Journal of Affective Disorders 389 (2025) 119760

Contents lists available at ScienceDirect

Journal of Affective Disorders

journal homepage: www.elsevier.com/locate/jad

Research paper

Depression intervention using AI chatbots with social cues: a randomized trial of effectiveness

Shuo Xu [a,*], Tiancong Ma [b]

[a] School of Design, South China University of Technology, Guangzhou Higher Education Mega Center, Panyu District, 510006 Guangzhou, China
[b] Department of Biomedical Engineering, The Hong Kong Polytechnic University, Hong Kong

ARTICLE INFO

Keywords:
Depression
Mental health interventions
Digital mental health
Social cues
Chatbot
mHealth

ABSTRACT

Background: Depression is a serious problem among college students, and chatbots are a popular intervention tool. Social cues are used in chatbot design, but their effectiveness in depression treatment remains to be verified. This study aimed to compare the effects of chatbots with high-social-cue (HSC) versus low-social-cue (LSC) designs on depressive symptoms.

Methods: An open-label randomized controlled trial was conducted over 16 weeks. Eighty-four college students with baseline Patient Health Questionnaire-9 (PHQ-9) scores ≥9 were randomly assigned to either an HSC group (text + voice + animations) or LSC group (text-only). Clinical outcomes, including PHQ-9, Generalized Anxiety Disorder scale (GAD-7), and Positive and Negative Affect Schedule (PANAS) scores, were collected every 4 weeks. Secondary measures included user satisfaction (Client Satisfaction Questionnaire-8, CSQ-8), therapeutic alliance (Working Alliance Inventory-Short Revised, WAI-SR), and self-reported adherence.

Results: Baseline characteristics did not differ significantly between groups. Intention-to-treat analysis revealed that the HSC group achieved greater reductions in PHQ-9 ($d = 0.63$, $P < 0.01$) and GAD-7 ($d = 0.50$, $P = 0.003$) scores compared to the LSC group. The HSC group also demonstrated higher adherence rates ($d = 0.82$, $P < 0.01$), CSQ-8 ($P = 0.02$), and WAI-SR scores ($P < 0.001$). LSC group.

Conclusion: Chatbots with high-social-cue designs significantly outperformed text-only versions in alleviating depression and anxiety, while enhancing adherence, satisfaction, and therapeutic alliance.

1. Introduction

College students have been more likely than the general public to experience depression in recent years, and this has emerged as a serious public health concern requiring immediate attention (Lei et al., 2016). Following the coronavirus disease 2019 (COVID-19) outbreak in 2020, the pressure and challenges faced by college students in preventing and treating depression escalated significantly (Liu et al., 2022a). Numerous studies have established links between depression and suicide attempts (Hawton et al., 2013), impaired academic performance (Andrews and Wilding, 2004; Hysenbegasi et al., 2005), reduced quality of life (Zhong et al., 2019), and interpersonal difficulties (Lee et al., 2021). Although most universities offer free mental health services, many students avoid seeking help due to gaps in psychological education and cognitive biases (Neathery et al., 2020; Ratnayake and Hyde, 2019). Research indicates that university students exhibit lower rates of engagement with in-person psychological services, primarily due to stigma associated with reduced anonymity (Gulliver et al., 2010; Reavley et al., 2014). This reluctance is further compounded by findings suggesting that offline interventions may inadvertently exacerbate anxiety through internalized shame (Jorm et al., 2007). Additionally, traditional interventions face logistical barriers, including reliance on limited professional resources and challenges in delivering timely care (Renton et al., 2014). Consequently, organizations often prioritize brief counseling sessions, which may inadequately address long-term mental health needs (Lee and Jung, 2018).

In response to these challenges, researchers have increasingly turned to Internet-based Psychological Interventions (IPIs), including chatbots (Andersson, 2016; Wang et al., 2018). Advances in mobile and internet technologies have enabled the deployment of IPIs for treating depression in college students, with completed studies demonstrating their feasibility (Palma-Gómez et al., 2020; Herrero et al., 2019; Bolinski et al.,

* Corresponding author at: School of Design, B11 Building, University Town Campus, South China University of Technology, Guangzhou Higher Education Mega Center, Panyu District, 510006 Guangzhou, China.
E-mail address: sdxushuo@mail.scut.edu.cn (S. Xu).

https://doi.org/10.1016/j.jad.2025.119760
Received 28 August 2024; Received in revised form 21 April 2025; Accepted 21 June 2025
Available online 23 June 2025

2018; Musiat et al., 2019; Harrer et al., 2018). For example, Harrer et al. conducted a randomized controlled trial of an app-based stress intervention, reporting significant reductions in anxiety and depression (Harrer et al., 2018). Ongoing research protocols, such as those by Palma-Gómez et al. and Musiat et al., are evaluating transdiagnostic and resilience-focused IPIs across international student populations (Palma-Gómez et al., 2020; Musiat et al., 2019). These interventions often integrate cognitive behavioral therapy (CBT) (Cook et al., 2019; Fitzsimmons-Craft et al., 2021), enabling users to conduct private, self-guided interventions at their convenience.

Artificial intelligence (AI)-driven chatbots have shown particular promise in IPI delivery. Over the past decade, AI has advanced rapidly, enabling chatbots like Woebot (Fitzpatrick et al., 2017), Tess (Fulmer et al., 2018), and Wysa (Inkster et al., 2018) to address mental health concerns with efficacy comparable to human therapists (Torous et al., 2020; Gratzer and Goldbloom, 2020; Vaidyam et al., 2019). Notably, therapeutic chatbots have demonstrated higher self-disclosure rates (Lee et al., 2020) and adherence (Vaidyam et al., 2019) than traditional IPIs. While chatbots have been applied to diverse populations—including adolescents (Huang et al., 2015), older adults (Ryu et al., 2020), and clinical patients (Greer et al., 2019)—their use for depression in college students remains underexplored (Vaidyam et al., 2019; Abd-Alrazaq et al., 2020). Critically, the design of these chatbots, particularly the incorporation of social cues (e.g., voice modulation, facial expressions), has emerged as a key factor influencing user engagement and outcomes (Agrawal and Williams, 2017; Lopez et al., 2017; Bailenson et al., 2001; Li, 2013).

Early studies highlighted the importance of social cues in human-robot interaction.For instance, an exercise-promoting chatbot demonstrated the formation of therapeutic alliances through human-like interactions (Bickmore et al., 2005), while embodied conversational agents (ECAs)—defined as computer-generated characters using verbal/nonverbal cues (Burton et al., 2016)—proved effective in psychological therapy via randomized trials (Burton et al., 2016; Suganuma et al., 2018). Theoretical frameworks, such as the media equation theory (Martin, 1997) and social agent theory (Atkinson et al., 2005), posit that even basic social cues (e.g., vocal tone, body language) can enhance users' perception of chatbots as relatable agents (Atkinson et al., 2005; Chidambaram et al., 2012; Roubroeks et al., 2009; Louwerse et al., 2005). Empirical studies corroborate that chatbots with human-like features elicit stronger social responses, improving user satisfaction and adherence (Andrist et al., 2013; Cooney et al., 2015; Eyssel and Hegel, 2012). These insights underscored the importance of designing ECAs with tailored social cues to optimize mental health interventions. However, the specific impact of social cues on depression interventions remains unclear, particularly in self-guided settings where user engagement is critical.

