_{adj}=0.008)	0.53 (P_{adj}=0.004)	0.73 (P_{adj}<0.001)	0.71 (P_{adj}<0.001)	0.42 (P_{adj}=0.24)	0.44 (P_{adj}=0.21)	0.22 (P_{adj}=0.23)	**Pears r**
W = 0, P < 0.001	W = 26, P_{adj} <0.001	W = 4.5, P_{adj} <0.001	W = 18.5, P_{adj} =0.004	W = 6, P_{adj} <0.001	W = 6.5, P_{adj} <0.001	W = 4.5, P_{adj} <0.001	W = 9, P_{adj} <0.001	W = 14, P_{adj} <0.001	W = 39.5, P_{adj}=0.004	**Clinic vs. MADF Chat**
W = 58, P = 0.003	W = 55, P_{adj} =0.05	W = 48, P_{adj} =0.04	W = 136.5, P_{adj} =0.96	W = 94, P_{adj} =0.09	W = 57, P_{adj} =0.27	W = 55.5, P_{adj} =0.05	W = 73.5, P_{adj} =0.06	W = 87, P_{adj} =0.82	W = 122.5, P_{adj} =0.80	**Clinic vs. MADF Chat (bias-corr)**

Semi-structured qualitative feedback

In open-ended feedback, anonymity emerged as a central theme. Many participants reported that the written format allowed them to speak more openly about sensitive experiences, facilitating honesty and disclosure without fear of judgment. Writing rather than speaking was described as helpful for structuring thoughts and reflecting on emotions. Patients further emphasized the CA's empathetic and responsive interaction style, describing a sense of being heard. The polite tone, reflective summaries, and targeted follow-up questions were frequently contrasted with static questionnaires and experienced as more engaging for self-reflection. Additional positive aspects included clear language, robustness to typos, conversational continuity, and the inclusion of crisis hotline information, which was perceived as reassuring.

Patients also identified areas for improvement. The most frequently mentioned issue was repetition, which some participants found tiring or disruptive to conversational flow. Some participants reported contextual understanding difficulties when responses contained multiple elements, as well as occasional response latency.

Discussion

In this study, we evaluated the feasibility, acceptability, and clinical agreement of *MADRS-Chat*, an LLM-based CA administering the MADRS in a real-world psychiatric setting. Two main findings emerged. First, *MADRS-Chat* was feasible and well accepted. Patients reported high ease of use and perceived alliance, particularly feeling treated respectfully, and clinicians rated responses as appropriate, accurate, and empathetic. Second, *MADRS-Chat* scores showed meaningful agreement with clinician and self-ratings but demonstrated a systematic positive bias in absolute severity. This bias was not observed in prior validation on clinician-led transcripts, suggesting a modality-dependent effect. Qualitative feedback indicated that the written, anonymous format elicited more detailed and emotionally expressive disclosures. Post-hoc calibration corrected this shift while preserving relative differences, supporting the interpretation that the discrepancy reflected calibration rather than loss of validity. Together, these results position *MADRS-Chat* as a complementary tool for structured symptom assessment, particularly for standardized monitoring or pre-visit screening.

Most prior conversational AI systems in mental health have focused on intervention-oriented applications[6,15,16], rather than structured clinical assessment[17,18]. Across these studies, users generally report high ease of use, perceived support, and appreciation for the low-threshold nature of chatbot interactions, while recurrent barriers include repetition, interaction fatigue, response latency, and declining engagement over time. Our findings align closely with this literature. *MADRS-Chat* demonstrated high acceptability and perceived alliance among patients, as well as strong clinician endorsement of response quality. At the same time, open-ended feedback highlighted similar limitations, particularly repetitiveness and occasional contextual misunderstandings. Importantly, these constraints may have different implications depending on the system's purpose. In intervention-focused systems, sustained engagement and perceived therapeutic benefit are central, making repetitiveness particularly detrimental.[19] In contrast, for structured assessment and monitoring, interactions are time-limited and goal-directed, potentially making such limitations more tolerable.

