Effect of an AI agent trained on a large language model (LLM) as an intervention for depression and anxiety symptoms in young adults: a 28-day randomized controlled trial

Yuqing Zhao, Wei Qian, Yaru Chen, Donghong Wu, Yujia Luo, Kankan Wu, Zhengkui Liu

Submitted to: Journal of Medical Internet Research
on: February 26, 2024

Table of Contents

Effect of an AI agent trained on a large language model (LLM) as an intervention for depression and anxiety symptoms in young adults: a 28-day randomized controlled trial

Yuqing Zhao[1, 2*]; Wei Qian[1, 3*]; Yaru Chen[1, 2]; Donghong Wu[4]; Yujia Luo[4]; Kankan Wu[1, 2]; Zhengkui Liu[1, 2]

[1]CAS Key Laboratory of Mental Health, Institute of Psychology Chinese Academy of Sciences Beijing CN
[2]Department of Psychology University of Chinese Academy of Sciences Beijing CN
[3]Department of Psychology Sofia University San Francisco US
[4]Beijing Zhongke Psychological Assistance Center Beijing CN
[*]these authors contributed equally

Corresponding Author:
Zhengkui Liu
CAS Key Laboratory of Mental Health, Institute of Psychology
Chinese Academy of Sciences
No.16, Lincui Road
Beijing
CN

Abstract

Background: Internet- and mobile-based psychological interventions (IMIs) have the advantages of being more accessible, less costly in terms of time and money, and more user friendly than traditional psychological interventions. Young adults face emotional problems in their daily lives that cannot be ignored, and youth are becoming the mainstay of the mobile internet. Large language modeling techniques are rapidly evolving and show great potential for understanding and generating natural language, enabling more effective interaction with human dialog than traditional techniques, and large language model (LLM)-based AI conversational agents may play an important role in intervening in young adults' negative emotions.

Objective: The aim of this study is to investigate whether an AI agent trained on a LLM can intervene in depression and anxiety in young people with negative emotions and to determine the effectiveness of such intervention.

Methods: This study is a 28-day randomized controlled trial (RCT). Residents were randomly assigned to an intervention group or a waiting group according to the order in which they were successfully contacted by the staff, and each user was asked to engage in a total of 28 days of dialog intervention with the Douyin companion bot and complete three psychological questionnaires (on Days 1, 14, and 28); however, the intervention group began to receive the dialog intervention after completing the first questionnaire, and the waiting group began to receive the dialog intervention after completing the third questionnaire. During the first four weeks, the waiting group was treated as a blank control. The two groups of subjects completed the three questionnaires at exactly the same point in time. Each user's depression, anxiety, and positive and negative emotions were measured using the Patient Health Questionnaire (PHQ-9), the Generalized Anxiety Disorder Scale (GAD-7) and the Positive and Negative Affect Schedule (PANAS), respectively.

Results: Data from 657 users who completed pre-, mid- and posttreatment measures were included in the analyses. The prevalence of mild depression and anxiety among participants at baseline was 44.3% and 23.4%, respectively. The results showed a significant reduction in negative moods in the intervention group after four weeks of the dialog intervention, which was not found when the dialog intervention was conducted for a fortnight. Repeated-measures ANOVA showed that the dialog intervention significantly reduced depression in the intervention group at two weeks and significantly reduced both depression and anxiety in the intervention group at four weeks.

Conclusions: Overall, LLM-based AI conversational agents can effectively alleviate the mild anxiety and depressive symptoms of young adults with negative emotions through dialog interventions.

(JMIR Preprints 26/02/2024:57766)
DOI: https://doi.org/10.2196/preprints.57766

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Effect of an AI agent trained on a large language model (LLM) as an intervention for depression and anxiety symptoms in young adults: a 28-day randomized controlled trial

Wei Qian[1,3,#], Yuqing Zhao[1,2,#], Yaru Chen[1,2], Donghong Wu[4], Yujia Luo[4], Kankan Wu[2,1,*], Zhengkui Liu[1,2,*]

[1] CAS Key Laboratory of Mental Health, Institute of Psychology, Chinese Academy of Sciences, Beijing, China

[2] Department of Psychology, University of Chinese Academy of Sciences, Beijing, China

[3] Department of Psychology, Sofia University, San Francisco, California, USA

[4] Beijing Zhongke Psychological Assistance Center, Beijing, China

Author Note

# Wei Qian and Yuqing Zhao contributed equally to the paper.

* Zhengkui Liu and Kankan Wu are co-corresponding authors. Correspondence concerning this article should be addressed to:

Zhengkui Liu (PhD), Key Laboratory of Mental Health, Institute of Psychology, Chinese Academy of Sciences, 16 Lincui Road, Chaoyang District, Beijing 100101, China. E-mail address: liuzk@psych.ac.cn.

Kankan Wu (MD), Key Laboratory of Mental Health, Institute of Psychology, Chinese Academy of Sciences, 16 Lincui Road, Chaoyang District, Beijing 100101, China. E-mail address: wukk@psych.ac.cn

Word count 5004 words

Number of figures: 1

Number of tables: 3

Paper Type: Original research

Effect of an AI agent trained on a large language model (LLM) as an intervention for depression and anxiety symptoms in young adults: a 28-day randomized controlled trial