This study aimed to compare a high-social-cue ECA (“Neil”) with a low-social-cue version, evaluating their impact on (1) depressive symptoms, (2) adherence, and (3) therapeutic alliance. In a 16-week randomized controlled trial involving 84 Chinese college students, we hypothesized that the high-social-cue intervention would outperform its low-social-cue counterpart across all metrics.

2. Methods

2.1. Participants and sample size

The inclusion criteria were as follows: (1) full-time college students; (2) aged ≥18 years; (3) PHQ-9 score ≥ 9; (4) smartphone ownership (iOS or Android); (5) willingness to use the chatbot intervention. The exclusion criteria were: (1) inability to read/write Chinese; (2) current psychiatric treatment (e.g., psychotherapy or pharmacotherapy); (3) baseline PHQ-9 score < 9 or a score of ≥2 on Item 9 (suicidal ideation).

Participants were recruited from four universities in Guangzhou, Hong Kong, and mainland China between December 1, 2022, and January 15, 2023 (original dates corrected for temporal consistency). Recruitment involved online advertisements and campus posters. Eligible students received cash compensation. Of 213 respondents, 84 met the criteria: South China University of Technology (n = 32), Guangdong University of Technology (n = 20), Jinan University (n = 20), and Hong Kong Polytechnic University (n = 12). The sample comprised undergraduates (n = 48) and postgraduates (n = 36) aged 18–28 years (M = 23.3, SD = 1.07), with slightly more males (51.2 %, n = 43) than females. All participants were native Chinese speakers.

A power analysis (α = 0.05, 1-β = 0.80) determined the required sample size based on prior studies showing a PHQ-9 reduction of 3.6 points (SD = 5) for chatbot interventions (Kroenke et al., 2001). The formula for a two-tailed independent t-test (Daniel and Cross, 2018) was:

$$n = \frac{2\left(Z_{\alpha/2} + Z_{\beta}\right)^2 \sigma^2}{\delta^2} \quad (1)$$

where: Zα/2 = 1.96 (α = 0.05), Zβ = 0.842 (β = 0.20), σ = 5 (pooled standard deviation), Δ = 3.6 (expected mean difference). This yielded n = 31 per group. Accounting for a 15 % attrition rate (Liu et al., 2022b), we recruited 37 participants per group (total N = 84). Given that psychological trials frequently employ lower thresholds (e.g., PHQ-9 ≥ 9) to encompass a wider spectrum of symptoms, utilizing the PHQ-9 to establish eligibility with a cutoff score of ≥9 aligns with the objectives of psychological intervention trials aimed at addressing a diverse array of depressive symptoms (von Glischinski et al., 2021).

Once the participant agrees to the informed consent form, they will be included in the trial. For safety reasons, participants will be provided with a free local mental health helpline number. The experimental methods used in this work were authorized by the local Institutional Review Board.

2.2. Interventions

2.2.1. Chatbot-delivered intervention

In this study, an ECA named “Neil” was developed and deployed via smartphone and web-based platforms. The ECA was accessible through a dedicated mobile application or web interface, compatible with Windows, macOS, Android, and iOS systems. Neil's conversational content was designed using principles of cognitive behavioral therapy (CBT) and validated by clinical psychologists. As shown in Fig. 1 (workflow) and Fig. 2 (system architecture), the ECA operated through a structured pipeline.

Neil was implemented using the open-source conversational AI framework RASA, supporting both text and voice interactions. For voice inputs, the open-source “Ali Paraformer” speech recognition model converted audio to text, which was then processed by the natural language understanding (NLU) module. This module performed three core functions: (1)Natural Language Processing: Extraction of keywords, semantic structures, and contextual cues. (2)Intent Classification: Machine learning models tagged user intents (e.g., “expressing sadness,” “seeking coping strategies”) against predefined categories. (3)Emotion Recognition: Identification of emotional states (e.g., anxiety, hopelessness) through lexical analysis (e.g., negative words like “worthless”), syntactic patterns, and speech prosody (for voice inputs).

Based on these analyses, the dialogue management module selected responses from a predefined CBT-aligned template library. Responses were dynamically adapted to user inputs (e.g., inserting “exam-related” advice when detecting academic stress) while strictly adhering to clinical guidelines. A rule-based natural language generation (NLG) module then produced output text.

During interactions, Neil assessed users' psychological states using validated scales and delivered CBT-based interventions. Key components included: (1) Psychoeducation about CBT principles, (2) Identification of automatic thoughts and irrational beliefs, (3) Development of

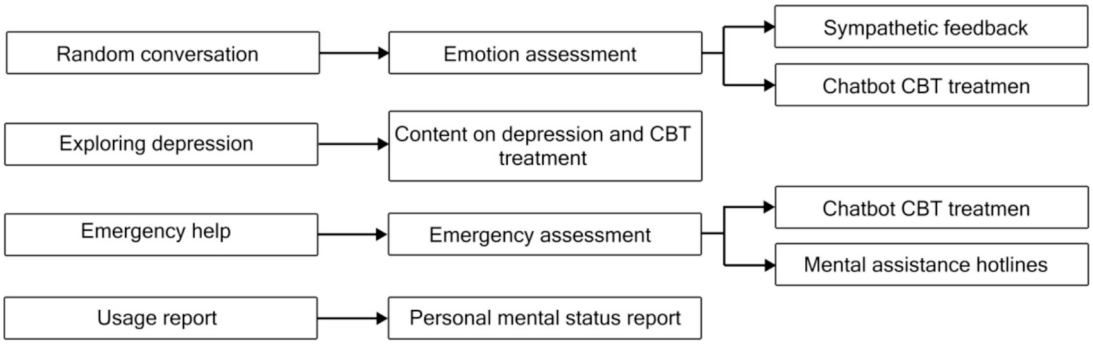

Fig. 1. The chatbot Neil's workflow.