In terms of assessment design, many prior approaches have adapted validated questionnaires[20] into structured conversational formats. While these studies demonstrate that conversational delivery can preserve overall psychometric properties, amplification of symptom severity has been observed in some CA-based assessments[18,21], underscoring that conversational formats may introduce modality-dependent effects requiring calibration. Evidence from intervention-focused CA research suggests that effectiveness depends less on whether systems are rule-based or LLM-based, and more on how structure, constraint, and interactional flexibility are combined[22]. Together, these findings support hybrid architectures that preserve clinical structure while allowing conversational flexibility[23], aligning closely with the design of *MADRS-Chat*, where conversational guidance and symptom scoring were architecturally separated. Together, these studies support a growing consensus that AI systems for psychiatric assessment benefit most from hybrid architectures, i.e., using LLM-based conversational interfaces for data collection while leaving scoring and potential interpretation to a clinically grounded, domain-specific model.

Although *MADRS-Chat* demonstrated strong agreement with human ratings, the conversational format introduced a systematic positive bias in absolute severity scores. Importantly, this bias was not observed during prior validation on clinician-led interview transcripts, suggesting that it was not intrinsic to the scoring model but emerged from the conversational assessment modality itself. The written, interactive format likely altered the linguistic input distribution, constituting a modality-specific covariate shift[24]. Notably, this interaction-induced bias was effectively corrected through a lightweight *post-hoc* calibration, aligning absolute score levels while preserving rank-order correlations. This indicates that the discrepancy reflected calibration rather than loss of clinical validity and highlights the necessity of bias-aware validation when conversational systems are used for symptom measurement[25]. Prior work in medical AI has shown that even when LLMs align closely with guidelines or expert judgments, they may exhibit stable directional distortions driven by contextual framing or interaction dynamics[25,26]. More broadly, risks in LLM-based mental health systems appear to increase when interaction, reasoning, and decision-making are tightly coupled[27]. In line with this, *MADRS-Chat* separates the conversational model from the scoring model[10], thereby constraining *where* and *how* bias can arise. Within this constrained scope, bias becomes measurable, interpretable, and correctable, supporting the view that to date conversational AI may be particularly suited for standardized symptom assessment and monitoring.

Lastly, it is important to consider *why* this systematic bias emerged. Our findings suggest that the conversational format itself shaped symptom disclosure. Compared with clinician-led interviews, patients produced more detailed, and emotionally expressive responses. This pattern aligns with qualitative evidence indicating that conversational AI is often experienced as emotionally safe and disinhibiting, partly due to perceived anonymity[28]. Clinically, such increased openness provides a plausible explanation for the observed inflation in severity scores. Notably, this interaction dynamic resembles elements of a digital therapeutic alliance[29], including perceived empathy, respect, and feeling heard, even in a system explicitly designed for assessment rather than therapy. Similar disclosure-enhancing effects have been reported in prior CA-based assessment work, where anonymity and scalability facilitated engagement with sensitive mental health topics[17].

Several limitations should be considered. First, the study was primarily designed to assess feasibility and acceptability, and the sample size was not powered for definitive conclusions regarding clinical validity. Second, detailed sociodemographic variables were not systematically collected. Although the clinical sample was heterogeneous, the absence of structured data on socioeconomic background, ethnicity, and gender or sexual minority status limits conclusions regarding equity and differential performance. In addition, participation may have introduced selection bias, as individuals willing to enroll could be more receptive to AI-based tools; however, financial compensation may have mitigated this effect. Third, the evaluation was cross-sectional and limited to a single session. The stability, sensitivity to change, and potential drift of symptom estimates during longitudinal use remain unknown and should be examined in future monitoring studies. Fourth, both the CA and transcript-based clinician ratings relied exclusively on written content, whereas several MADRS items traditionally incorporate non-verbal cues such as facial expression, prosody, and psychomotor behavior. Consequently, in-person clinician ratings might differ from transcript-based assessments. Future work should explore multimodal extensions. Finally, calibrating *MADRS-Chat* to clinician ratings enhances comparability with current clinical standards but does not eliminate potential bias inherent to clinician judgment itself.