Abstract

Background: Internet- and mobile-based psychological interventions (IMIs) have the advantages of being more accessible, less costly in terms of time and money, and more user friendly than traditional psychological interventions. Young adults face emotional problems in their daily lives that cannot be ignored, and youth are becoming the mainstay of the mobile internet. Large language modeling techniques are rapidly evolving and show great potential for understanding and generating natural language, enabling more effective interaction with human dialog than traditional techniques, and large language model (LLM)-based AI conversational agents may play an important role in intervening in young adults' negative emotions.
Objective: The aim of this study is to investigate whether an AI agent trained on a LLM can intervene in depression and anxiety in young people with negative emotions and to determine the effectiveness of such intervention.
Methods: This study is a 28-day randomized controlled trial (RCT). Residents were randomly assigned to an intervention group or a waiting group according to the order in which they were successfully contacted by the staff, and each user was asked to engage in a total of 28 days of dialog intervention with the Douyin companion bot and complete three psychological questionnaires (on Days 1, 14, and 28); however, the intervention group began to receive the dialog intervention after completing the first questionnaire, and the waiting group began to receive the dialog intervention after completing the third questionnaire. During the first four weeks, the waiting group was treated as a blank control. The two groups of subjects completed the three questionnaires at exactly the same point in time. Each user's depression, anxiety, and positive and negative emotions were measured using the Patient Health Questionnaire (PHQ-9), the Generalized Anxiety Disorder Scale (GAD-7) and the Positive and Negative Affect Schedule (PANAS), respectively.
Results: Data from 657 users who completed pre-, mid- and posttreatment measures were included in the analyses. The prevalence of mild depression and anxiety among participants at baseline was 44.3% and 23.4%, respectively. The results showed a significant reduction in negative moods in the intervention group after four weeks of the dialog intervention, which was not found when the dialog intervention was conducted for a fortnight. Repeated-measures ANOVA showed that the dialog intervention significantly reduced depression in the intervention group at two weeks and significantly reduced both depression and anxiety in the intervention group at four weeks.
Conclusions: Overall, LLM-based AI conversational agents can effectively alleviate the mild anxiety and depressive symptoms of young adults with negative emotions through dialog interventions.

Key words: LLM-based; AI conversational agents; mental health; depression; anxiety; youth

Introduction

Internet- and mobile-based psychological interventions (IMIs) have the advantage of being more accessible and less costly in terms of time and money than traditional psychological interventions [1] [2], with the added advantage that people in need do not avoid seeking help because of stigma [3]. The efficacy of Internet interventions has been demonstrated in several studies [4]. Such anonymized online interventions can be effective at reducing the risk of individuals avoiding medical care due to fear of social pressure (i.e., experiencing stigmatization). However, even though there has been substantial evidence that Internet-based interventions can be effective at changing behavior and providing help in treating mental disorders, some traditional interventions for depression and anxiety symptoms still have low adoption rates and high attrition rates [5][6].

Conversational agents are software programs (including chatbots and robots) that use artificial intelligence to simulate a conversation with a user via text or speech, and they have demonstrated many application benefits in mental health [7]. Compared to other digital mental health interventions, conversational agents are particularly promising due to their own strong interactivity and ability to simulate therapeutic conversations and control their intensity [8]. Increasing access to

information via the internet and cell phones now also highlights the potential for conversational agents to provide autonomous, interactive and critical mental health support [9]. It has been shown that it is not the involvement of a therapist or professional resource that truly makes a difference in psychological interventions but rather non-directive interpersonal encounters [9] and that conversational agents can provide just the kind of companionship that is completely free of value judgments.

Young adults face emotional problems, such as depression and anxiety, in their daily lives that can have a significant impact on their lives and cannot be ignored [10]. According to the American Psychological Association, anxiety and depression are common emotional reactions that lead to a very similar set of symptoms, including difficulty sleeping, fatigue, muscle tension, and irritability [11]. Among young adults, the prevalence of depression ranges from 13.5% to 48.3%, and the prevalence of anxiety ranges from 23.6% to 50% [11][12][13].

Recent surveys show that youth are becoming the mainstay of the mobile internet and short video industry [14], and conversational agents can often play an important role in relation to well-known platforms or products (e.g., Apple's Siri). After seeing the great promise of conversational agents for psychological interventions and the emotional problems faced by youth groups, [Hidden because of the anonymity requirements of peer review] Corporate Social Responsibility ([Hidden] CSR) developed an AI conversational agent called Douyin Xinqing (Douyin companion bot). Users can chat with it, and it replies in the form of text, etc. The Douyin companion bot guides the user through a chat about their emotions and shows companionship and empathy when appropriate. The underlying theory involves schools of thought such as humanistic psychology, positive psychology, solution-focused brief therapy (SFBT) and problem management plus (PM+) and can provide users with warm companionship that is completely nonjudgmental and totally accepting. To test the actual intervention effect produced by users' conversations with the companion bot on the alleviation of their emotional problems, [Hidden because of the anonymity requirements of peer review] CSR joined forces with the [Hidden because of the anonymity requirements of peer review] to launch a 28-day randomized controlled trial (RCT). We hypothesized that after a 28-day dialog intervention, compared with the control group, the intervention group would experience a significant reduction in depression and anxiety, an increase in subjectively perceived positive emotions about life, and a decrease in negative emotions.

Method

Recruitment and Procedure

The study design and procedures were approved by the ethics review committee of the [Hidden because of the anonymity requirements of peer review]. We recruited participants by dropping in-app messages to users in Douyin, targeting young adults with symptoms of depression or anxiety. Participants first read the instructions and were informed that this was a long-term intervention and that their main task was to talk to an AI robot and complete several questionnaires. Willing participants were asked to complete a simple recruitment questionnaire that included basic demographic information and the Patient Health Questionnaire (PHQ-9) and the Generalized Anxiety Disorder Scale (GAD-7), which measure symptoms of depression and anxiety. The inclusion criteria included being 18-25 years old, proficient in Chinese, not dyslexic; scoring at least 5 on the PHQ-9 or GAD-7; and not having a serious physical or diagnosed mental illness. At the end of the questionnaire, participants were asked to manually check the box for informed consent to ensure that they were aware of all aspects of the study and were willing to have conversations with the AI robot used for research purposes.