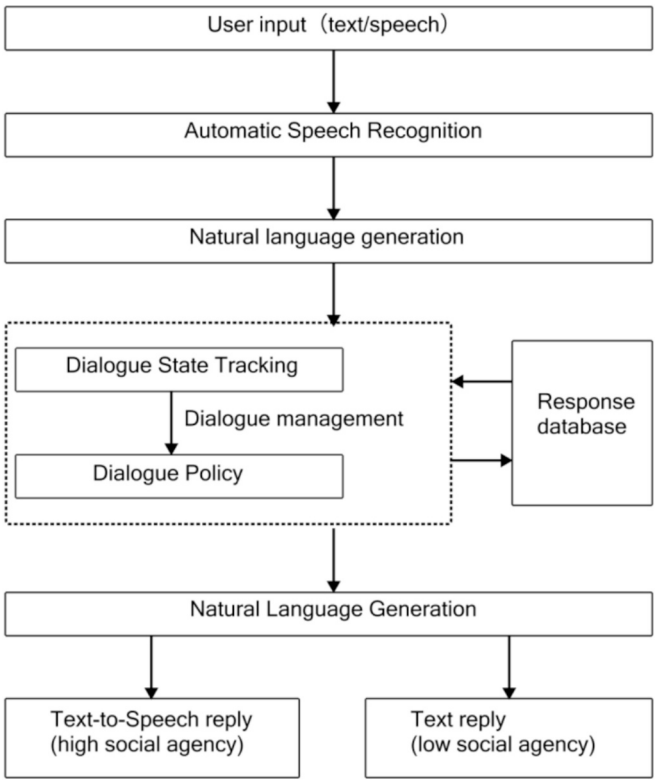

Fig. 2. The chatbot Neil's structure.

positive coping strategies, (4) Consolidation of therapeutic outcomes. Users could also engage in informal conversations with Neil, which provided empathetic feedback, or record psychological states to generate personalized reports.

To mitigate risks associated with uncontrolled AI outputs, Neil avoided large language models (LLMs). Instead, it used a rule-based template system to ensure adherence to CBT protocols and ethical standards (e.g., preventing harmful or non-scientific remarks). This design choice prioritized clinical safety over generative flexibility.

2.2.2. Social cues

As mentioned above, two chatbot variants based on the Neil platform were designed for this study. The two versions differed in the quantity of integrated social cues, defined as follows:(1) LSC condition: The chatbot used text-based empathetic communication (e.g., "I understand this must be difficult for you") displayed on mobile screens. As illustrated in Fig. 3A, interactions involved static text dialogues without vocal or animated elements.(2) HSC condition: While maintaining equivalent textual content, this version incorporated multimodal social cues through three enhancements: Emotional tone modulation (e.g., softened pitch during empathetic responses) implemented via Variational Inference for Text-to-Speech (VITS) technology; facial animations: Eye expressions (e.g., squinting to convey concern) synchronized with dialogue; body language: Head nods and gestures (e.g., waving) rendered through Unreal Engine 5 (UE5) -generated character animations (Fig. 3B).

Both conditions shared the identical conversational core, ensuring matched responses to participant inquiries. During interactions, the high-agency system output synchronized voice and animation, whereas the low-agency condition required manual text reading by users.

2.2.3. Participants administration

Either the HSC test group or the LSC control group received participants at random. A smartphone app or link to Neil, a therapeutic chatbot with HSC, was given to participants allocated to the HSC test group. They were instructed to use the chatbot for 16 weeks. Correspondingly, participants in the LSC control group were provided with a mobile app link to Neil, a therapeutic chatbot with LSC, and were asked to use the chatbot for 16 weeks.

The ECAs' intervention was the only kind of treatment that participants were allowed to seek during the experiment. However, for the safety of the participants, the research team also provided them with a free local emergency psychological assistance hotline. If participants report the need for emergency psychological assistance, professionals will provide them with help and continue to pay attention to the safety of the participants. In this case, the trial data of the relevant participants will not be used for statistical analysis.

2.3. Measures

First, baseline data were collected from participants. Baseline measurements included age, education level, depression severity (assessed using the PHQ-9), anxiety severity (GAD-7 (Spitzer et al., 2006)), and positive/negative affect (PANAS questionnaire (Watson et al., 1988)). These data were collected via online questionnaires distributed after participants confirmed enrollment.

The primary outcomes were reductions in depression (PHQ-9) and anxiety (GAD-7) scores. Secondary outcomes included user satisfaction (Client Satisfaction Questionnaire-8, CSQ-8 (Kelly et al., 2018)), therapeutic alliance (Working Alliance Inventory-Short Revised, WAI-SR (Munder et al., 2010)), and self-reported adherence (4-point Likert scale: 1 = no use, 2 = use ≤50 % of days, 3 = use >50 % of days, 4 = use almost daily).

Participants were assessed at five timepoints over 16 weeks (T_1 = baseline, T_2 = 4 weeks, T_3 = 8 weeks, T_4 = 12 weeks, T_5 = 16 weeks). The 4-week interval was chosen to minimize short-term symptom fluctuations. At T_1–T_5, PHQ-9, GAD-7, PANAS, and adherence data were collected; at T_5, CSQ-8 and WAI-SR were administered. All questionnaires were distributed electronically by the research team.

