Marcus: A Chatbot for Depression Screening Based on the PHQ-9 Assessment

Evaluating accuracy and effectivess in college students based on gender and age during the pandemic

Patrick Toulme
AWS Artificial Intelligence
Amazon
Arlington, Virginia, USA
ptoulme@amazon.com

Jude Nanaw
Department of Computer Science
University of Virginia
Charlottesville VA, USA
jn7tez@virginia.edu

Panagiotis Apostolellis
Department of Computer Science
University of Virginia
Charlottesville VA, USA
panaga@virginia.edu

***Abstract*—College students are a population particularly susceptible to anxiety and depression, with financial struggles and social stigmatization creating a barrier to seeking psychological support. The recent pandemic has exacerbated these issues, with lockdowns and remote instruction creating extra stress factors and making access to consultation services even harder. Online depression screening tools have tried to address such problems and the development of chatbots for the detection and even therapeutic use of anxiety symptoms has been on the rise. This work reports findings from testing Marcus, a depression screening chatbot based on a popular depression assessment tool, the Patient Health Questionnaire (PHQ-9). Our results indicate that Marcus was comparable to the online version of PHQ-9 in detecting depression based on produced scores, using a within-subjects experimental design with predominantly college students in the USA. Nonetheless, the chatbot was not found to be the most effective method based on comparing participant preferences and initiation rates. Implications of our findings for the development of similar computer-based screening tools are discussed, as well as recommendations for future work in this area.**

Keywords-chatbot design; medical application; computer-based depression screening; user study.

I. INTRODUCTION

Among the various mental health issues faced by millions of individuals worldwide, depression is considered one of the most prevalent with over 280 million people of all ages globally affected by the disorder, being especially prevalent among 15-29-year-olds [1]. Continuous struggle with depression has proven to have personal adverse effects on individuals and has been linked to issues such as low socioeconomic status [2] , family functioning [3], and diminished social support [3]. Among the population age composition, university students are a particular group with exceptionally high depression rates [4]. Despite the detrimental impact of incessant depression including an increased risk of suicide by a factor of 20 among depressed populations compared to non-depressed populations, [5], a significant portion of these individuals do not seek treatment. Studies have found that the most common reasons for not seeking treatment among university students include the stigma associated with mental health issues [6], as well as feeling a lack of perceived need [7] and insufficient mental health education [8]. Moreover, the lack of easy access to depression treatment poses another barrier for university students. Economic barriers are a major cause for not seeking screening or treatment, as associations exist between low socioeconomic status and depression rates [9]. With depression being incredibly widespread, all aspects of treatment are in high demand and a public need, including both therapeutic interventions as well as screening and assessment needs.

To counteract the significant barrier of stigmatization and financial ability, which is especially prominent for university students, online depression screening tools and chatbots have been suggested as a viable solution [10]. The development of such chatbots removes the need for interpersonal communication in this preliminary step of the treatment process, making screening more accessible. The PHQ-9 is the most viable depression screening tool in the industry of primary care and uses a four-point scale to reveal the tendency to depression ranging from minimal to severe depression [11]. With the advancement of artificial intelligence and Natural Language Processing (NLP), there have been various developments of mental health chatbots that focus on anxiety screening using the PHQ-9 [12]–[14], as well as the therapeutic treatment and reduction of depression-like symptoms [15][16].

Additionally, the COVID-19 pandemic drove a fundamental change within the healthcare delivery system due to unprecedented challenges such as social distancing guidelines and stay-at-home orders [17]. In large part, the pandemic accelerated the rise of telehealth, where all healthcare services, including screening, were provided remotely–and oftentimes via virtual agents. Moreover, the pandemic saw the increased development of chatbots that screened for COVID-19, which delivered consistent and accurate results while also providing sustained service at a low operating cost [18]. These systems exemplified the potential of telehealth and chatbots in terms of overcoming obstacles, such as costs and physical barriers that prevent individuals from receiving care.

Inspired by the increased demands caused by the pandemic for easy access to anxiety screening explicitly for college students, we developed Marcus, a chatbot for depression screening, In the current study, we briefly describe our mobile-based chatbot and report results from a study with university students in the United States. We hypothesized that our virtual screening tool will be equally effective in detecting depression as the traditional online PHQ-9. Findings and design implications for similar computer-based screening tools are discussed in light of prior research.

Section II presents a review of related work including background on the PHQ-9, as well as existing virtual agents utilized for depression screening and therapeutic methods.

 ISBN: 978-1-68558-078-0

Section III includes the driving research questions and details regarding the design of the chatbot. Section IV presents the research methods in regards to participants and the data collection process. Section V describes our results including the various statistical analysis used and graphical representations of major findings. Section VI discusses our main findings in light of prior research and acknowledges study limitations. Section V summarizes our research work and presents directions for future research based on our findings.

II. Background

This research builds on past work related to chatbots and virtual agents designed to screen for depression or act as a therapeutic agent to reduce depression-like symptoms.

A. Diagnostic Evaluations: Framework and Administration

The original PHQ-9 is viewed as a dual-purpose instrument, which provides not only provisional depression diagnoses, but also grade depressive symptom severity [19]. Over the years, the nine-item scale has been validated as a depression screening tool in the primary care industry and remains widely utilized [20]. The assessment itself features a straightforward scoring methodology, where each question is rated by the patient with a frequency from 0 to 3 – with 0 indicating "not at all" and 3 indicating "nearly every day". A summation of the scores is performed to provide a final score from 0 to 27, with a higher score representing a greater level of severity.

A study conducted with university students in Iran examined the validity and reliability of the PHQ-9 alongside other assessments [20]. Students completed the self-administered version of the PHQ-9, as well as psychiatrist interviews to determine depression level – with the tool ultimately demonstrating a satisfactory internal consistency. Another study conducted in Kosovo measured depression during the COVID-19 outbreak with an online version of the PHQ-9 administered through Google Forms [21]. Results exhibited a higher-than-average percentage of participants with moderate to severe symptoms. Additionally, a pilot evaluation investigated links between depression references on social media (Facebook) by undergraduate students and depression on a clinical scale using an online version of the PHQ-9 [22]. Over 70% of eligible profile owners participated in the online PHQ-9 survey, indicating a successful administration. Results of the study demonstrated a positive association between the two, with participants who scored higher into a depression category on the PHQ-9, being more likely to display depression symptom references on Facebook.