Participants who completed the questionnaire and met the inclusion criteria were contacted by staff via phone or text message and were divided equally into the AI conversation intervention group and the waiting group according to the order in which they were successfully contacted (i.e., the first participants who were successfully contacted were assigned to the intervention group, and the second participants were assigned to the waiting group). Participants in the intervention group received a 28-

day AI conversation intervention after completing the baseline questionnaire. We asked users in the intervention group to use the web link every day, upload the start and end times of their conversations with the AI bot and screenshots of their chats; users who chatted for more than 5 minutes were considered to have successfully punched in for that day. Participants who clocked more than 8 days in the first 14 days were allowed to complete the second questionnaire at the two-week mark. Similarly, participants who quit more than eight times between days 15 and 28 could complete the third questionnaire at the four-week mark. Therefore, all users in the intervention group who completed the third questionnaire received at least 16 AI conversation interventions.

Only the users in the waiting group had the task of waiting for the first four weeks, and they completed the three questionnaires at exactly the same point in time as those in the intervention group. Only after the first four weeks of intervention did the waiting group begin to receive the 28-day conversational intervention, and no additional questionnaires were interspersed in between. This corresponds to each participant completing three questionnaires and submitting up to 28 punch cards. The reward for completing one questionnaire was 10 RMB, and the reward for one day's punch card was 2 RMB, which means that each subject who participated in the entirety of the study was rewarded 86 RMB.

Interventions

Douyin Xinqing is an AI dialog bot nicknamed Xiao Qing (hereinafter uniformly referred to as the Douyin companion bot). Users can start chatting with the Douyin companion bot by clicking on the “Direct Message” on the homepage of their accounts.

After a user sends any message to the Douyin companion bot, it first asks the user about his or her mood and guides him or her to start chatting around his or her emotions. According to the situation described by the user, the companion bot expresses empathy, asks questions to guide the user to talk, and provides suggestions at the appropriate time. Users can also record their emotions every day and look back regularly to observe changes in their inner feelings. In addition, the Douyin companion bot can also identify crisis situations such as suicide and severe mental illness and, if necessary, guide users to call the hotline to obtain professional psychological assistance.

The Douyin companion bot is trained based on a large language model (LLM) on Volcano Ark. A refined multiturn dialog corpus of emotional care scenarios is added to the model so that the model can learn how to understand and reply to the user's emotional topics for multiturn dialog. The training methods include supervised fine-tuning (SFT) and reinforcement learning from human feedback (RLHF):

SFT: In response to simulated messages from the user (Question), a team of experts was invited to write high-quality responses (Answer), and the Q-A data were used to construct a multiturn dialog corpus for fine-tuning (Finetune). The expert team consisted of counsellors, social workers, emotion experts, etc.

RLHF: Many real-life experts ranked the quality of the responses, used the entire ranking to generate a reward model, and continually iterated seeking optimization of the model's responses.

The team's internal and external psychologists also formed an evaluation group to regularly evaluate and score the dialog data of the simulated user and model chats, accounting for multiple dimensions such as empathy, talkativeness, usefulness, and safety to ensure the training effect. The team also established a three-level classification tree of emotional topics and graded them according to light, medium, and heavy to cover all categories and levels of topics.

The refined Multiturn dialog corpus is based on the Problem Management Plus (PM+), a short-term, effective and practical psychosocial program developed by the World Health Organization (WHO) [15], and combines humanistic and SFBT features with an anthropomorphic tone. The corpus includes the following features:

1. Emphasizing the feeling of dialog, leading users to talk through multiple rounds of dialog, and avoiding directly giving hard and generic solutions to users' problems;
2. Make users feel accepted through empathy, making them willing to open up, to achieve total

acceptance without preaching and to provide a harbor for users to talk about their emotions;
3. Focusing on the present without dwelling on the past, with the goal of alleviating negative emotions, and with the flexibility for users to start and exit the dialog at any time;
4. Focusing on emotional companionship rather than counseling, while emphasizing its AI status and not misleading users.

Measures

Depression was assessed with the Patient Health Questionnaire (PHQ-9), which was developed by Kroenke et al. [16], translated into Chinese and revised by Wang et al. [17]. The PHQ-9 contains nine items scored on a scale from 0 (not at all) to 3 (almost every day) based on the self-reported frequency of depressive symptoms in the past 2 weeks. The total score ranges from 0 to 27, and scores above 10 indicate probable depression symptoms [18]. The internal consistency reliability of the scale was acceptable according to the three tests in this study, with Cronbach's α values of 0.84, 0.86, and 0.88.

Anxiety was assessed with the Chinese version of the Generalized Anxiety Disorder Scale (GAD-7) [19]. The 7 items are rated on a 4-point scale ranging from 0 (never) to 3 (almost every day) to assess the frequency of anxiety symptoms in the past 2 weeks. The total score ranges from 0 to 21, and according to the established criteria [20], scores above 10 indicate probable anxiety symptoms. The internal consistency reliability of the scale was acceptable according to the three tests in this study, with Cronbach's α values of 0.88, 0.90, and 0.90.

Positive and negative moods were assessed with the Positive and Negative Affect Schedule (PANAS), which was developed by Watson et al. [21], translated into Chinese and revised by Qiu et al. [22]. The scale was categorized into two dimensions, positive (PA) and negative (NA), and each dimension was scored independently, with 9 items per dimension. All the items are rated on a 5-point scale ranging from 1 (few) to 5 (very much). The higher the subject's score on a subscale is, the stronger the associated emotion experienced. According to the three tests of this study, the Cronbach's alphas were 0.92, 0.93, and 0.94 for the positive mood subscale and 0.88, 0.89, and 0.90 for the negative mood subscale.

Statistical analysis

The data were analyzed using SPSS 26.0. First, descriptive statistics, independent sample *t* tests and one-way ANOVA were used to examine differences in sociodemographic information among groups at baseline. Second, repeated-measures analysis of covariance was used to analyze randomized controlled trials with baseline and follow-up measurements. Third, the Wilcoxon signed-rank test was used to examine the differences in depression, anxiety and cognitive function between baseline and postintervention in each group.