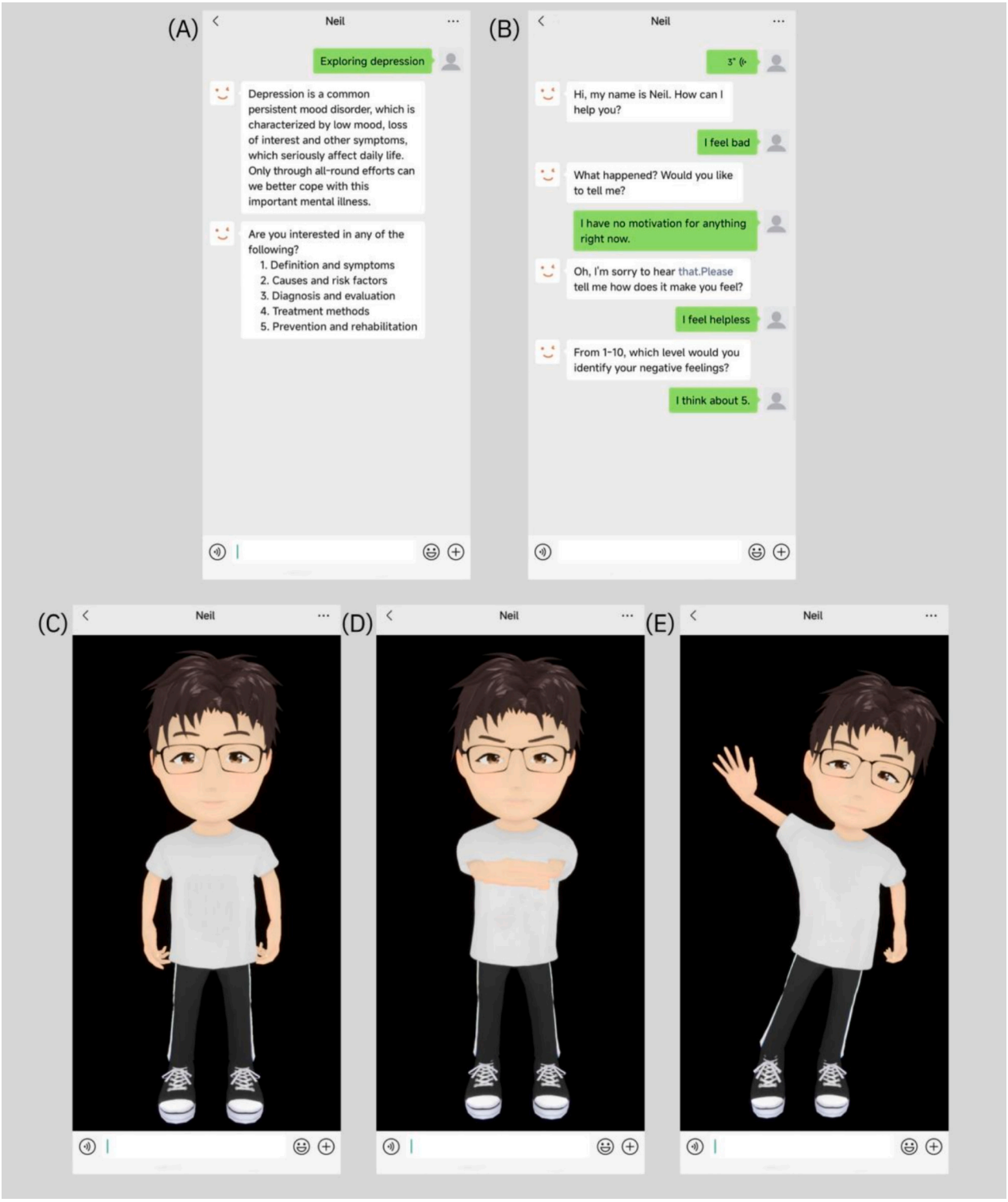

Fig. 3. Examples of LSC: (A) A question-and-answer system for exploring depression is provided while helping users understand depression; (B) A demonstration of CBT-based treatment for depression, supports text messages and voice messages. Examples of HSC:(C) A character in a resting state; (D) A serious animation with a frown and crossed arms; (E) An animation of a waving hand.

After the intervention, participants completed an open-ended question: "What was your best and worst experience interacting with Neil?"

2.4. Randomization

Because the HSC and LSC groups were presented differently and this was obvious to the participants, a blinded design was not possible. We made use of randomization. Using the random number generator in the SPSS v.27 (IBM Corp., Armonk, NY), a random number between $0 < n \leq 1$ was allocated to each participant. The HSC group was allocated to participants who were assigned a random number between $0 < n \leq 0.5$. The LSC group was allocated to the remaining participants who were randomly assigned a number between $0.5 < n \leq 1$.

2.5. Statistical methods

The significance level was set at $p = 0.05$. All analyses were performed using SPSS v.27. Baseline characteristics (e.g., age, sex, education, PHQ-9/GAD-7 scores) were compared between groups via analysis of variance (ANOVA) and chi-square tests to confirm randomization balance.

All outcomes were analyzed on an intention-to-treat (ITT) basis, where participants were retained in their originally assigned groups regardless of adherence or dropout. This approach preserved randomization benefits and minimized attrition bias (Fisher et al., 2017). For primary outcomes (PHQ-9/GAD-7 reductions), univariate analysis of covariance (ANCOVA) was conducted to assess group effects after adjusting for baseline scores. Missing data were assumed to be missing at random (MAR) and handled using multiple imputation with five iterations.

Cohen's d was calculated to quantify the magnitude of group differences. Secondary outcomes (CSQ-8, WAI-SR, adherence ratings) were compared using independent *t*-tests. Finally, participant feedback was analyzed via word frequency analysis.

3. Results

3.1. Participant flow

The experiment lasted 16 weeks, beginning on February 1, 2024. At the end of the trial, there were 16 participants who were not followed up, with a total loss rate of 19.05 % (16/84). Fig. 4 depicts the participants' general flow. Chi-square analysis and independent t-test were used to analyze the baseline data of participants who withdrew from the study and those who completed the study. Group membership ($\chi^2 = 0.31$; $P = 0.58$), gender ($\chi^2 = 0.44$; $P = 0.51$), age ($t = 0.38$; $P = 0.71$), education ($t = 1.06$; $P = 0.29$), PHQ-9 ($t = 0.99$; $P = 0.32$), GAD-7 ($t = 0.55$; $P = 0.59$), and PANAS positive ($t = 0.15$; $P = 0.88$) and negative ($t = 0.98$; $P = 0.33$) affect scores were among the factors that did not show any significant differences.

3.2. Baseline data

Age, gender, education level, and clinical characteristics were compared between the groups at baseline (T_1) using ANOVA and chi-square analysis, as indicated in Table 1, and no significant differences were detected in any of the aspects.

Table 1
Participant and variable demographics at baseline (T_1).

	HSC group[a]	LSC group[a]	χ^2/t	P
Age (years)	23.55(1.14)	23.14(0.97)	1.74	0.10
Gender				
Male	23(54.76)	20(47.62)	0.43	0.51
Female	19(45.24)	22(52.38)		
Education (years)	16.57(1.07)	16.19(0.93)	1.72	0.09
Scale, mean(SD)				
Depression (PHQ-9)	13.52(3.45)	13.21(3.17)	0.42	0.67
Anxiety (GAD-7)	15.10(3.25)	15.61(3.04)	0.75	0.45
Positive affect	28.17(7.92)	29.33(10.29)	0.58	0.57
Negative affect	28.93(8.15)	28.50(9.03)	0.18	0.86

[a] The numbers are mean (standard deviation) or n (%).

3.3. ITT analysis

Intention-to-treat (ITT) analysis was conducted for primary outcomes (PHQ-9 and GAD-7 scores) measured at T_5. Analysis of covariance (ANCOVA) revealed that the HSC group achieved significantly greater reductions in depression ($d = 0.63$; $F = 37.15$, $P < 0.01$) and anxiety ($d = 0.50$; $F = 9.55$, $P < 0.01$) compared to the LSC group (Table 2). Temporal trends of PHQ-9 and GAD-7 scores from T_1 to T_5 are illustrated in Fig. 5. After Bonferroni correction for multiple comparisons, reductions in PHQ-9 ($P < 0.01$) and GAD-7 ($P = 0.03$) remained significant. No significant between-group differences were observed in positive or negative affect scores ($P > 0.05$).

Within-group analyses demonstrated that the HSC group exhibited significant reductions in GAD-7 ($t = 2.581$, $P = 0.01$) and PHQ-9 ($t = 6.094$, $P = 0.001$) scores from T_1 to T_5. In contrast, the LSC group showed no significant improvement in GAD-7 scores ($P = 0.14$), although PHQ-9 scores decreased significantly ($t = 2.99$, $P = 0.004$).

3.4. Ancillary analyses

To find out if there was a main effect, patients who finished the trial underwent multivariate analysis of variance (MANOVA).The findings indicated that time ($F = 15.06$; $P < 0.01$) and group membership ($F = 23.35$; $P < 0.01$) significantly effected PHQ-9 scores; group membership ($F = 14.69$; $P < 0.01$) and time ($F = 2.52$; $P < 0.04$) significantly effected GAD-7 scores.