Prior work has also investigated the effect of gender on depression diagnosis, both in the general population and specifically college students [23]. Research has indicated that when compared to males, females account for a larger proportion of patients with depression [24]. Moreover, studies have shown that the gender differences regarding depression rates are more significant at younger ages than when at an adult age [25]. In addition to the gender depression variance being linked by hormonal differences, studies on the respective clinical aspects have pointed to socialization differences also being a factor in depression rates [26].

B. Computer-Based Methods for Depression Screening

The PHQ-9 has seen adaptations into chatbot versions of the assessment in recent years. The virtual agent implementation aims to bring various benefits to the forefront, including the ability to screen for depressive symptoms remotely on a large scale and at a low cost [13]. The study utilized a chatbot named "Tess", which inquired about the nine PHQ-9 criteria posed on the questionnaire, seeking to discover relationships between demographic variables and PHQ-9 scores by administering the assessment to adults and older adults. The chatbot demonstrated strong reliability in both results and completion rate. While results indicated a correlation between demographic characteristics and PHQ-9 score, the associated effect of this was deemed as weak. As the study primarily recruited adults and older adults (above the age of 65), this posed the limitation of a lack of focus on a population that is more susceptible to depression or depression-like symptoms, such as university students. Moreover, the study only administers the PHQ-9 through the Tess chatbot and not through another means (i.e., online survey or paper format) for validation purposes, thus posing another limitation.

Another study conducted in Spain delivered the development of "Perla," a conversational agent that performed a depression screening interview based on the PHQ-9 [12]. The review found that the chatbot was preferred by internet users more than the form-based questionnaire, and the results were consistent enough to deem the chatbot as a valid alternative to traditional self-report methods. Opportunities for expanding on the study include further research into user preferences, such as having a conversational agent with human face and name characteristics to provide further appeal. Moreover, another work identified employees and those in the workplace as a group with specifically high potential for exposure to mental health problems [14]. The study featured the development of a fully automated chatbot "Viki," which evaluated workers for risks of suffering from depression, anxiety, stress, and burnout. Results found that the conversation and gamification style of the chatbot delivered potential for greater engagement and effectiveness.

Additional studies considered groups with higher possible exposure to depression by screening via a chatbot. During the COVID-19 pandemic, frontline workers were especially impacted, and new tools were necessary to identify individuals that were in need of treatment, especially among those who feared stigma around mental health [27]. The study considered a text interface as well as a conversational speech chatbot based on the PHQ-9 for evaluation, with feasibility based on the Technology Acceptance Model [28]. Results of the study demonstrated that most participants found the chatbot to be acceptable, with perceived usefulness and prior depression-like symptoms being the two most significant factors in predicting the inclination of participants to use the chatbot [27].

C. Therapeutic Chatbots to Reduce Depression-Like Symptoms

In addition to screening chatbots based on the adaptation of the PHQ-9 assessment, strides have also been made into the development of chatbots that act as therapeutic agents. With the purpose of improving mental health via the reduction of

 ISBN: 978-1-68558-078-0

depression-like symptoms, therapeutic virtual agents utilize the PHQ-9 as a measurement tool both at baseline and post-treatment stages [29]. A study conducted with university students administered the PHQ-9 – in addition to other questionnaires as measures of separate clinical variables – at baseline and every 4 weeks throughout the 16-week period [29]. During the period, students were randomly assigned to receive either therapy from the chatbot or minimal level bibliotherapy. Results demonstrated that the chatbot-delivered intervention was significantly more efficacious than bibliotherapy, with PHQ-9 scores being reduced further with the virtual agent. In another study, various precursors to depression and other mental health disorders were identified [30]. Through the development of a virtual agent named "Elomia," the study delivered therapy to university students who indicated some level or susceptibility to depression. Results revealed that users of Elomia described a reduction in anxiety and depression symptoms – in addition to 70% of users who noted returning to the chatbot in moments of stress or other related symptoms.

Moreover, additional works have placed an emphasis on groups who have the potential to be more susceptible to mental health issues. In one study, chatbot-based treatment was provided to a post-partum population [16]. Participants either received treatment via chatbot or through traditional means with the PHQ-9 and General Anxiety Disorder (GAD-7) among other assessments, were administered to establish a baseline. While scores did not differ significantly between the varying treatment groups, a large proportion of women (74%) indicated use of the chatbot – demonstrating a greater willingness to interact with the virtual agent. Research conducted in India identified the student population of ages 15-25 as sufferers of mental health issues for a variety of reasons including method of education and the high expectations from family and friends [10]. The study additionally noted the lack of willingness of those affected by depression or precursors of depression (e.g., other emotional issues, social anxiety) to voice their circumstances. Therefore, the development of "CareBot"—a therapeutic virtual agent—aimed at providing similar support as that of a counselor or therapist. While the chatbot was not deemed a viable solution as a substitute for a psychologist, the tool did serve as a provider of conversation allowing users to speak out their problems.

Technologies such as DialogFlow and backend applications like NodeJS and Firebase are prevalent in the development of PHQ-9 based screening and therapeutic chatbots [31]. However, we identified a gap in how such technologies can be harnessed to fulfill their significant potential for an explicitly sensitive group to depression, college students, especially during the high-stress environment of the recent pandemic.

III. MARCUS: CHATBOT FOR DEPRESSION SCREENING

This section presents our research questions and the design of Marcus, the chatbot, for testing those questions.

A. Research Questions

Considering there is no published research on the use of chatbots, including their accuracy and effectiveness, explicitly for college students in the U.S., we decided to develop a mobile-based chatbot as a testbed for responding to the following two research questions:

1) Is Marcus an accurate depression screening method as compared to a more conventional online tool, especially for college students? What is the effect of gender, if any?

2) Is Marcus perceived as an effective tool for depression screening by U.S. university students, taking into account interactions with the chabot and selft-reported measures?

Our hypothesis for the first question is based on prior research, especially since we followed the work by [12], and is that the chatbot will be as accurate in detecting depression as the online PHQ-9. We had no specific direction for our hypothesis regarding the effect of gender because no prior work investigated an effect of gender on chatbot vs traditional screening accuracy. As for the second question, we assumed Marcus would be perceived positively by most participants as reported by prior work using screening [12] and therapeutic chatbots [10] [30].