In conclusion, this work positions *MADRS-Chat* within a growing but still underdeveloped space between traditional clinician-led assessment and fully autonomous mental-health CAs.

Our findings suggest that structured symptom assessment represents a clinically tractable application of conversational AI in psychiatry. However, conversational assessment is not a neutral translation of existing formats: the interaction itself shapes symptom disclosure. Increased openness may enhance sensitivity but can also systematically shift severity estimates, underscoring the need for bias-aware validation and calibration. By separating conversational interaction from symptom scoring, MADRS-Chat provides a pragmatic framework for integrating LLMs into psychiatric workflows while preserving transparency and clinical oversight.

Declarations

Competing Interests

SW and TK are affiliated with the Centre for Digital Health Interventions, a joint initiative of the Institute for Implementation Science in Health Care, University of Zurich; the Department of Management, Technology, and Economics at ETH Zurich; and the Institute of Technology Management and School of Medicine at the University of St.Gallen. CDHI is funded in part by CSS, a Swiss health insurer; Mavie Next, an Austrian care provider; and MTIP, a Swiss digital health investor. TK is also co-founder of Pathmate Technologies, a university spin-off company that creates and delivers digital clinical pathways. However, CSS, Mavie Next, MTIP, and Pathmate Technologies were not involved in this research. TK has neither shares nor any formal role with Pathmate Technologies. SO hs co-founder of DeepPsy (deeppsy.io) a start-up that enables biomarker- and data-driven psychiatry. All other authors report no competing interests.

Author Contribution

SW: Conceptualization, Methodology, Software, Validation, Formal analysis, Data Curation, Writing – Original Draft, Visualization; BP: Data Curation, Data Acquisition; NP: Data Curation, AC: Data Curation, LA: Data Curation, LK: Data Curation, Data Acquisition, JGM: Data Curation, TW: Data Curation, GK: Data Curation, TK: Methodology, Funding acquisition; BK: Funding acquisition, Resources SO: Conceptualization, Methodology, Supervision, Resources, Funding acquisition

Acknowledgement

This work was supported by the Swiss National Science Foundation (SNF Grant 205913). We thank Manuel Schegg, Schwarna Maria Raja and Vivienne Walser for their help in patient screening and recruitment. We want to thank Matthias Hirsch-Hoffmann for his valuable IT support, including the setup of the virtual machine, deployment assistance, and generating the secure access infrastructure for the conversational agent implementation. We want to thank Yamina Ehrt, Diana Pintea and Lion Comfort for their help in the clinician ratings of the conversational agent-derived interviews.

Data Availability

Patient interview transcripts cannot be shared due to patient privacy but can be shared upon reasonable request.

Code availability

The fully python-based code for the conversational agent can be shared upon reasonable request. The conversational agent is available in German, English, French, Italian, Portuguese, and Spanish.

Funding

This work was supported by the Swiss National Science Foundation (SNF [CRSII5 205913]).