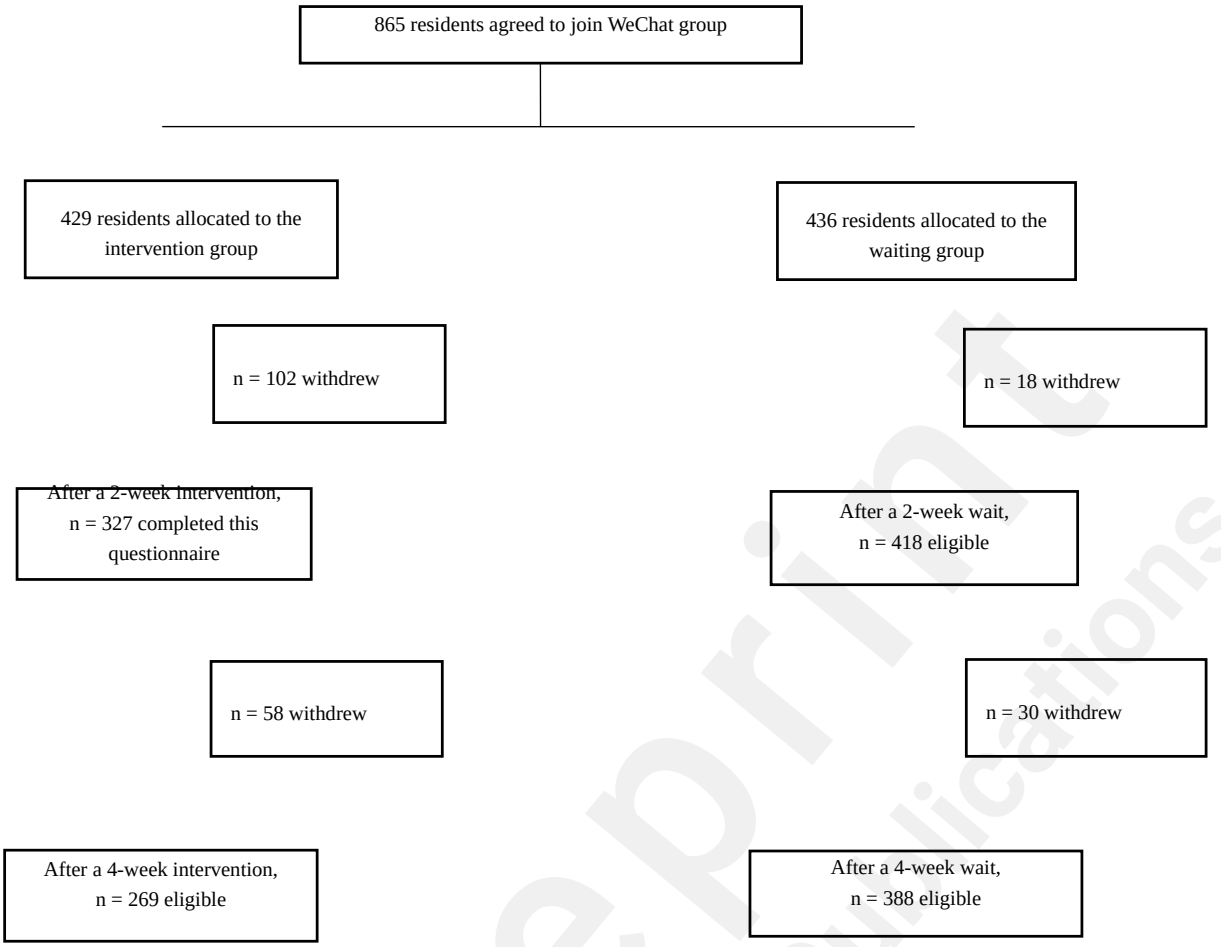
Results

Figure 1 shows the participant flow throughout the study. Between November 13, 2023, and November 19, 2023, a total of 865 people responded to the staff's friend requests and expressed their willingness to participate in the study. Of these, 429 and 436 participants were randomized into intervention and waiting groups, respectively, based on the order of successful contact, and completed the baseline questionnaire on November 20th (see Figure 1). The second questionnaire was administered on December 4, and 76.2% (327/429) and 95.9% (418/436) of the participants in the intervention group and waiting group, respectively, completed the second questionnaire. The third questionnaire was administered on December 18, and another 82.3% (269/327) and 92.8% (388/418) of participants in the intervention and waiting groups, respectively, completed the third questionnaire.

Attrition

The total attrition rate for the study was 24.0% (657/865), and the attrition rates were not the same between the two groups, with the intervention group having a higher attrition rate (37.3% vs. 11.0%; $\chi 2 = 11.26$; $P = .001$). A chi-square test and t test of information pertaining to participants who completed the baseline questionnaire and those who completed the last questionnaire showed that

there was a significant difference in the sex ratio ($\chi 2$ = 37.346, *df* = 1, *P* = .001) between subjects who dropped out and those who continued to participate in the study, with males having a higher



attrition rate. There was a significant difference in age (20.59 vs. 21.07; *t* = -2.897, *df* = 863, *P* = .004), and residents who withdrew from the study were significantly older. There was a significant difference in depression (10.19 vs. 11.15; *t* = -2.065, *df* = 312.46, *P* = .04) and negative affect (21.30 vs. 22.65; *t* = -2.308, *df* = 324.24, *P* = .02) at the first test, with the subjects with attrition having more depressive symptoms and more negative subjective perceptions. There were no significant differences in anxiety (7.71 vs. 8.19; *t* = -1.252, *df* = 316.032, *P* = .21) or positive affect (21.99 vs. 22.64; *t* = -1.198, *df* = 863, *P* = .23) at the first test; i.e., there was no structured attrition.