B. Marcus Chatbot Design

When designing Marcus, we faced two important design decisions: the visual representation of the chatbot and the format of user input. Considering we were addressing a younger audience of college students we opted for simulating a more realistic representation of the chatbot, using a young male image that aligns with the name Marcus, to trigger a higher affinity with the chatbot [32]. As for user input, we started with pre-defined multiple-choice options resembling the four levels of the original PHQ-9 about the frequency of experienced depression symptoms. However, this type of input was not deemed natural enough, as we wanted to simulate the chatting that our young population is used to in their everyday digitally enabled interactions. A brief and informal pilot study with undergraduate students at the university using an earlier implementation of Marcus confirmed this assumption. Therefore, we chose to use free-text input and employ NLP to categorize the text input into one of the PHQ-9 levels, despite understanding the challenges involved in this approach in terms of classification accuracy [33].

Based on the decision above, Marcus was developed using a BERT machine learning model and a variety of APIs and platforms. BERT stands for Bidirectional Encoder Representations from Transformers and was developed in 2018 by researchers at Google Artificial Intelligence as a large service-based model for NLP [34]. BERT models are trained for a variety of language tasks, such as sentiment analysis, which Marcus uses to analyze the user's response for positive or negative sentiment. The backend of Marcus is using Google DialogFlow [35], which handles the NLP through a system of intents, entities, and phrase training. A BERT machine learning model is then generated based on the developer-provided training data to handle all inference requests by the user. Intents are the question-response pairs that are expected in the conversation flow, and entities are groups of words detected within the conversation with an integer value assigned. These entities are what the BERT model uses when scoring the users' responses after an inference call is made.

 ISBN: 978-1-68558-078-0

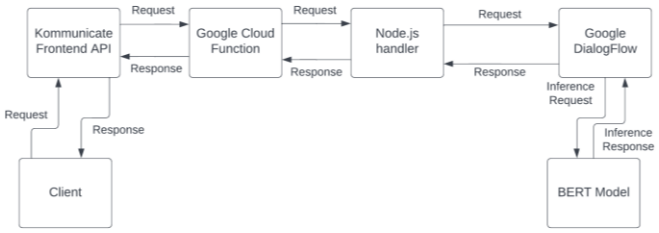

Figure 1. Schematic representation of technologies used for Marcus.

The BERT model performs sentiment analysis on the user's input and assigns a numeric score (i.e., 0-3 from the original PHQ-9 four-item scale) to the user's phrase based on the entities matched in the training data provided.

The technical workflow of Marcus (Figure 1) involves routing the user's natural language through Google Dialog-Flow fulfillment with each intent linked to a different scripted method in the fulfillment code base. This workflow was written in Node.js and is hosted on a serverless cloud function on Google Cloud Platform. An inference call to DialogFlow is then made from each of these workflows to score the user's response. The numeric score output by the BERT model is then routed to the Node.js handler, which outputs a success code 200 response to the Google Cloud Function, which in turn triggers the next question to be displayed to the user via the frontend API and finally back to the client. Kommunicate, a tool for automating conversations with a chatbot [36], was used for the frontend implementation of Marcus, as it allowed interfacing with all the technologies used and enabled deployment on multiple platforms. The original dataset of phrases with corresponding PHQ-9 scores comes from Perla (translated from Spanish), which is also a BERT-modeled chatbot version of PHQ-9 [12]. Additional phrases were scored and added to the dataset through multiple iterations of piloting the chatbot with undergraduate university students, including the researchers.

Marcus was developed for iOS and as a web application. iOS was chosen due to the ease of development, large support system, large number of APIs available, and the large prevalence of iOS devices amongst college students [37]. iOS was also chosen as the primary user platform for its consistent user interface across devices and overall ease of use for the user. The web application serves as an alternative for participants who do not have access to an iOS device. Marcus is embedded in the iOS application through the Kommunicate API using native Swift coding and providing the chatbot's visual interface (Figure 2). We slightly revised the wording of subsequent questions after the first one to make the conversation appear more natural, since the questionnaire reads differently in a multiple-choice format (e.g., all questions were preceded with a phrase like "How often over the past two weeks…"). We also considered inserting extra prompts and phrases between questions to increase the naturalness of the conversation, but after consulting with a psychometrician from the university's counselling and psychological services, we were advised against this practice, as it would potentially compromise the validity of the instrument.

Figure 2. Screenshot of typical Marcus conversation on an iPhone.

IV. METHODS

The research study was approved by the Institutional Review Board of the University of Virginia with protocol IRB-SBS#4005/2022-01-20. The period of data collection was between March and June of 2022, during a semester when the university was transitioning to removing protections and the use of masks in classrooms.

A. Participants

Participants were recruited mainly through email listservs at the University of Virginia and consisted of mostly Engineering and Psychology undergraduate students. LinkedIn and additional social media were also used but college students were prioritized to address the research questions regarding depression screening in college aged populations. A total of 187 people started the survey, but 57 participants did not consent or dropped before being shown either the PHQ-9 or Marcus and are excluded from analysis. Out of the remaining 130 people, 72 participants reported demographic information, including age, gender, and education level. The majority of participants identified as female (N = 46, 63.89%) and male (N = 23, 31.94%), but there were also three participants who identified as neither (N = 3, 4.17%). The largest age group was participants who reported being 18-24 years old (N = 65, 90.28%), which falls within the demographic under investigation. Most participants reported that their highest level of education received was an undergraduate college degree. Nine extra participants, for a total of 81, completed both assessments but had incomplete demographic information. For the 72 participants who fully completed the survey, the demographic information results are shown in Table I.

TABLE I. DEMOGRAPHIC BREAKDOWN OF PARTICIPANT DATA

Demo-graphic	N (%)	Marcus Mean	Marcus SD	PHQ-9 Mean	PHQ-9 SD
Gender					
Female	46 (63.89)	11.61	7.18	12.02	6.73
Male	23 (31.94)	6.74	4.85	7.26	4.44
Other	2 (2.78)	13.50	14.85	12.00	14.14
Decline[a]	1 (1.39)	13.00	0.00	9.00	0.00
Age					
18	5 (6.94)	14.60	7.06	13.00	8.34
19	14 (19.44)	11.57	6.65	13.36	6.31
20	20 (27.78)	10.35	7.24	10.65	6.03
21	22 (30.56)	7.55	6.04	8.05	6.40
22	2 (2.78)	6.50	9.19	5.00	5.66
23-29[b]	3 (4.17)	9.00	9.00	7.00	4.36
30 and over[b]	6 (8.33)	13.50	7.71	13.33	6.15
Highest Education Received					
High School	5 (6.94)	15.60	6.02	14.60	8.44
Undergrad	61 (84.72)	9.62	6.74	10.13	6.32
Graduate	6 (8.33)	10.67	9.16	10.33	7.15

a. Due to the limited sample size, $N = 3$, "Decline" was grouped with "Other" during our analysis.

b. Based on the limited sample size for some age groups, these ages are reported as intervals.