References

1. Malhi, G. S., Mann, J. J. & Depression *Lancet* **392**, 2299–2312 (2018).
2. Montgomery, S. A. & Åsberg, M. A. New Depression Scale Designed to be Sensitive to Change. *Br. J. Psychiatry*. **134**, 382–389 (1979).
3. Montgomery, S. A. & Åsberg, M. A. New Depression Scale Designed to be Sensitive to Change. *Br. J. Psychiatry*. **134**, 382–389 (1979).
4. Geijer, J., Baigi, A. & Aiff, H. Inter-rater reliability among psychiatrists when assessing depression according to the Montgomery–Åsberg Depression Rating Scale. *Nord J. Psychiatry*. **75**, 607–613 (2021).
5. König, H., König, H. H. & Konnopka, A. The excess costs of depression: a systematic review and meta-analysis. *Epidemiol. Psychiatr Sci.* **29**, e30 (2020).
6. Heinz, M. V. et al. Randomized Trial of a Generative AI Chatbot for Mental Health Treatment. *NEJM AI* **2**, (2025).
7. Vaidyam, A. N., Wisniewski, H., Halamka, J. D., Kashavan, M. S. & Torous, J. B. Chatbots and Conversational Agents in Mental Health: A Review of the Psychiatric Landscape. *Can. J. Psychiatry*. **64**, 456–464 (2019).
8. Shah, J. et al. Development and usability testing of a chatbot to promote mental health services use among individuals with eating disorders following screening. *Int. J. Eat. Disord*. **55**, 1229–1244 (2022).
9. Dosovitsky, G. & Bunge, E. L. Bonding With Bot: User Feedback on a Chatbot for Social Isolation. *Front. Digit. Health*. **3**, 735053 (2021).
10. Weber, S. et al. Using a fine-tuned large language model for symptom-based depression evaluation. *Npj Digit. Med.* **8**, 598 (2025).
11. Davis, F. D. & Perceived Usefulness Perceived Ease of Use, and User Acceptance of Information Technology. *MIS Q.* **13**, 319 (1989).
12. Venkatesh, T. & Xu Consumer Acceptance and Use of Information Technology: Extending the Unified Theory of Acceptance and Use of Technology. *MIS Q.* **36**, 157 (2012).
13. Hong, G., Smith, M. & Lin, S. The AI Will See You Now: Feasibility and Acceptability of a Conversational AI Medical Interviewing System. *JMIR Form. Res.* **6**, e37028 (2022).
14. Falkenström, F., Hatcher, R. L., Skjulsvik, T., Larsson, M. H. & Holmqvist, R. Development and validation of a 6-item working alliance questionnaire for repeated administrations during psychotherapy. *Psychol. Assess.* **27**, 169–183 (2015).
15. Feng, Y. et al. Effectiveness of AI-Driven Conversational Agents in Improving Mental Health Among Young People: Systematic Review and Meta-Analysis. *J. Med. Internet Res.* **27**, e69639 (2025).
16. Davis, A. et al. Evaluating the Feasibility and Acceptability of a Mobile Mental Health Intervention for Adolescent Depression and Anxiety. *Clin. Pediatr. (Phila)*. **65**, 53–62 (2026).
17. Hungerbuehler, I., Daley, K., Cavanagh, K., Garcia Claro, H. & Kapps, M. Chatbot-Based Assessment of Employees' Mental Health: Design Process and Pilot Implementation. *JMIR Form. Res.* **5**, e21678 (2021).
18. Jin, Z., Hu, J., Bi, D., Zhao, K. & Yu, H. Evaluating the Efficacy of AI-Based Interactive Assessments Using Large Language Models for Depression Screening: Development and Usability Study. *JMIR Form. Res.* **10**, e78401 (2026).