Figure 1. Recruitment and follow-up of residents in both groups

Description

Most of the residents were female (406 [61.8%] of 657), with a mean age of 20.59 (SD 2.00) years. Most residents were of rural origin (485 [73.8%] of 657) and were enrolled in college (235 [35.8%] of 657) or university (328 [49.9%] of 657). According to the cutoff scores, 291 (44.3%) residents had probable depression, and 154 (23.4%) had probable anxiety. The mean PHQ-9, GAD-7, PA and NA scores were 10.20 (SD = 5.28), 7.71 (SD = 4.42), 22.00 (SD = 6.78) and 21.32 (SD = 6.85), respectively. The baseline characteristics were similar across groups, except for sex (Table 1).

Table 1 Demographic characteristics

	Total	Intervention	Waiting	F	χ^2	*P* value
Residents enrolled	657	269	388			
Sex					*5.073*	*.024*
Female	406 (61.8%)	176 (65.4%)	230 (59.3%)			
Male	251 (38.2%)	93 (34.6%)	158 (40.7%)			

Age, years	20.59 (2.00)	20.55 (2.12)	20.61(1.92)	0.84		.36
Location of residence					0.604	.44
City	172 (26.2%)	75 (27.9%)	97 (25.0%)			
Rural area	485 (73.8%)	194 (72.1%)	291 (75.0%)			
Education level					2.657	.45
Less than university or college	77 (11.7%)	28 (10.4%)	49 (12.6%)			
University	328 (49.9%)	136 (50.6%)	192 (49.5%)			
College	235 (35.8%)	100 (37.2%)	135 (34.8%)			
Graduate level or above	17 (2.6%)	5 (1.8%)	12 (3.1%)			
Baseline data						
Probable depression	291 (44.3%)	113 (42.0%)	178 (45.9%)		0.964	.33
Probable anxiety	154 (23.4%)	62 (23.0%)	92 (23.7%)		0.939	.33
PHQ-9 score	10.20 (5.28)	9.84 (5.14)	10.44 (5.37)	2.051		.15
GAD-7 score	7.71 (4.42)	7.63 (4.50)	7.77 (4.38)	0.164		.69
PA score	22.00 (6.78)	21.94 (6.53)	22.04 (6.95)	0.034		.85
NA score	21.32 (6.85)	21.45 (7.18)	21.22 (6.61)	0.196		.66

Effects of the AI dialogic intervention

The results of the repeated-measures ANOVA indicated a significant main effect of group on each scale on the 14th and 28th days. After 2 weeks of the intervention, the main effects of time (P = .007, η_p^2=0.011) and the main effects of the time-by-group interaction (P = .001, η_p^2=0.016) on the PHQ-9 score were significant. After 4 weeks of the intervention, the main effects of the time by group interaction on the PHQ-9 score (P = .037, η_p^2=0.005) and GAD-7 score (P = .043, η_p^2=0.005) were significant (Table 2).

The Wilcoxon signed-rank test results examining the differences in depression, anxiety, positive emotion and negative emotion between baseline and postintervention in each group are shown in Table 3. After 2 weeks of the AI dialogic intervention and a 2-week wait, scores on all scales were not significantly different from those at baseline for either the intervention or waiting groups. After 4 weeks of intervention and waiting, the PHQ-9 score (t=-2.284, P = .023), GAD score (t=-3.107, P = .002) and NA score (t=-3.602, p < .001) of the residents in the intervention group and the PA score (t=-3.639, p < .001) of the residents in the waiting group were significantly lower than those at baseline.

Table 2 Effect estimates from repeated-measures analysis of covariance

		ANCOVA [a]			ANCOVA [b]		
		F	*P* value	η_p^2	*F*	*P* value	η_p^2
PHQ-9 score	Group	*45.626*	*<.001*	*0.065*	*38.016*	*<.001*	*0.055*
	Time	*7.241*	*.007*	*0.011*	1.696	.18	0.003
	Time×Group	*10.253*	*.001*	*0.016*	*3.296*	*.04*	*0.005*
GAD-7 score	Group	*35.329*	*<.001*	*0.051*	*35.742*	*<.001*	*0.052*
	Time	0.432	.51	0.001	2.203	.11	0.003
	Time×Group	0.740	.39	0.001	*3.155*	*.04*	*0.005*
PA score	Group	*83.058*	*<.001*	*0.113*	*189.14*	*<.001*	*0.225*
	Time	2.587	.11	0.004	0.679	.51	0.001
	Time×Group	3.236	.07	0.005	1.319	.27	0.002
NA score	Group	*100.31*	*<.001*	*0.134*	*136.43*	*<.001*	*0.173*
	Time	0.051	.82	<0.001	0.074	.93	<0.001
	Time×Group	0.015	.90	<0.001	0.180	.84	<0.001

Notes: [a] ANCOVA results comparing assessment scores of the 2-week intervention group and waiting group over time, controlling for sex, number of words in conversation, number of days in conversation and average number of words in conversation per day at 14 days. [b] ANCOVA results comparing assessment scores of the 4-week intervention group and waiting group over time, controlling for sex, number of words in conversation, number of days in conversation and average number of words in conversation per day at 28 days.

Table 3 Effects of within-subject intervention

	Intervention Group (N=269)						
	Baseline	2-week	*t*	*p* value	4-week	*T*	*p* value
PHQ-9 score	9.84 (5.14)	9.70 (5.71)	-0.497	.62	*9.13 (5.77)*	*-2.284*	*.02*
GAD-7 score	7.63 (4.49)	7.43 (4.58)	-0.801	.42	*6.86 (4.60)*	*-3.107*	*.002*
PA score	21.94 (6.53)	22.10 (6.88)	0.449	.65	22.41 (7.18)	1.064	.29
NA score	21.46 (7.18)	21.08 (7.29)	-0.944	.35	*19.94 (6.93)*	*-3.602*	*<.001*
	Waiting Group (N=388)						
	Baseline	2-week	*t*	*p* value	4-week	*t*	*p* value
PHQ-9 score	10.44 (5.37)	10.17 (5.24)	-1.198	.23	10.18 (5.31)	-1.114	.27
GAD-7 score	7.77 (4.38)	7.70 (4.32)	-0.377	.71	7.54 (4.24)	-1.137	.26
PA score	22.04 (6.95)	21.72 (7.12)	-1.124	.26	*21.07 (7.36)*	*-3.639*	*<0.001*
NA score	21.22 (6.61)	21.03 (6.56)	-0.585	.56	20.66 (7.13)	-1.751	.08

[a] Data are presented as the means (SD).