B. Procedure

Participants were introduced to the research study through a Qualtrics survey, with the first page acting as an opt-in informed consent. Each participant was asked to take the PHQ-9 online embedded in the Qualtrics survey as well as the chatbot version of the screening, in a randomized order handled by the survey tool. Participants were given the option to experience Marcus either on an iOS device (provided with the link to download the app from the App Store) or through the web application (provided with a link to the web interface on Kommunicate). The multiple-choice PHQ-9 score was recorded automatically from the participants' responses on Qualtrics, while they had to manually enter the score outputted by Marcus to Qualtrics for the chatbot version. Participants were also asked demographic questions at the end of the survey, such as their age, gender, education level, location, and employment status. The participants' preference of screening method was also recorded, both regarding the perceived comfort, honesty and accuracy of interaction with the tools (including a neutral/no preference option), as well as an open-ended text to justify the reasons for their preference.

V. RESULTS

The results presented in this section aim at both addressing our key research questions and uncovering any extra insights that will inform our iteration of the chatbot. A variety of statistical tools were used to conduct analysis on the results, such as descriptive statistics, T-tests, analysis of variance (ANOVA), and correlation coefficients. Whenever normality is not reported, the distribution was found to be normal based on a histogram analysis. We started by checking the internal consistency of the online PHQ-9 responses, *Cronbach's a* = 0.904; we had no question-level records from Marcus, as the tool simply output the total score.

A. Accuracy of Chatbot for Measuring Depression

A two-tailed paired samples T-test was conducted on data from the 81 participants who completed screenings through both Marcus and the PHQ-9. The results, $t(80) = -0.971$, $p = 0.355$, found that there is no significant difference between the two screening scores, with Marcus average score ($M = 10.05$, $SD = 7.17$) being very similar compared to the online PHQ-9 average score ($M = 10.44$, $SD = 6.74$). The paired samples correlations indicated a highly significant correlation between the tools ($r = 0.863$, $p < 0.001$). The number of depression cases marked by Marcus and the PHQ-9 based on score classification varied slightly. According to Marcus, 11.11% ($N = 9$) of participants were identified as having a severe risk of suffering from a depression-related disorder, while data from the PHQ-9 questionnaire indicated that 12.35% ($N = 10$) of participants were at a severe risk. Marcus additionally found that 34.57% ($N = 28$) of participants classified as having moderate or moderately severe risk of depression compared to 38.27% ($N = 31$) for the PHQ-9. Complete score classification results between the screening tools for participants reporting their gender ($N = 72$) is shown in Figure 3.

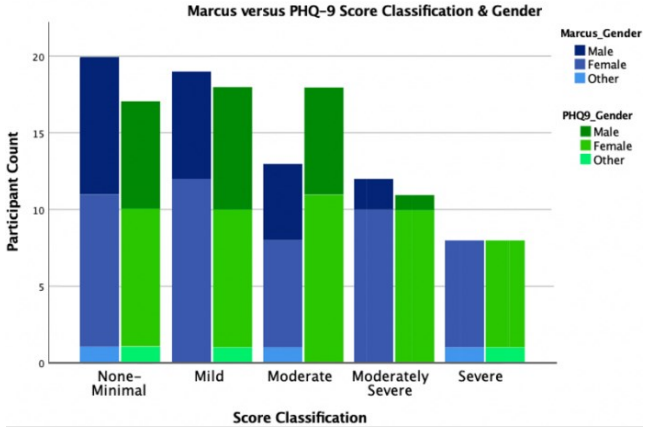

Figure 3. Marcus vs online PHQ-9 score classification by gender ($N = 72$).

Descriptive statistics were calculated for male and female participants that successfully completed the Marcus screening and the PHQ-9 assessment (see Table I). The average score across the two tools indicated a large discrepancy between the screening score of male ($M = 7.00$) and female ($M = 11.82$) participants. Investigating the significance of this relationship we ran a one-way ANOVA with bootstrapping (resampling the dataset across 1000 simulated samples with a 95% confidence intervals) due to violating the assumption of homogeneity of variance as shown by a Levene's test $F(1,67) = 5.696$, $p = 0.020$. Participants who reported their gender as "Other" or declined to respond were excluded due to the very small sample size ($N = 3$). Because of no significant difference found between Marcus and PHQ-9 scores based on the t-Test, the average of the two scores for each participant was used as the dependent variable in the ANOVA. We found a statistically significant difference between the depression screening scores, $F(1,67) = 9.904$, $p = 0.002$, with male participants having a significantly lower score as compared to females; the effect size was fairly large *Hedges' g* = 0.72 (preferred over Cohen's *d* due to unequal variance).

TABLE II. TOOL ORDER, INTIATION, COMPLETIION, AND PLATFORM

Tool order	N (%)	Initiated[a] (%)	Completed (%)	iOS Application (%)	Kommunicate (%)
Presented First					
Marcus	71 (54.62)	71 (100.00)	49 (69.01)	31 (63.27)	18 (36.73)
PHQ-9	59 (45.38)	59 (100.00)	49 (83.05)		
Set to Appear Second					
Marcus	59 (45.38)	49[b] (83.05)	32 (65.31)	22 (68.75)	10 (31.25)
PHQ-9	71 (54.62)	49[c] (69.01)	49 (100)		
Overall					
Marcus	130 (100.00)	120[b] (92.31)	81 (67.50)	53 (65.43)	28 (34.57)
PHQ-9	130 (100.00)	108[c] (83.08)	98 (90.74)		

a. The times a participant reached the survey page to download the app or follow the website link.

b. Ten (10) participants never reached the Marcus access screen and therefore were not counted.

c. Twenty-two (22) participants never reached the PHQ-9 questions and therefore were not counted.

Even though the study had a wide range of participants' age, the majority of them were between 18 and 21 years old (since undergraduates were mainly recruited). Once more, after averaging the scores across the two tools, we noticed a large variance between the screening score of younger undgraduates of ages 18 ($M = 13.80$) and 19 ($M = 12.46$) compared to older undergraduate students of ages 20 ($M = 10.50$) and 21 ($M = 7.80$). Examining further the significance of this relationship, we ran a one-way ANOVA with four levels, one for every one of the typical ages for U.S. college students; a Levene's test $F(3,57) = 0.199$, $p = 0.897$, showed that the homogeneity of variance assumption was met. A non-statistically significant difference between depression scores among the four college-age groups was discovered, $F(3,57) = 2.257$, $p = 0.092$; the above mean differences constituted a small effect size $\eta^2 = 0.106$.