Based on the results of the independent *t*-test conducted with T_5 completers, it was found that the HSC group had a significantly better therapeutic alliance (as measured by WAI-SR) than the LSC control group ($t = 3.68$; $P < 0.001$), and the HSC group had a significantly

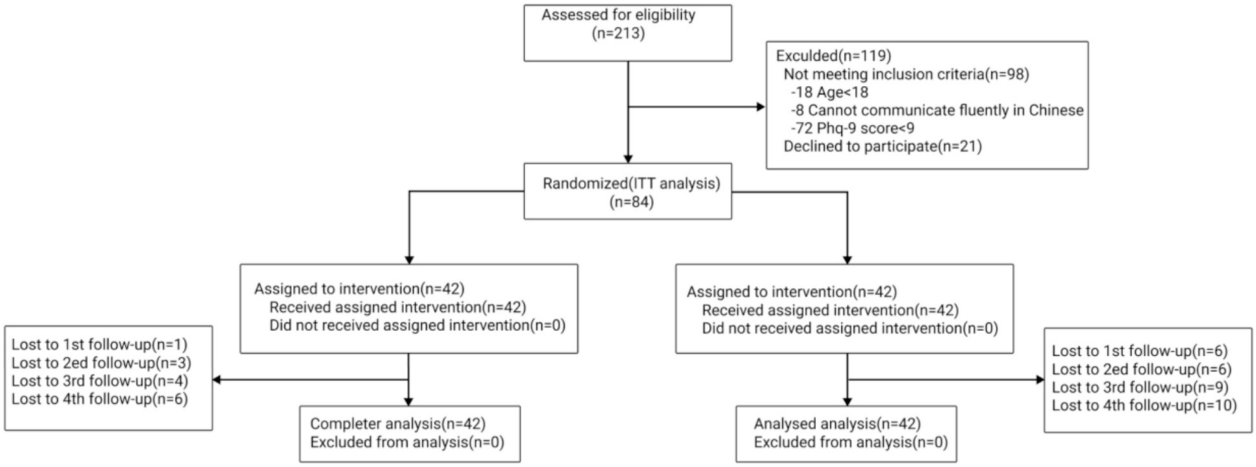

Fig. 4. The flow of participants.

Table 2
ITT analysis at T_5.

	HSC group		LSC group		F	P	d[c]
	T_5[a]	95%CI[b]	T_5[a]	95%CI[b]			
PHQ-9	9.05(0.25)	8.56–9.55	11.19(0.25)	10.70–11.69	37.15	<0.01*	0.63
GAD-7	14.60(0.22)	13.15–14.06	13.60(0.22)	14.14–15.05	9.55	0.003*	0.50
Positive affect	27.94(0.98)	26.00–29.89	29.88(0.98)	27.94–31.82	1.96	0.17	0.31
Negative affect	28.75(0.47)	27.81–29.69	27.66(0.47)	26.72–28.60	2.67	0.11	0.18

* The result is significant at the 0.01 level.
[a] The numbers are pooled mean (standard error).
[b] 95 % Confidence Interval.
[c] Using means and standard errors, Cohen's d was demonstrated for between-subjects effects at T_5.

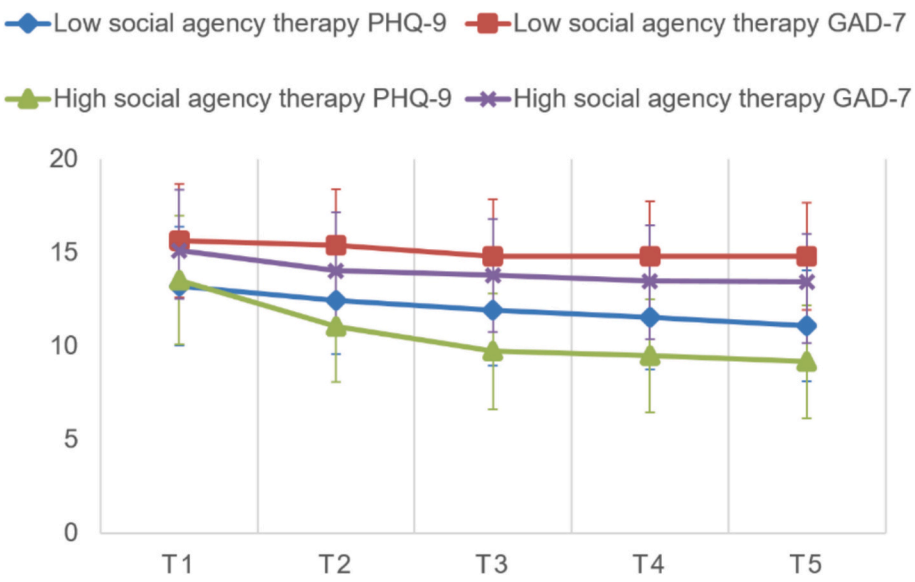

Fig. 5. Clinical variables during the trail.

higher CSQ-8 score ($t = 2.48$; $P = 0.02$).

Self-reported adherence rates for both groups were summarized in Table 3, reporting participant engagement frequencies from T_1 to T_5. An independent t-test revealed that the HSC group had significantly higher adherence rates ($t = 3.76$, $P < 0.01$, $d = 0.82$) than the LSC group. As shown in Fig. 6, adherence trends diverged between groups:(1)LSC group: Adherence increased during the first 8 weeks before declining in the subsequent 8 weeks. (2)HSC group: Adherence declined initially but stabilized after 8 weeks.

Word frequency analysis of open-ended responses ("What were your best and worst experiences with Neil?") was presented in Table 4. Key themes included:(1)HSC group: Participants praised the anthropomorphic design (e.g., emotional expressions, dynamic dialogues) for enhancing emotional involvement ($n = 108/38$ responses) and self-disclosure willingness ($n = 138/38$).(2)LSC group: Users emphasized functional efficiency ($n = 11/47$) but criticized unnatural interactions ($n = 18/47$).

Notably, three participants expressed greater trust in Neil compared to similar tools but remained more inclined to seek professional psychological assistance.

Table 3
The number of participants that engage at each frequency.

		Never used	≤50 % of days	>50 % of days	Daily use
LSC group	T_1	0	38	4	0
	T_2	0	36	6	0
	T_3	0	24	18	0
	T_4	8	34	0	0
	T_5	13	29	0	0
HSC group	T_1	0	11	31	0
	T_2	0	30	12	0
	T_3	1	40	1	0
	T_4	0	38	4	0
	T_5	1	41	0	0

4. Discussion

This randomized controlled trial rigorously examined the efficacy of therapeutic chatbots featuring high social cues—such as voice, animations, and nonverbal gestures—compared to text-only chatbots for self-guided depression management among college students. The findings indicated that participants engaging with the high social cue chatbot experienced significantly greater reductions in depression (as measured by the PHQ-9) and anxiety (as assessed by the GAD-7) scores, in contrast to those in the low social cue control group ($p < 0.01$). Additionally, individuals utilizing the high social cue chatbot reported higher adherence rates, a stronger therapeutic alliance (as indicated by the WAI-SR), and greater overall satisfaction (CSQ-8). These results substantiate the social cue hypothesis, suggesting that anthropomorphic design elements enhance user engagement and foster trust (Louwerse et al., 2005). Interestingly, we also found that although both the HSC and LSC groups had significant effects in reducing PHQ-9 scores, the GAD-7 scores of the HSC group decreased significantly from T1 to T5, while the GAD-7 scores of the LSC group did not improve significantly. For anxiety, social cues may have an immediate soothing effect. Voice intonation (such as a gentle voice) and animation (such as a smiling face) can directly reduce anxiety-related physiological arousal (such as