Discussion

Principal Results

The purpose of this study was to investigate whether an AI companion bot trained on a large language model and developed to intervene in the emotional problems of young adults can provide sufficient companionship and support through dialog with users and ultimately alleviate their emotional problems effectively. To this end, we designed a 28-day randomized controlled trial to verify the effectiveness of the intervention.

Users with mild symptoms of depression or anxiety were screened for participation in the study (according to the cutoff, the prevalence of mild depression and anxiety at baseline was 44.3% and 23.4%, respectively). The results showed that after 14 days of dialog intervention, there was no significant difference from baseline in any of the moods in the intervention or waiting groups. After 28 days of the dialog intervention, the intervention group showed a significant decrease in depression, anxiety, and negative mood and no significant change in positive mood, whereas the waiting group showed a significant decrease in positive mood. The natural decrease in positive mood in the waiting group may be because, in the initial instructions, we disclosed to each participant that this was an AI dialog intervention study. The intervention group received the information at the first opportunity, whereas the waiting group experienced a long wait during the first 28 days. This can be disappointing for participants who joined the study with curiosity about AI dialog but were randomly assigned to the waiting group.

The results of the repeated-measures ANOVA showed that the dialog intervention significantly reduced depression in the intervention group at two weeks and significantly reduced both depression and anxiety in the intervention group at four weeks. This finding suggests that the dialog intervention may have a more immediate relieving effect on depression (at two weeks, but significant relief at four weeks) but that it takes some time for the dialog intervention to have an intervention effect on anxiety (four weeks). This result may complement the results of a recently published meta-analysis showing that for depression, intervention studies lasting 0-8 weeks and 9-16 weeks resulted in greater intervention effects than those with durations greater than or equal to 17 weeks [23]. This suggests that, for dialog intervention studies targeting depression, the duration of the intervention can be a predictive variable and that the effects of such interventions may be cumulative and ultimately significant over time. Two weeks may be a meaningful point in time, and dialog interventions lasting two weeks can play a role in alleviating depression. Although two weeks of dialog intervention was unable to significantly mitigate users' anxiety, this outcome conflicts with the findings of previous research. A CBT-based dialog intervention study showed that after a two-week intervention, there was a significant reduction in users' anxiety compared to baseline [24]. This may have been determined by the difference in anxiety of the subject groups at baseline. The results of previous meta-analyses have shown [23] that dialog interventions yield better results in groups with high anxiety at baseline than in groups with low anxiety. In contrast, the participants in the present study had a GAD score of 7.71 at baseline, which was much lower than the approximately 18.5 reported in the study by Fitzpatrick et al. (overall levels were not reported; 19.02 for the information control group vs. 18.05 for the intervention group). The results of the present study complement the findings

of the meta-analysis, and the fact that our dialog intervention was able to take effect at 4 weeks in the less anxious group in turn corroborates the effectiveness of the Douyin companion bot.

Strengths and limitations

Most previous studies have taken CBT as the underlying intervention logic, supplemented with positive psychology components. The control groups were also mostly controlled by adopting learning e-books [24][25], or no control group was set up [26]. In these studies, there were also few measures available to ensure that subjects had stable and consistent daily clock-in behavior throughout the intervention process, and valid intervention information could be captured only through subsequent qualitative analyses to ensure that subjects were truly communicating and interacting emotionally with the conversational agent. All of the above shortcomings were addressed in our study. First, the intervention logic behind the Douyin companion bot was not CBT; it was trained based on a large language model on Volcano Ark, and the intervention logic combines humanistic psychology, positive psychology, positive mindfulness meditation, focused solution and PM+, among other schools of thought and is itself novel and original. In our study, 59.1% of the participants who completed the clock-in task (those who clocked in ≥16 times during the 28-day dialog task) indicated that they felt that the dialog with the dialog with the Douyin companion bot was fun; 56.5% indicated that monetary incentives encouraged them to persevere with the dialog task; 50.9% of the participants indicated that they felt accompanied; 40.9% indicated that they felt comforted and understood; and 27.2% indicated that they felt valued. Importantly, even though the percentage of individuals reporting that they participated in the study out of a desire to obtain money was greater than 50%, we should not assume that this biases the results. First, with the period of this study (28 days for the intervention group and 56 days for the waiting group), a monetary incentive of RMB 86 is a very small amount per day; second, the monetary incentive is just a way for us to maintain our participants, who needed only to complete the questionnaire results truthfully and submit their clock-in records to be paid for their participation, which did not affect the scores of their questionnaires or the content of their chats with the Douyin companion bot. In addition to monetary incentives, a considerable number of participants also felt the dialog process was fun and had the warm experience of being empathized with, understood and accompanied. Second, the use of a nonblank control may prevent the intervention from being accurately separated, which was not a problem in our study. In fact, a previous study found that after two weeks of dialog intervention and e-book learning, anxiety decreased significantly in both the intervention and control groups [24], which prevented the researchers from being able to determine whether it the dialog intervention was working. In our study, the so-called control group was, in fact, a waiting group. The waiting group received the dialog intervention in its entirety after 28 days of blank control, which on the one hand allowed us to ensure that no other intervention components (e.g., e-book learning) were working on the control group; on the other hand, it ensured that our study was ethical and humane, i.e., that every subject wishing to take part in the dialog intervention could experience this process entirely without being randomly being assigned to the control group and thus missing the opportunity. Finally, we set a requirement for the number of clock-ins performed for participants in the intervention group. Only subjects who met the requirement (≥8 days in a fortnight) were able to complete the second and third questionnaires. Participants were reminded to clock-in every day. On the one hand, previous research has shown that conversational agents with regular reminders are more effective in intervening on subjects' emotional problems than those without reminders [23]. On the other hand, it also ensures that our subjects participated in the dialog intervention for depressed and anxious moods in a stable and consistent manner.