B. *Effectiveness of Chatbot Compared to Online PHQ-9*

We defined effectiveness based on the initiation and completion rates of the screening sessions per tool. Initiation was originally defined as reaching the survey page with the link to download the mobile app or visit the screening tool website. Completion was based on recording a PHQ-9 score (calculated by Qualtrics) or a chatbot score (entered by the participant). The achieved response rate enabled obtaining complete screening data from 62.31% (81/130) of the overall sample. An additional 17 participants did only the PHQ-9 assessment and not the chatbot screening, while 32 participants completed neither. In addition, the order of tool presentation was also taken into account to investigate the effect of the randomized approach—regarding which tool was presented first—on completion rate. The majority of participants were presented with Marcus first ($N = 71$, 54.62%), while only 59 participants started with the online PHQ-9 (45.38%). The full data on completion rate, initiation rate, order of tool presentation, and chatbot platform are presented in Table II.

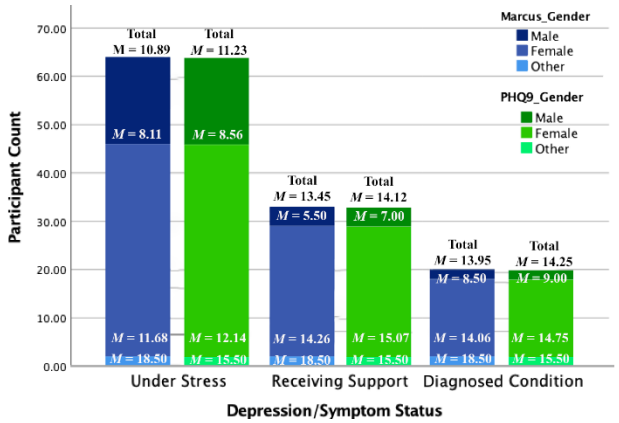

Figure 4. Marcus vs online PHQ-9 score by gender and stress factors; "Under Stress" includes only participants who somewhat/strongly agreed in that question; participant count does not add to the total of $N = 72$.

Tool effectiveness additionally considered a mean score comparison between Marcus and the PHQ-9 for the 72 participants who provided a response regarding if they were under considerable stress, had received recent mental health support, or were diagnosed with a mental or psychological condition over the last year. A majority replied that they "Strongly" or "Somewhat" agreed to being under considerable stress ($N = 64$, 88.89%); almost half reported having received mental health support ($N = 33$, 45.83%); while the majority were not diagnosed with a mental health condition ($N = 50$, 69.44%). Mean comparisons based on tool and gender for the self-reported stress assessment and their associated PHQ-9 scores are shown in Figure 4.

C. *Perceived Preference of Screening Tool*

A total of 72 participants reported an overall preferred tool for screening. A majority preferred the online PHQ-9 ($N = 44$, 61.11%) with the remainder of participants being either neutral ($N = 15$, 20.83%) or preferring Marcus ($N = 13$, 18.06%).

In addition to submitting their holistic tool preference, participants were also requested to report their screening tool preferences in terms of comfort, honesty, and accuracy. The results from the 70 participants who rated Marcus and the PHQ-9 on these three factors are shown in Figure 5.

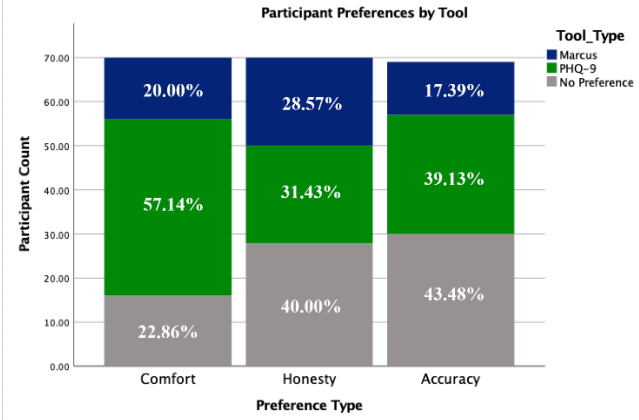

Figure 5. User preference between Marcus vs online PHQ-9 across three factors for $N = 70$ ("Accuracy" was completed by one less participant).

Participant comments regarding overall tool preference resulted in various common sentiments across individuals who partook in the study. Those who preferred the PHQ-9 assessment made note of the questionnaire being a much faster screening method with the multiple-choice option allowing for straightforward responses, which participants felt were being correctly correlated to a numeric score. Participants who preferred the Marcus screening noted their preference for the open-ended response system and that the chatbot "felt more natural." Concerns with Marcus included a lack of clarity on acceptable responses and the accuracy of the scoring methodology from the chatbot, while some also stated that the tool lacked realism and diversity. Many indicated a neutral stance that the tools were similar.

Regarding user comfort, results were generally mixed with some participants reporting that Marcus felt "more personable" and made them feel "more comfortable because it was like talking to a human." However, others who found the PHQ-9 to be a more comfortable tool noted its easier process and not feeling comfortable talking to a person, which Marcus simulates. Regarding the ability to provide honest answers, Marcus was stated to have "more flexibility" regarding responses and provided the ability to convey more information. Users additionally noted the positives in the flexibility for answers to be more vague or specific with the chatbot, as opposed to the preset responses with the PHQ-9.

VI. Discussion

Here we discuss our findings for addressing our research questions, also comparing with results from similar studies and presenting opportunities for improving our chatbot.

A. Psychometric Properties [RQ1]

Marcus' output score correlated significantly with the output score by the control PHQ-9, as indicated by the paired samples T-test. Additionally, the overall classification rates of the two screening methods were correlated, even though Marcus had the tendency to underscore participant responses (see Figure 3). The correlation between mean scores and classification rates between Marcus and the PHQ-9 control survey indicates that Marcus is a relatively accurate method for depression screening for a college population, performing comparably to similar chatbots [12][13]. Marcus' overall lower classification scores and slightly higher standard deviation indicates that Marcus' BERT model was not 100% accurate in translating user responses to entity scores. However, literature on PHQ-9 indicates that the instrument appears to expectedly have increased specificity and declining sensitivity in the middle part of the scale—between mild and moderate depression [38]—the classification levels where Marcus seemed to be mostly misaligned with the online PHQ-9.