19. Aschentrup, L. et al. Effectiveness of gamified digital interventions in mental health prevention and health promotion among adults: a scoping review. *BMC Public. Health*. **24**, 69 (2024).
20. Luo, X. et al. AI in Medical Questionnaires: Scoping Review. *J. Med. Internet Res.* **27**, e72398 (2025).
21. Teferra, B. G. et al. Leveraging large language models for automated depression screening. *PLOS Digit. Health*. **4**, e0000943 (2025).
22. Du, Q., Ren, Y., Meng, Z., He, H. & Meng, S. The Efficacy of Rule-Based Versus Large Language Model-Based Chatbots in Alleviating Symptoms of Depression and Anxiety: Systematic Review and Meta-Analysis. *J. Med. Internet Res.* **27**, e78186–e78186 (2025).
23. Thirunavukarasu, A. J. et al. Large language models in medicine. *Nat. Med.* **29**, 1930–1940 (2023).
24. Chang, Y., Liu, Y. C., Huang, S. S. & Hsu, W. Y. Assessing bias in AI-driven psychiatric recommendations: A comparative cross-sectional study of chatbot-classified and CANMAT 2023 guideline for adjunctive therapy in difficult-to-treat depression. *Psychiatry Res.* **348**, 116501 (2025).
25. Schmidgall, S. et al. Evaluation and mitigation of cognitive biases in medical language models. *Npj Digit. Med.* **7**, 295 (2024).
26. Gallegos, I. O. et al. Bias and Fairness in Large Language Models: A Survey. *Comput. Linguist*. **50**, 1097–1179 (2024).
27. Ke, Y. et al. Mitigating Cognitive Biases in Clinical Decision-Making Through Multi-Agent Conversations Using Large Language Models: Simulation Study. *J. Med. Internet Res.* **26**, e59439 (2024).
28. Siddals, S., Torous, J. & Coxon, A. It happened to be the perfect thing: experiences of generative AI chatbots for mental health. *Npj Ment Health Res.* **3**, 48 (2024).
29. Malouin-Lachance, A., Capolupo, J., Laplante, C. & Hudon, A. Does the Digital Therapeutic Alliance Exist? Integrative Review. *JMIR Ment Health*. **12**, e69294–e69294 (2025).