This study has several limitations. First, even though our study included two follow-up surveys within four weeks and drew conclusions about the immediate intervention effect and the long-term intervention effect of the Douyin companion bot, the four-week duration itself was still a short-term intervention, and a longer dialog intervention might have provided more effective clues about the long-term intervention effects of the Douyin companion bot. Second, there was a significant sex

difference between the intervention and waiting groups of our recruited participants, and overall, there were far more females than males (406 vs. 251). This was in part due to our recruitment process. We rotated the participants into intervention and waiting groups in the order in which they contacted the staff, which ensured that our groupings were randomized but was not an unbiased way of grouping. After the grouping was completed, we could not artificially adjust the sex ratio between the two groups. Therefore, we controlled for sex as a covariate in all subsequent data analyses. Overall, the number of females participating in the study was much higher than that of males, which may also reflect, to some extent, the greater willingness of females than males to participate in Internet-based intervention studies, which has been similarly inferred in previous studies [27]. Third, there was structured attrition of our subjects, with the results of the chi-square test indicating that males, older subjects, and subjects with more negative emotions at baseline were more likely to leave the study. This result supports the idea that girls are more willing to participate in internet-based intervention studies; on the other hand, our dialog intervention study, although not heavy in terms of daily tasks, was a project that was maintained for 4 weeks and gave little monetary incentive per day. Older participants were likely to have less daily time and more earning power; therefore, a small monetary incentive did not encourage them to persist in participating in the study for 28 days. Additionally, subjects who had more negative moods at baseline may have perceived less fun during the conversations, which became one of the reasons they disengaged from the study.

Conclusions

LLM-based AI conversational agent can effectively alleviate the depression and anxiety of young adults with mild negative emotions through a four-week dialog intervention and can have an intervention effect on alleviating depression in users after two weeks. As a convenient and easily accessible conversational agent, the AI companion bot provides users with companionship and comfort while lowering the threshold for people to seek psychological help and obtain social support. Importantly, the main purpose of an AI companion bot is to provide companionship, understanding and empathy to young adults in need; it currently has major limitations, can intervene in only mild depression and anxiety, and cannot address mood disorders or more serious psychological problems.

Acknowledgements

We extend our sincere thanks to all participants and organizations that supported this research. Furthermore, we are grateful for the assistance of Fan Yang, Maidina Nijiati, Xue Zhang, Changhua Liu, Qian Chen, Meilin Chen, Yixuan Liu, Yijin Huang, Meixuan Lv and Pu Gong in the acquisition of experimental materials as well as in data collection.

Data Availability

Supporting data for this study can be requested through the corresponding author.

Conflicts of Interest

None declared.

Reference

[1] Barney, L. J., Griffiths, K. M., Jorm, A. F., & Christensen, H. (2006). Stigma about depression and its impact on help-seeking intentions. *Australian & New Zealand Journal of Psychiatry*, *40*(1), 51-54. doi:10.1080/j.1440-1614.2006.01741.x

[2] Ebert, D. D., Van Daele, T., Nordgreen, T., Karekla, M., Compare, A., Zarbo, C., ... & Baumeister, H. (2018). Internet-and mobile-based psychological interventions: applications, efficacy, and potential for improving mental health. *European Psychologist*. https://doi.org/10.1027/1016-9040/a000318

[3] Yap, M. B. H., Reavley, N., & Jorm, A. F. (2013). Where would young people seek help for mental disorders and what stops them? Findings from an Australian national survey. *Journal of affective disorders*, *147*(1-3), 255-261. https://doi.org/10.1016/j.jad.2012.11.014

[4] Wang, Q., Zhang, W., & An, S. (2023). A systematic review and meta-analysis of Internet-based self-help interventions for mental health among adolescents and college students. *Internet interventions*, *34*, 100690. https://doi.org/10.1016/j.invent.2023.100690

[5] Enrique, A., Palacios, J. E., Ryan, H., & Richards, D. (2019). Exploring the Relationship Between Usage and Outcomes of an Internet-Based Intervention for Individuals With Depressive Symptoms: Secondary Analysis of Data From a Randomized Controlled Trial. *Journal of medical Internet research*, *21*(8), e12775. https://doi.org/10.2196/12775. PMID: 31373272

[6] Rollman, B. L., Belnap, B. H., Abebe, K. Z., Spring, M. B., Rotondi, A. J., Rothenberger, S. D., & Karp, J. F. (2018). Effectiveness of online collaborative care for treating mood and anxiety disorders in primary care: a randomized clinical trial. *JAMA psychiatry*, *75*(1), 56-64. doi:10.1001/jamapsychiatry.2017.3379

[7] Hoermann, S., McCabe, K. L., Milne, D. N., & Calvo, R. A. (2017). Application of synchronous text-based dialogue systems in mental health interventions: systematic review. *Journal of medical Internet research*, *19*(8), e267. https://doi.org/10.2196/jmir.7023. PMID: 28784594

[8] Gaffney, H., Mansell, W., & Tai, S. (2019). Conversational agents in the treatment of mental health problems: mixed-method systematic review. *JMIR mental health*, *6*(10), e14166. doi: 10.2196/14166. PMID: 31628789

[9] Scholten, M. R., Kelders, S. M., & Van Gemert-Pijnen, J. E. (2017). Self-guided web-based interventions: scoping review on user needs and the potential of embodied conversational agents to address them. *Journal of medical Internet research*, *19*(11), e383. doi: 10.2196/jmir.7351. PMID: 29146567