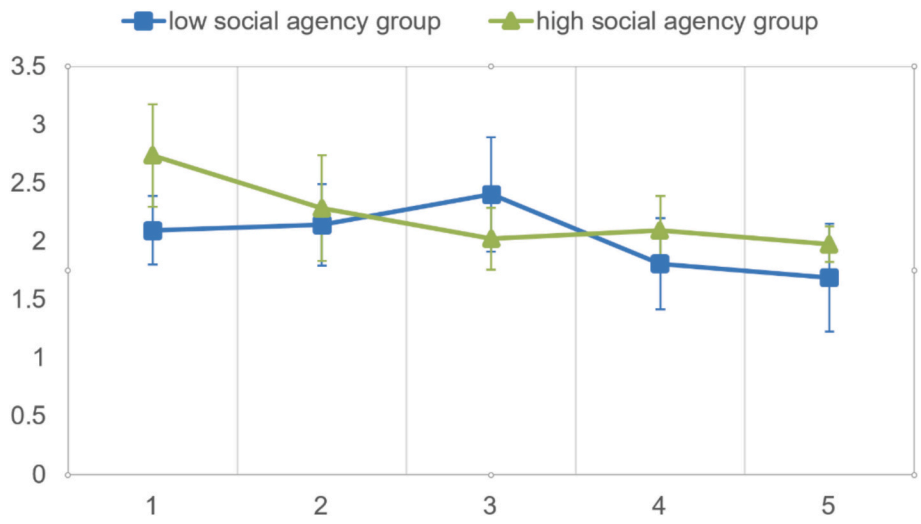

Fig. 6. Self-reported adherence rate.

Table 4
Participant feedback results statistics.[a]

Question	User feedback	LSC group	HSC group
The best experiences about using Neil	Easy to access	(17/38)[b]	(15/40)[b]
	Quick response	(22/38)	(17/40)
	More self-disclosure	(9/38)	(12/40)
	Empathetic feedback	(10/38)	(17/40)
	Humanized wording	(12/38)	(11/40)
	Professional	(12/38)	(8/40)
	Turn-based dialogue	(14/38)	(6/40)
	Anonymous safety	(10/38)	(2/40)
The worst experiences about using Neil	Lack of personalization	(20/38)	(14/40)
	Unnatural	(17/38)	(13/40)
	Stipid	(19/38)	(6/40)
	General	(16/38)	(11/40)
	Irrelevant content	(8/38)	(9/40)
	Repetitive dialogue	(9/38)	(5/40)
	Expertise is difficult to understand	(8/38)	(5/40)

[a] Several responses include more than one theme and were counted more than once.
[b] Numbers are (counted number/total number of participants).

tension), while plain text lacks such immediate emotion regulation functions.

Adherence trends in the study highlighted the challenges associated with fully automated interventions, as both groups demonstrated declining engagement over the 16-week period. However, the high social cue group maintained a significantly higher level of adherence. Existing digital treatments and telemedicine interventions often suffer from low compliance, unsustainability, and a lack of flexibility (Duarte et al., 2017; Gilbody et al., 2017). In contrast, AI chatbots improve adherence through several innovative mechanisms: (1) On-demand support flexibility, providing 24/7 instant interaction capabilities that transcend the time and space limitations of traditional digital interventions; (2) Personalized service continuity, utilizing machine learning to analyze user data (including electronic medical records, real-time physiological indicators, and interaction histories) to dynamically generate customized intervention plans (Galvão Gomes da Silva et al., 2018; Stephens et al., 2019); (3) Enhanced interaction stickiness, employing natural language processing technology (Milne-Ives et al., 2020) to simulate human dialogue and create a safe communication environment that mitigates stigma—particularly crucial for the ongoing management of sensitive health issues; and (4) Technology integration and scalability, where user engagement is bolstered through immersive experiences such as virtual reality (Gaffney et al., 2019), which prove more effective than the one-way educational models of traditional digital interventions. These characteristics enable AI chatbots to address the prevalent issue of "user churn" in conventional digital healthcare, facilitating a more sustained user engagement pathway compared to asynchronous communication methods like email or video, thanks to instant feedback mechanisms such as real-time behavior monitoring, goal setting, and knowledge delivery.

The significance of this study lies in its validation of the social cue hypothesis. Through randomized controlled trials, it demonstrates that HSC chatbots (incorporating elements like voice and animation) can more effectively alleviate symptoms of depression and anxiety compared to LSC chatbots (text-only), thereby supporting the applicability of social cue theory within digital therapy. This research expands the theory of human-computer interaction by revealing the critical role of non-verbal cues (such as nodding and facial expressions) in establishing a therapeutic alliance, thus providing empirical evidence for the design of Embodied Conversational Agents (ECAs).

Nonetheless, the study acknowledges certain limitations, particularly a small and homogeneous sample comprised of tech-savvy college students, which may restrict the generalizability of the findings. Although the high social cue design led to improved outcomes, technical limitations prevented the full replication of human gestures, such as subtle facial expressions, potentially limiting users' ability to accurately interpret nonverbal cues. Future research should aim to validate these findings across more diverse populations, isolate the effects of specific social cues (e.g., voice versus animations), and explore multimodal integrations with wearable sensors or virtual reality to enhance user immersion.

It is also crucial to recognize the inherent risks associated with self-administered AI systems. For instance, free input from users may contain ambiguous or unstructured information (such as emotional descriptions), which can compromise the accuracy of medical data records. To address this, the practice of doctor-patient dialogue can be employed, utilizing declarative questions (e.g., "You mentioned that the headache lasted for three days, correct?") to verify the system's understanding of the user's intent. Additionally, there exists a risk that patients may delay treatment for genuine emergencies due to their reliance on digital health applications. To mitigate this concern, an emergency service feature should be integrated into the application, providing a clear one-click option for users to contact emergency services, accompanied by a prompt such as, "For emergency assistance, please click here to reach the rescue center."

5. Conclusion

This study compared the effectiveness of a high-social-cue therapeutic chatbot (featuring voice, facial animations, and gestures) against a low-social-cue conventional version (text-only) in delivering self-help depression interventions to college students. The results demonstrated that the high-social-cue chatbot produced significantly greater reductions in depression severity, higher user adherence, and stronger therapeutic alliance than its low-social-cue counterpart. These findings extended the theoretical basis of the social cue hypothesis, contributed to human-computer interaction frameworks, and provided actionable insights for designing therapeutic chatbots and other AI health agents.

CRediT authorship contribution statement

Shuo Xu: Writing – original draft, Project administration, Funding acquisition, Conceptualization, Writing – review & editing, Validation, Methodology, Formal analysis. **Tiancong Ma:** Validation, Methodology, Conceptualization, Visualization, Software, Data curation.