Furthermore, gender seemed to be a predictor of depression screening score for both Marcus and the online PHQ-9, with female participants scoring higher, i.e., classified as having "severe" or "moderately severe" depression (Figure 3), compared to male participants—a significant finding based on an ANOVA. Our findings are in line with prior research, which has shown that females are more susceptible to depression than males [39], especially in a college setting [23]. Moreover, our score differences based on age, even if non-significant, indicate a tendency of younger adolescents to express higher stress and depression symptoms, as found by other studies [25]. It is possible that a larger sample might have been able to statistically confirm this inclination.

B. Completion Rate and Preference [RQ2]

Our measure of effectiveness compared the initiation and completion rates of using the chatbot versus the online PHQ-9 (Table II). The analysis of survey data showed that Marcus had an overall higher initiation rate (92.31%) than PHQ-9 (83.08%), but much lower completion rate (67.50% vs 90.74%). Considering participants had to follow an external link to either visit the online Kommunicate interface or download the iOS app to interact with Marcus, we speculated that the recorded rates—based on participants simply reaching the survey page with the link—were not accurate. A follow-up analysis of the Kommunicate logs allowed us to identify 38 participants who had no timestamped interactions with the chatbot within a 3-hour window following the survey initiation. This means that they either did not open the web-based chatbot link or did not download the app, depending on their selected platform. Only one participant was identified with a matching conversation but dropped halfway through the conversation. Based on this added analysis, the corrected initiation rate was found to be lower at 63.08% (82/130), but the corrected completion rate was much higher at 98.76% (81/82). The corrected rates are similar to findings by other studies [12]–[14], while the drop-out rate can be explained by the overhead needed to visit an external page, which some participants probably found detrimental to their participation.

Examining the correlation of participants' self-reported level of stress or having received support or even having been diagnosed with depression, with their PHQ-9 scores from both tools (Figure 4), we can clearly see that both tools were excellent in their assessment of depression-like symptoms. This is very similar to the association of depression references in social media with high scores of assessed depression using the PHQ-9 found by [22]. In terms of user preference, our findings were mixed despite most participants expressing a preference for the online PHQ-9 in most factors (Figure 5). We note that the online PHQ-9 was perceived as more comfortable and accurate due to the ease of use and speed of operation of completing a multiple-choice assessment, as opposed to Marcus, which was perceived equally honest, probably due to its anthropomorphism [32].

Nonetheless, it is also interesting to note that a couple of participants felt uncomfortable with the human representation of Marcus; one expressed a general discomfort with virtual chatbots, while another one felt unease due to Marcus represented as a white male. We recognize a tradeoff between the potential increased social presence and emotional affinity versus discomfort afforded by a human-like chatbot. Some comments confirmed findings from other studies about the increased freedom offered by a chatbot as a screening tool [12] [14], while a few participants complained about the chatbot not understanding their free-text responses. Similar challenges have been reported by other researchers [33][40], with better training data needed for optimizing chatbot performance.

 ISBN: 978-1-68558-078-0

C. Limitations

Despite comparing our chatbot with the online PHQ-9, we do not claim that either of the two tools were able to accurately detect depression; therefore, by "accuracy" we actually examined how close our tool came to the PHQ-9 as the standard in clinical practice. A higher sample size would have potentially increased the power of our findings, especially in terms of examining the differences in depression scores between college year of study. Similarly, a more diverse demographic in terms of geographic location, gender, race, and ethnicity would have allowed us to draw more generalizable results. Further analysis in terms of the number of questions presented to participants and natural language prompts recognized by the chatbot, could also serve to better identify the factors affecting completion rate and user preference; however, due to IRB restrictions we could not assign a unique ID to each participant to match survey data with chatbot interactions. Finally, the chatbot's limited training dataset can partly explain why multiple prompts by users could not be interpreted to intents by the BERT model, compromising accuracy (misclassification of some responses) and comfort (not understanding some answers). Some users reported the latter was a significant negative factor in their experience using Marcus.

VII. Conclusions and Future Work

The present work includes findings from a study with 130 participants in the U.S., comparing accuracy and effectiveness of two ways of administering the PHQ-9 depression screening instrument. Our chatbot, Marcus, was found comparable to the online PHQ-9 in terms of scores but the slight tendency to underscore participant responses produced a lower classification in some cases as compared to the traditional instrument. User comments and rating of the two tools indicated a preference for the online PHQ-9 in all factors but honesty, where Marcus was equally preferred. This leaves room for improving Marcus in terms of comfort, which is an important aspect of any health assessment interaction for the results to be accurate [41]. Overall, Marcus was found to be an effective first step for accessible depression screening, especially during the challenging times of a pandemic when college students were affected the most and were struggling to access emotional support.

Future work is focused towards three directions: a) improving classification accuracy of Marcus, b) accessing a wider, more diverse demographic of college students, and c) customizing the chatbot to be more inclusive for different populations. Regarding the first goal, additional supervised learning should be performed on the Marcus' BERT model along with more refined configuration values (e.g., entity scores based on identified intents). A more diverse population will allow us to run a between-subjects experiment—a better approach for such a sensitive health practice as depression screening—and gain insights about the use and perception of the chatbot based on factors like socioeconomic status, education, race, gender, etc. In line with this goal, we plan to test how customization of Marcus' representation (name, image) might affect accessibility by different demographics and increase comfort level, similar to how a research-based mobile app like Healthy Minds includes different meditation guides to accommodate multiple user preferences [42]. We anticipate such a follow-up study will provide insights to health and HCI researchers working in the domain of creating inclusive technologies for medical applications.

Acknowledgments

We would like to thank Nicole Ruzek, director of the university's Counseling & Psychological Services, and Bethany Teachman, professor at the department of Psychology and director of the Program for Anxiety, Cognition and Treatment (PACT) Lab, for their guidance and advice in the experimental design of this work. We also want to acknowledge the significant support by Raul Arrabales for providing the data from the Perla chatbot implementation.

References

[1] Institute of Health Metrics and Evaluation, "Global Health Data Exchange (GHDx)." http://ghdx.healthdata.org/gbd-results-tool?params=gbd-api-2019-permalink/d780dffbe8a381b25e1416884959e88b (accessed Mar. 18, 2023).

[2] V. Lorant, D. Deliège, W. Eaton, A. Robert, P. Philippot, and M. Ansseau, "Socioeconomic inequalities in depression: A meta-analysis," *Am J Epidemiol*, vol. 157, no. 2, 2003, doi: 10.1093/aje/kwf182.

[3] R. Shao *et al.*, "Prevalence of depression and anxiety and correlations between depression, anxiety, family functioning, social support and coping styles among Chinese medical students," *BMC Psychol*, vol. 8, no. 1, pp. 1–19, Dec. 2020, doi: 10.1186/s40359-020-00402-8.