Figures

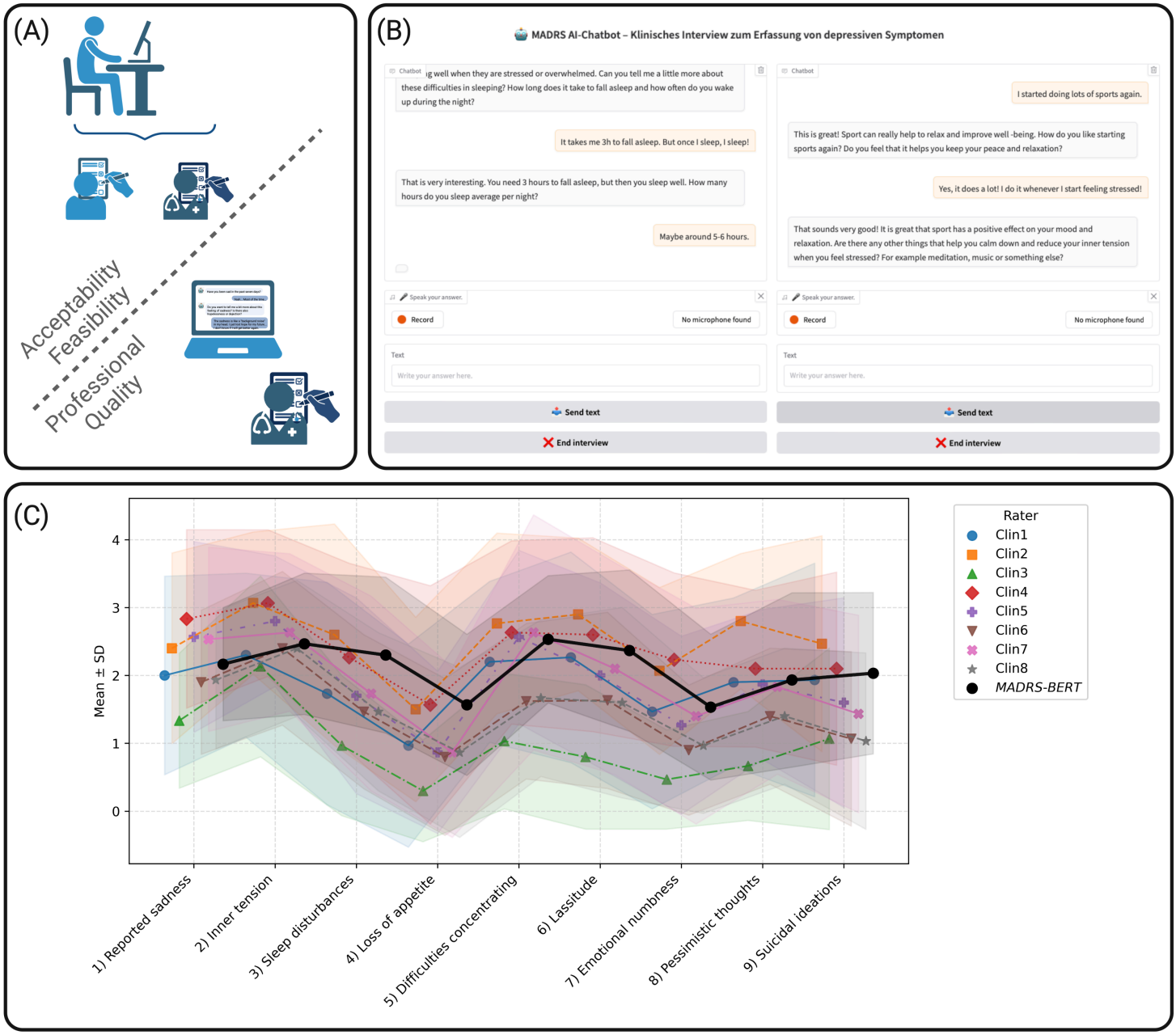


Figure 1

Study overview, conversational agent interface, and external validation of *MADRS-BERT*. (A) Schematic of the study workflow. Patients completed a structured MADRS interview using *MADRS-Chat*; feasibility and acceptability were assessed from both patient and clinician perspectives, and clinicians additionally evaluated the professional quality of conversational agent generated interviews based on transcript review. **(B)** Example screenshots of the *MADRS-Chat*interface illustrating the guided assessment (e.g., items on sleep and inner tension). A video can be found in Supplementary Material. **(C)** External validation of MADRS-BERT: mean MADRS item scores across topics shown as mean ± standard deviation (SD), with each line representing one rater (•black = *MADRS-BERT*; colored = clinicians) and shaded bands indicating ±1 SD.