Ethics approval

The study design and protocol were approved by the Institutional Ethics Committee of South China University of Technology. The research was carried out in conjunction with the Declaration of Helsinki. Written informed consent was obtained from all participants.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

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.

Acknowledgments

The authors thank all participants who participated in this study.

References

Abd-Alrazaq, A.A., Rababeh, A., Alajlani, M., et al., 2020. Effectiveness and safety of using chatbots to improve mental health: systematic review and meta-analysis [J]. J. Med. Internet Res. 22 (7), e16021.
Agrawal, S., Williams, M.-A., 2017. Robot authority and human obedience: A study of human behaviour using a robot security guard, F.
Andersson, G., 2016. Internet-delivered psychological treatments [J]. Annu. Rev. Clin. Psychol. 12 (1), 157–179.
Andrews, B., Wilding, J.M., 2004. The relation of depression and anxiety to life-stress and achievement in students [J]. Br. J. Psychol. 95 (4), 509–521.
Andrist, S., Spannan, E., Mutlu, B., 2013. Rhetorical robots: making robots more effective speakers using linguistic cues of expertise, F. IEEE.
Atkinson, R.K., Mayer, R.E., Merrill, M.M., 2005. Fostering social agency in multimedia learning: examining the impact of an animated agent's voice [J]. Contemp. Educ. Psychol. 30 (1), 117–139.
Bailenson, J.N., Blascovich, J., Beall, A.C., et al., 2001. Equilibrium theory revisited: mutual gaze and personal space in virtual environments [J]. Presence: Teleoperators & Virtual Environments 10 (6), 583–598.
Bickmore, T., Gruber, A., Picard, R., 2005. Establishing the computer–patient working alliance in automated health behavior change interventions [J]. Patient Educ. Couns. 59 (1), 21–30.
Bolinski, F., Kleiboer, A., Karyotaki, E., et al., 2018. Effectiveness of a transdiagnostic individually tailored Internet-based and mobile-supported intervention for the indicated prevention of depression and anxiety (ICare Prevent) in Dutch college students: study protocol for a randomised controlled trial [J]. Trials 19, 1–13.
Burton, C., Szentagotai Tatar, A., McKinstry, B., et al., 2016. Pilot randomised controlled trial of Help4Mood, an embodied virtual agent-based system to support treatment of depression [J]. J. Telemed. Telecare 22 (6), 348–355.
Chidambaram, V., Chiang, Y.-H., Mutlu, B., 2012. Designing persuasive robots: how robots might persuade people using vocal and nonverbal cues, F.
Cook, L., Mostazir, M., Watkins, E., 2019. Reducing stress and preventing depression (RESPOND): randomized controlled trial of web-based rumination-focused cognitive behavioral therapy for high-ruminating university students [J]. J. Med. Internet Res. 21 (5), e11349.
Cooney, S., Dignam, H., Brady, N., 2015. Heads first: visual aftereffects reveal hierarchical integration of cues to social attention [J]. PloS One 10 (9), e0135742.
Daniel, W.W., Cross, C.L., 2018. Biostatistics: A Foundation for Analysis in the Health Sciences [M]. John Wiley & Sons.
Duarte, A., Walker, S., Littlewood, E., et al., 2017. Cost-effectiveness of computerized cognitive–behavioural therapy for the treatment of depression in primary care: findings from the Randomised Evaluation of the Effectiveness and Acceptability of Computerised Therapy (REEACT) trial [J]. Psychol. Med. 47 (10), 1825–1835.
Eyssel, F., Hegel, F., 2012. (s) he's got the look: Gender stereotyping of robots 1 [J]. J. Appl. Soc. Psychol. 42 (9), 2213–2230.
Fisher, L.D., Dixon, D.O., Herson, J., et al., 2017. Intention to treat in clinical trials [M]. In: Statistical issues in drug research and development. Routledge, pp. 331–350.
Fitzpatrick, K.K., Darcy, A., Vierhile, M., 2017. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): a randomized controlled trial [J]. JMIR Mental Health 4 (2), e7785.
Fitzsimmons-Craft, E.E., Taylor, C.B., Newman, M.G., et al., 2021. Harnessing mobile technology to reduce mental health disorders in college populations: a randomized controlled trial study protocol [J]. Contemp. Clin. Trials 103, 106320.
Fulmer, R., Joerin, A., Gentile, B., et al., 2018. Using psychological artificial intelligence (Tess) to relieve symptoms of depression and anxiety: randomized controlled trial [J]. JMIR Mental Health 5 (4), e9782.
Gaffney, H., Mansell, W., Tai, S., 2019. Conversational agents in the treatment of mental health problems: mixed-method systematic review [J]. JMIR mental health 6 (10), e14166.
Galvão Gomes da Silva, J., Kavanagh, D.J., Belpaeme, T., et al., 2018. Experiences of a motivational interview delivered by a robot: qualitative study [J]. J. Med. Internet Res. 20 (5), e116.
Gilbody, S., Brabyn, S., Lovell, K., et al., 2017. Telephone-supported computerised cognitive–behavioural therapy: REEACT-2 large-scale pragmatic randomised controlled trial [J]. Br. J. Psychiatry 210 (5), 362–367.
von Glischinski, M., von Brachel, R., Thiele, C., et al., 2021. Not sad enough for a depression trial? A systematic review of depression measures and cut points in clinical trial registrations [J]. J. Affect. Disord. 292, 36–44.
Gratzer, D., Goldbloom, D., 2020. Therapy and e-therapy—preparing future psychiatrists in the era of apps and chatbots [J]. Acad. Psychiatry 44, 231–234.
Greer, S., Ramo, D., Chang, Y.-J., et al., 2019. Use of the chatbot "vivibot" to deliver positive psychology skills and promote well-being among young people after cancer treatment: randomized controlled feasibility trial [J]. JMIR Mhealth Uhealth 7 (10), e15018.
Gulliver, A., Griffiths, K.M., Christensen, H., 2010. Perceived barriers and facilitators to mental health help-seeking in young people: a systematic review [J]. BMC Psychiatry 10, 1–9.
Harrer, M., Adam, S.H., Fleischmann, R.J., et al., 2018. Effectiveness of an internet-and app-based intervention for college students with elevated stress: randomized controlled trial [J]. J. Med. Internet Res. 20 (4), e136.
Hawton, K., Comabella I, C.C., Haw, C., et al., 2013. Risk factors for suicide in individuals with depression: a systematic review [J]. J. Affect. Disord. 147 (1–3), 17–28.
Herrero, R., Mira, A., Cormo, G., et al., 2019. An internet based intervention for improving resilience and coping strategies in university students: study protocol for a randomized controlled trial [J]. Internet Interv. 16, 43–51.
Huang, J., Li, Q., Xue, Y., et al., 2015. Teenchat: a chatterbot system for sensing and releasing adolescents' stress, F. Springer.
Hysenbegasi, A., Hass, S.L., Rowland, C.R., 2005. The impact of depression on the academic productivity of university students [J]. J. Ment. Health Policy Econ. 8 (3), 145.
Inkster, B., Sarda, S., Subramanian, V., 2018. An empathy-driven, conversational artificial intelligence agent (Wysa) for digital mental well-being: real-world data evaluation mixed-methods study [J]. JMIR Mhealth Uhealth 6 (11), e12106.
Jorm, A.F., Wright, A., Morgan, A.J., 2007. Where to seek help for a mental disorder? [J]. Med. J. Aust. 187 (10), 556–560.
Kelly, P.J., Kyngdon, F., Ingram, I., et al., 2018. The client satisfaction Questionnaire-8: psychometric properties in a cross-sectional survey of people attending residential substance abuse treatment [J]. Drug Alcohol Rev. 37 (1), 79–86.
Kroenke, K., Spitzer, R.L., Williams, J.B.W., 2001. The PHQ-9: validity of a brief depression severity measure [J]. J. Gen. Intern. Med. 16 (9), 606–613.
Lee, R.A., Jung, M.E., 2018. Evaluation of an mhealth app (destressify) on university students' mental health: pilot trial [J]. JMIR Mental Health 5 (1), e8324.
Lee, Y.-C., Yamashita, N., Huang, Y., 2020. Designing a chatbot as a mediator for promoting deep self-disclosure to a real mental health professional [J]. Proceedings of the ACM on Human-Computer Interaction 4 (CSCW1), 1–27.
Lee, T.S.-H., Wu, Y.-J., Chao, E., et al., 2021. Resilience as a mediator of interpersonal relationships and depressive symptoms amongst 10th to 12th grade students [J]. J. Affect. Disord. 278, 107–113.
Lei, X.-Y., Xiao, L.-M., Liu, Y.-N., et al., 2016. Prevalence of depression among Chinese university students: a meta-analysis [J]. PloS One 11 (4), e0153454.
Li, C.-Y., 2013. Persuasive messages on information system acceptance: a theoretical extension of elaboration likelihood model and social influence theory [J]. Comput. Hum. Behav. 29 (1), 264–275.