[4] E. Sheldon *et al.*, "Prevalence and risk factors for mental health problems in university undergraduate students: A systematic review with meta-analysis," *J Affect Disord*, vol. 287, pp. 282–292, 2021, doi: 10.1016/j.jad.2021.03.054.

[5] J. P. Lépine and M. Briley, "The increasing burden of depression," *Neuropsychiatr Dis Treat*, vol. 7, no. Suppl. 1, pp. 3–7, May 2011, doi: 10.2147/NDT.S19617.

[6] K. O. Conner *et al.*, "Mental health treatment seeking among older adults with Depression: The impact of stigma and race," *American Journal of Geriatric Psychiatry*, vol. 18, no. 6, pp. 531–543, 2010, doi: 10.1097/JGP.0b013e3181cc0366.

[7] L. H. Andrade *et al.*, "Barriers to mental health treatment: Results from the WHO World Mental Health surveys," *Psychol Med*, vol. 44, no. 6, pp. 1303–1317, 2014, doi: 10.1017/S0033291713001943.

[8] M. Neathery, E. J. Taylor, and Z. He, "Perceived barriers to providing spiritual care among psychiatric mental health nurses," *Arch Psychiatr Nurs*, vol. 34, no. 6, pp. 572–579, 2020, doi: 10.1016/j.apnu.2020.10.004.

[9] K. Daley, I. Hungerbuehler, K. Cavanagh, H. G. Claro, P. A. Swinton, and M. Kapps, "Preliminary Evaluation of the Engagement and Effectiveness of a Mental Health Chatbot," *Front Digit Health*, vol. 2, pp. 1–7, Nov. 2020, doi: 10.3389/fdgth.2020.576361.

[10] R. Crasto, L. Dias, D. Miranda, and D. Kayande, "CareBot: A mental health chatbot," in *2nd International Conference for Emerging Technology (INCET 2021)*, 2021, pp. 1–5. doi: 10.1109/INCET51464.2021.9456326.

[11] B. Arroll *et al.*, "Validation of PHQ-2 and PHQ-9 to Screen for Major Depression in the Primary Care Population," *The Annals of Family Medicine*, vol. 8, no. 4, pp. 348–353, Jul. 2010, doi: 10.1370/afm.1139.

[12] R. Arrabales, "Perla: A Conversational Agent for Depression Screening in Digital Ecosystems. Design, Implementation and Validation," *arXiv preprint*, Aug. 2020, [Online]. Available: https://arxiv.org/abs/2008.12875 (accessed Mar. 23, 2023).

 ISBN: 978-1-68558-078-0

[13] G. Dosovitsky, E. Kim, and E. L. Bunge, "Psychometric Properties of a Chatbot Version of the PHQ-9 With Adults and Older Adults," *Front Digit Health*, vol. 3, pp. 1–8, Apr. 2021, doi: 10.3389/fdgth.2021.645805.

[14] I. Hungerbuehler, K. Daley, K. Cavanagh, H. G. Claro, and M. Kapps, "Chatbot-based assessment of employees' mental health: Design process and pilot implementation," *JMIR Form Res*, vol. 5, no. 4, pp. 1–11, Apr. 2021, doi: 10.2196/21678.

[15] M. C. Klos, M. Escoredo, A. Joerin, V. N. Lemos, M. Rauws, and E. L. Bunge, "Artificial intelligence⇓based chatbot for anxiety and depression in university students: Pilot randomized controlled trial," *JMIR Form Res*, vol. 5, no. 8, pp. 1–9, Aug. 2021, doi: 10.2196/20678.

[16] S. Suharwardy *et al.*, "116: Effect of an automated conversational agent on postpartum mental health: A randomized, controlled trial," *Am J Obstet Gynecol*, vol. 222, no. 1, p. S91, 2020, doi: 10.1016/j.ajog.2019.11.132.

[17] J. Wosik *et al.*, "Telehealth transformation: COVID-19 and the rise of virtual care," *Journal of the American Medical Informatics Association*, vol. 27, no. 6, pp. 957–962, 2020, doi: 10.1093/jamia/ocaa067.

[18] A. R. Dennis, A. Kim, M. Rahimi, and S. Ayabakan, "User reactions to COVID-19 screening chatbots from reputable providers," *Journal of the American Medical Informatics Association*, vol. 27, no. 11, pp. 1727–1731, 2020, doi: 10.1093/jamia/ocaa167.

[19] K. Kroenke and R. L. Spitzer, "The PHQ-9: A new depression diagnostic and severity measure," *Psychiatr Ann*, vol. 32, no. 9, pp. 509–515, 2002, doi: 10.3928/0048-5713-20020901-06.

[20] M. Ghazisaeedi, H. Mahmoodi, I. Arpaci, S. Mehrdar, and S. Barzegari, "Validity, Reliability, and Optimal Cut-off Scores of the WHO-5, PHQ-9, and PHQ-2 to Screen Depression Among University Students in Iran," *Int J Ment Health Addict*, vol. 20, no. 3, pp. 1824–1833, 2022, doi: 10.1007/s11469-021-00483-5.

[21] N. Fanaj and S. Mustafa, "Depression measured by PHQ-9 in Kosovo during the COVID-19 outbreak: an online survey," *Psychiatr Danub*, vol. 33, no. 1, pp. 95–100, Apr. 2021, doi: 10.24869/psyd.2021.95.

[22] M. A. Moreno *et al.*, "A Pilot Evaluation of Associations Between Displayed Depression References on Facebook and Self-reported Depression Using a Clinical Scale," *J Behav Health Serv Res*, vol. 39, no. 3, pp. 295–304, Jul. 2012, doi: 10.1007/s11414-011-9258-7.

[23] A. K. Boggiano and M. Barrett, "Gender differences in depression in college students," *Sex Roles*, vol. 25, no. 11–12, pp. 595–605, Dec. 1991, doi: 10.1007/BF00289566/METRICS.

[24] L. Zhao *et al.*, "Gender Differences in Depression: Evidence From Genetics," *Frontiers in Genetics*, vol. 11. Frontiers Media S.A., Oct. 15, 2020. doi: 10.3389/fgene.2020.562316.

[25] K. M. Kiely, B. Brady, and J. Byles, "Gender, mental health and ageing," *Maturitas*, vol. 129, pp. 76–84, Nov. 2019, doi: 10.1016/j.maturitas.2019.09.004.