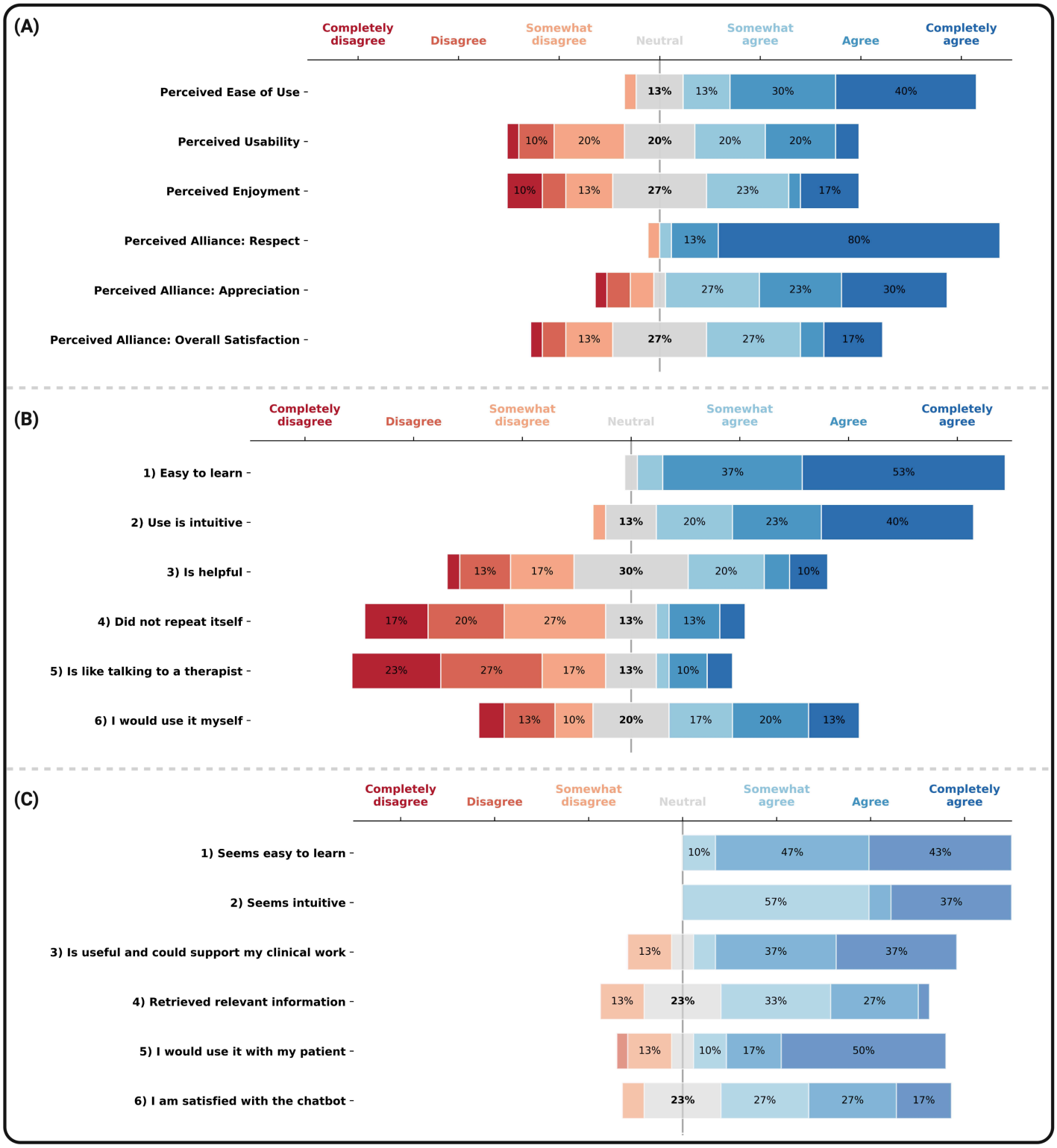


Figure 2

User acceptance, session alliance, and satisfaction with *MADRS-Chat*. Diverging stacked Likert distributions summarizing acceptance and satisfaction ratings on a 7-point Likert scale. **(A)**Patient-reported user acceptance and session alliance, capturing perceived ease of use, usability, enjoyment, and alliance with the conversational agent. **(B)** Patient satisfaction with *MADRS-Chat*, including perceived helpfulness, interaction quality, and willingness to use the system. **(C)** Clinician satisfaction ratings evaluating the conversational agent from a professional perspective, including perceived usefulness, information quality, and overall satisfaction. Percentages indicate the proportion of responses per Likert category (1–7), with negative, neutral, and positive response options displayed from left to right (only values above 10% are displayed).