[10] Zivin, K., Eisenberg, D., Gollust, S. E., & Golberstein, E. (2009). Persistence of mental health problems and needs in a college student population. *Journal of affective disorders*, *117*(3), 180–185. https://doi.org/10.1016/j.jad.2009.01.001

[11] Ramón-Arbués, E., Gea-Caballero, V., Granada-López, J. M., Juárez-Vela, R., Pellicer-García, B., & Antón-Solanas, I. (2020). The prevalence of depression, anxiety and stress and their associated factors in college students. *International journal of environmental research and public health*, *17*(19), 7001. **https://doi.org/10.3390/ijerph17197001**

[12] Fernández Rodríguez, C., Soto López, T., & Cuesta Izquierdo, M. (2019). Needs and demands for psychological care in university students. *Psicothema*. doi: 10.7334/psicothema2019.78

[13] Shah, T. D., & Pol, T. (2020). Prevalence of depression and anxiety in college students. *Journal of Mental Health and Human Behaviour*, *25*(1), 10-13. *DOI:* 10.4103/jmhhb.jmhhb_16_20

[14] Kaye, D. B. V., Chen, X., and Zeng, J. (2021). The co-evolution of two Chinese mobile short video apps: parallel platformization of Douyin and TikTok. *Mobile Media Commun.* 9, 229–253. doi: 10.1177/2050157920952120

[15] Dawson, K. S., Bryant, R. A., Harper, M., Tay, A. K., Rahman, A., Schafer, A., & Van Ommeren, M. (2015). Problem Management Plus (PM+): a WHO transdiagnostic psychological intervention for common mental health problems. *World Psychiatry*, *14*(3), 354. doi: 10.1002/wps.20255

[16] Kroenke, K., Spitzer, R. L., & Williams, J. B. (2001). The PHQ-9: validity of a brief depression severity measure. *Journal of general internal medicine*, *16*(9), 606-613. https://doi.org/10.1046/j.1525-1497.2001.016009606.x

[17] Wang, W., Bian, Q., Zhao, Y., Li, X., Wang, W., Du, J., Zhang, G., Zhou, Q., & Zhao, M. (2014). Reliability and validity of the Chinese version of the Patient Health Questionnaire (PHQ-9) in the general population. *General Hospital Psychiatry*, *36*(5), 539–544. https://doi.org/10.1016/j.genhosppsych.2014.05.021

[18] Kocalevent, R., Hinz, A., & Brähler, E. (2013). Standardization of the depression screener Patient Health Questionnaire (PHQ-9) in the general population. *General Hospital Psychiatry*, *35*(5), 551–555. https://doi.org/10.1016/j.genhosppsych.2013.04.006

[19] He, X., Li, C., Qian, J., & Wu, W. (2010). Reliability and validity of a generalized anxiety scale in general hospital outpatients. *ResearchGate*. https://doi.org/10.3969/j.issn.1002-0829.2010.04.002

[20] Löwe, B., Decker, O., Müller, S., Brähler, E., Schellberg, D., Herzog, W., & Herzberg, P. Y. (2008). Validation and standardization of the Generalized Anxiety Disorder screener (GAD-7) in the general population. *Medical Care*, *46*(3), 266–274. https://doi.org/10.1097/mlr.0b013e318160d093

[21] Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: the PANAS scales. *Journal of personality and social psychology*, *54*(6), 1063. https://doi.org/10.1037/0022-3514.54.6.1063

[22] Qiu, L., Zheng, X., & Wang, Y. F. (2008). Revision of the positive affect and negative affect scale. *Chinese Journal of Applied Psychology*, *14*(3), 249-254.

[23] He, Y., Yang, L., Qian, C., Li, T., Su, Z., Zhang, Q., & Hou, X. (2023). Conversational agent interventions for mental health problems: systematic review and meta-analysis of randomized controlled trials. *Journal of Medical Internet Research*, *25*, e43862. doi: 10.2196/43862. PMID: 37115595

[24] 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. *JMIR mental health*, *4*(2), e7785. doi: 10.2196/mental.7785. PMID: 28588005

[25] Fulmer, R., Joerin, A., Gentile, B., Lakerink, L., & Rauws, M. (2018). Using psychological artificial intelligence (Tess) to relieve symptoms of depression and anxiety: randomized controlled trial. *JMIR mental health*, *5*(4), e9782. doi: 10.2196/mental.9782. PMID: 30545815

[26] Gabrielli, S., Rizzi, S., Bassi, G., Carbone, S., Maimone, R., Marchesoni, M., & Forti, S. (2021). Engagement and effectiveness of a healthy-coping intervention via chatbot for university students during the COVID-19 pandemic: mixed methods proof-of-concept study. *JMIR mHealth and uHealth*, *9*(5), e27965. doi: 10.2196/27965. PMID: 33950849

[27] Su, W., Fang, X., Miller, J. K., & Wang, Y. (2011). Internet-based intervention for the treatment of online addiction for college students in China: a pilot study of the Healthy Online Self-hel** Center. *Cyberpsychology, Behavior, and Social Networking*, *14*(9), 497-503. http://doi.org/10.1089/cyber.2010.0167

Abbreviations

LLM: large language model
IMIs: Internet- and mobile-based psychological interventions
RCT: randomized controlled trial
PHQ-9: the Patient Health Questionnaire
GAD-7: the Generalized Anxiety Disorder Scale
PANAS: the Positive and Negative Affect Schedule
PA: positive dimension in PANAS
NA: negative dimension in PANAS
CSR: Corporate Social Responsibility
SFBT: solution-focused brief therapy
PM+: problem management plus
SFT: supervised fine-tuning
RLHF: reinforcement learning from human feedback
WHO: World Health Organization
ANOVA: analysis of variance
ANCOVA: analysis of covariance
CBT: cognitive-behavioral therapy