Liu, X.-Q., Guo, Y.-X., Zhang, W.-J., et al., 2022a. Influencing factors, prediction and prevention of depression in college students: a literature review [J]. World J. Psychiatry 12 (7), 860.
Liu, H., Peng, H., Song, X., et al., 2022b. Using AI chatbots to provide self-help depression interventions for university students: a randomized trial of effectiveness [J]. Internet Interv. 27, 100495.
Lopez, A., Ccasane, B., Paredes, R., et al., 2017. Effects of using indirect language by a robot to change human attitudes, F.
Louwerse, M.M., Graesser, A.C., Lu, S., et al., 2005. Social cues in animated conversational agents [J]. Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition 19 (6), 693–704.
Martin, C.D., 1997. The media equation: how people treat computers, television and new media like real people and places [book review] [J]. IEEE Spectr. 34 (3), 9–10.
Milne-Ives, M., de Cock, C., Lim, E., et al., 2020. The effectiveness of artificial intelligence conversational agents in health care: systematic review [J]. J. Med. Internet Res. 22 (10), e20346.
Munder, T., Wilmers, F., Leonhart, R., et al., 2010. Working Alliance Inventory-Short Revised (WAI-SR): psychometric properties in outpatients and inpatients [J]. Clin. Psychol. Psychother. Int. J. Theory Pract. 17 (3), 231–239.
Musiat, P., Potterton, R., Gordon, G., et al., 2019. Web-based indicated prevention of common mental disorders in university students in four European countries–study protocol for a randomised controlled trial [J]. Internet Interv. 16, 35–42.
Neathery, M., Taylor, E.J., He, Z., 2020. Perceived barriers to providing spiritual care among psychiatric mental health nurses [J]. Arch. Psychiatr. Nurs. 34 (6), 572–579.
Palma-Gómez, A., Herrero, R., Baños, R., et al., 2020. Efficacy of a self-applied online program to promote resilience and coping skills in university students in four Spanish-speaking countries: study protocol for a randomized controlled trial [J]. BMC Psychiatry 20, 1–15.
Ratnayake, P., Hyde, C., 2019. Mental health literacy, help-seeking behaviour and wellbeing in young people: implications for practice [J]. Educ. Dev. Psychol. 36 (1), 16–21.
Reavley, N.J., McCann, T.V., Cvetkovski, S., et al., 2014. A multifaceted intervention to improve mental health literacy in students of a multicampus university: a cluster randomised trial [J]. Soc. Psychiatry Psychiatr. Epidemiol. 49, 1655–1666.
Renton, T., Tang, H., Ennis, N., et al., 2014. Web-based intervention programs for depression: a scoping review and evaluation [J]. J. Med. Internet Res. 16 (9), e3147.
Roubroeks, M., Midden, C., Ham, J., 2009. Does it make a difference who tells you what to do? Exploring the effect of social agency on psychological reactance, F.
Ryu, H., Kim, S., Kim, D., et al., 2020. Simple and steady interactions win the healthy mentality: designing a chatbot service for the elderly [J]. Proceedings of the ACM on human-computer interaction 4 (CSCW2), 1–25.
Spitzer, R.L., Kroenke, K., Williams, J.B.W., et al., 2006. A brief measure for assessing generalized anxiety disorder: the GAD-7 [J]. Arch. Intern. Med. 166 (10), 1092–1097.
Stephens, T.N., Joerin, A., Rauws, M., et al., 2019. Feasibility of pediatric obesity and prediabetes treatment support through Tess, the AI behavioral coaching chatbot [J]. Transl. Behav. Med. 9 (3), 440–447.
Suganuma, S., Sakamoto, D., Shimoyama, H., 2018. An embodied conversational agent for unguided internet-based cognitive behavior therapy in preventative mental health: feasibility and acceptability pilot trial [J]. JMIR Mental Health 5 (3), e10454.
Torous, J., Myrick, K.J., Rauseo-Ricupero, N., et al., 2020. Digital mental health and COVID-19: using technology today to accelerate the curve on access and quality tomorrow [J]. JMIR Mental Health 7 (3), e18848.
Vaidyam, A.N., Wisniewski, H., Halamka, J.D., et al., 2019. Chatbots and conversational agents in mental health: a review of the psychiatric landscape [J]. Can. J. Psychiatry 64 (7), 456–464.
Wang, K., Varma, D.S., Prosperi, M., 2018. A systematic review of the effectiveness of mobile apps for monitoring and management of mental health symptoms or disorders [J]. J. Psychiatr. Res. 107, 73–78.
Watson, D., Clark, L.A., Tellegen, A., 1988. Development and validation of brief measures of positive and negative affect: the PANAS scales [J]. J. Pers. Soc. Psychol. 54 (6), 1063.
Zhong, X., Liu, Y., Pu, J., et al., 2019. Depressive symptoms and quality of life among Chinese medical postgraduates: a national cross-sectional study [J]. Psychol. Health Med. 24 (8), 1015–1027.