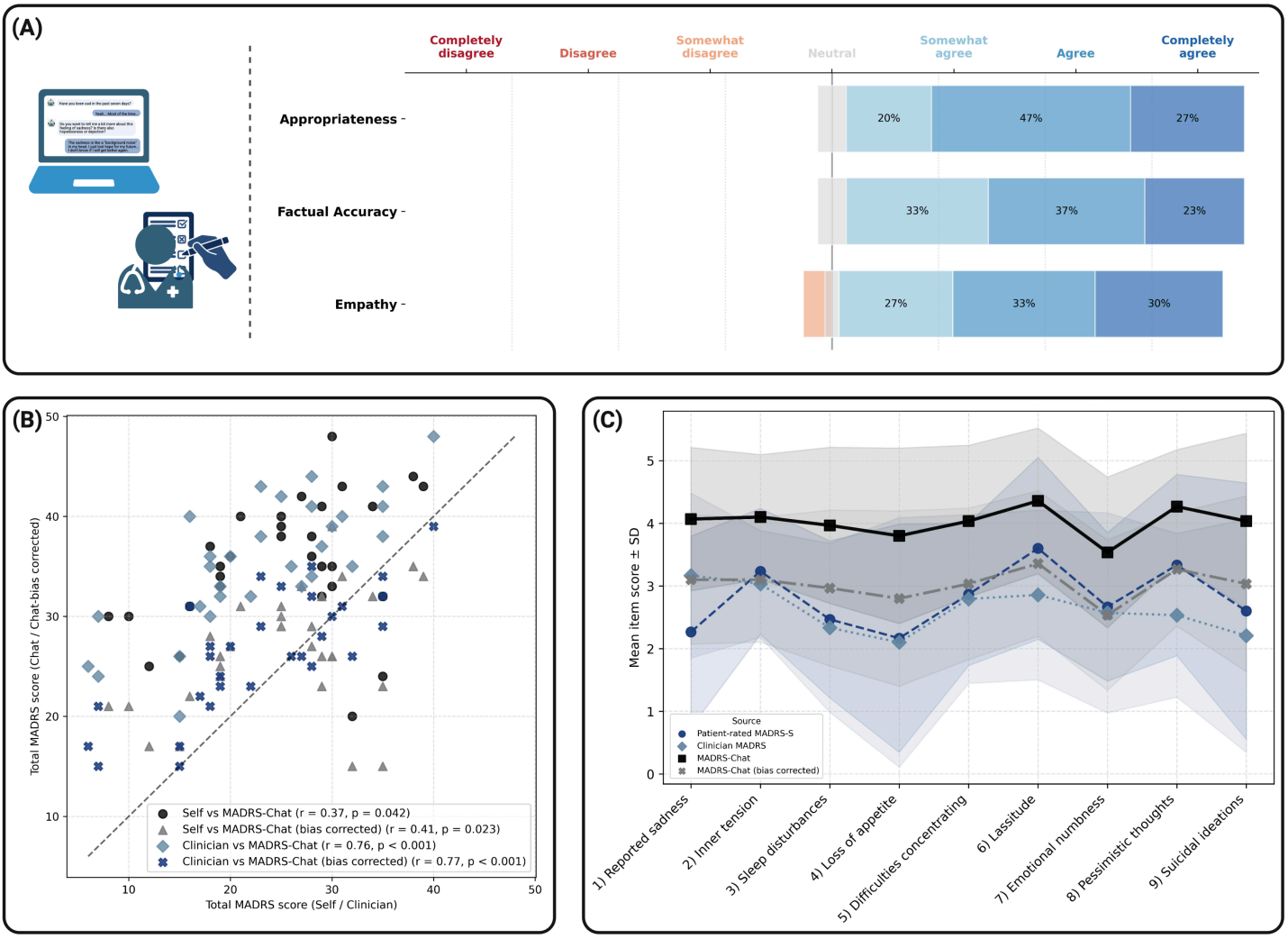


Figure 3

Clinical validity of *MADRS-Chat* compared with self- and clinician-rated MADRS. (A) Clinician-rated evaluation of *MADRS-Chat* responses across all MADRS items, summarised over the dimensions *Appropriateness*, *Factual Accuracy*, and *Empathy*. Percentages reflect the proportion of ratings per Likert category (1–7), aggregated across all MADRS topics (only values above 10% are displayed). Item-level (per MADRS topic) Likert distributions are provided in Supplementary Figure 3A-C. **(B)** Scatter plots illustrating agreement between *MADRS-Chat* and human ratings. Panels show pairwise comparisons between self-rated MADRS and *MADRS-Chat*, self-rated MADRS and bias-corrected *MADRS-Chat*, clinician-rated MADRS and *MADRS-Chat*, and clinician-rated MADRS and bias-corrected *MADRS-Chat*. Each point represents one subject; solid lines indicate the line of identity. **(C)** Mean MADRS item scores across subjects for self-rated MADRS, clinician-rated MADRS, *MADRS-Chat*, and bias-corrected *MADRS-Chat*. Lines indicate mean item scores per MADRS topic, with shaded bands representing ± standard deviation (SD).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryMaterialMADRSChatDemoEN.mp4
- SupplementaryMaterialMADRSChatWeber.docx
- SupplementaryMaterialMADRSChatDemoDE.mp4