[26] M. Altemus, N. Sarvaiya, and C. Neill Epperson, "Sex differences in anxiety and depression clinical perspectives," *Front Neuroendocrinol*, vol. 35, no. 3, pp. 320–330, Aug. 2014, doi: 10.1016/j.yfrne.2014.05.004.

[27] K. Kosyluk *et al.*, "Mental Distress, Label Avoidance, and Use of a Mental Health Chatbot: Results from a U.S. Survey," Nov. 2022. doi: 10.31124/ADVANCE.21431079.V1.

[28] V. Venkatesh, "Determinants of Perceived Ease of Use: Integrating Control, Intrinsic Motivation, and Emotion into the Technology Acceptance Model," *Information Systems Research*, vol. 11, no. 4, pp. 342–365, 2000, doi: 10.1287/isre.11.4.342.11872.

[29] H. Liu, H. Peng, X. Song, C. Xu, and M. Zhang, "Using AI chatbots to provide self-help depression interventions for university students: A randomized trial of effectiveness," *Internet Interv*, vol. 27, p. 100495, Mar. 2022, doi: 10.1016/j.invent.2022.100495.

[30] O. Romanovskyi, N. Pidbutska, and A. Knysh, "Elomia chatbot: The effectiveness of artificial intelligence in the fight for mental health," in *CEUR Workshop Proceedings of the 5th International Conference on Computational Linguistics and Intelligent Systems*, 2021, pp. 1–10.

[31] G. Giunti, M. Isomursu, E. Gabarron, and Y. Solad, "Designing Depression Screening Chatbots," in *Studies in Health Technology and Informatics*, vol. 284, pp. 259–263, Dec. 2021. doi: 10.3233/SHTI210719.

[32] D.-C. Toader *et al.*, "The Effect of Social Presence and Chatbot Errors on Trust," *Sustainability*, vol. 12, no. 1, p. 256, Dec. 2019, doi: 10.3390/su12010256.

[33] V. Dogra *et al.*, "A Complete Process of Text Classification System Using State-of-the-Art NLP Models," *Comput Intell Neurosci*, vol. 2022, pp. 1–26, Jun. 2022, doi: 10.1155/2022/1883698.

[34] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *arXiv preprint*, Oct. 2018, [Online]. Available: http://arxiv.org/abs/1810.04805 (accessed Mar. 24, 2023).

[35] Google Cloud Documentation, "DialogFlow ES." https://cloud.google.com/dialogflow/es/docs/training (accessed Feb. 24, 2023).

[36] "Kommunicate." https://www.kommunicate.io/ (accessed Mar. 09, 2023).

[37] Piper Sandler, "Taking Stock With Teens," 2021. [Online]. Available: https://www.pipersandler.com/teens (accessed: Feb. 28, 2023).

[38] K. Kroenke, R. L. Spitzer, and J. B. W. Williams, "The PHQ-9," *J Gen Intern Med*, vol. 16, no. 9, pp. 606–613, Sep. 2001, doi: 10.1046/j.1525-1497.2001.016009606.x.

[39] L. Zhao *et al.*, "Gender Differences in Depression: Evidence From Genetics," *Front Genet*, vol. 11, pp. 1–15, Oct. 2020, doi: 10.3389/fgene.2020.562316.

[40] C. R. Zgraggen, S. B. Kunz, and K. Denecke, "Crowdsourcing for creating a dataset for training a medication chatbot," in *Public Health and Informatics: Proceedings of MIE 2021*, IOS Press, 2021, pp. 1102–1103. doi: 10.3233/SHTI210364.

[41] C. Wensley, M. Botti, A. McKillop, and A. F. Merry, "Maximising comfort: how do patients describe the care that matters? A two-stage qualitative descriptive study to develop a quality improvement framework for comfort-related care in inpatient settings," *BMJ Open*, vol. 10, no. 5, p. e033336, May 2020, doi: 10.1136/bmjopen-2019-033336.

[42] S. B. Goldberg *et al.*, "Testing the Efficacy of a Multicomponent, Self-Guided, Smartphone-Based Meditation App: Three-Armed Randomized Controlled Trial," *JMIR Ment Health*, vol. 7, no. 11, p. e23825, Nov. 2020, doi: 10.2196/23825.

 ISBN: 978-1-68558-078-0

This page break denotes the conclusion of the published study "Marcus: A Chatbot for Depression Screening Based on the PHQ-9 Assessment."

The following page is an in-progress report of the follow-up study, which was commenced shortly after the completion of the first paper.

Follow-Up Study (In Progress): Development of a New Chatbot

Creating a new virtual agent to improve classification accuracy and provide customization

I. Objectives And Status

A. Improvng Classification Accuracy

With the follow-up study building upon the now-published work "Marcus: A Chatbot for Depression Screening Based on the PHQ-9 Assessment," we homed in on the limitations as points of improvement. Given the within-subjects design of the first study, we were able to determine the accuracy of the Marcus chatbot by comparing scores to the PHQ-9 via a Paired-Samples T Test. Results of the test demonstrated that the chatbot had an average score of 10.05 compared to 10.44 for the PHQ-9, indicating generally adequate accuracy for Marcus.

However, further analysis revealed that, while having a comparable average score overall, Marcus misclassified the severity of depression of multiple participants. As it pertains to the PHQ-9, the score intervals are as follows: 0-4 = None - Minimal Depression, 5-9 = Mild Depression, 10-14 = Moderate Depression, 15-19 = Moderately Severe Depression, and 20-27 = Severe Depression. With this being the case, the Marcus chatbot being off by one or two points in an instance resulted in a misclassified depression severity rating. This limitation was attributed to a few different factors, with the most significant ones being the need for more robust training data, as well as the limitations of Google DialogFlow.

In our follow-up study, our approach to improving this has been to migrate to a different technology to aid in the machine learning process – TensorFlow. With TensorFlow being a machine learning library for building and training models compared to DialogFlow, which is more of a platform for building conversational interfaces, there will be more opportunities for customization. In summary, TensorFlow allows for the building of a custom model tailored for specific needs (depression screening), while DialogFlow provides pre-built conversational components that can be customized.

Moreover, the utilization of TensorFlow will bolster accuracy due to the machine learning models built with the technology having been shown to be highly accurate for a wide range of tasks. By comparison, the natural language processing (NLP) capabilities of DialogFlow, while relatively accurate, can be much more precise. Lastly, with TensorFlow, there is a greater degree of flexibility with full control over the model architecture and training process. This means that there is more opportunity to experiment with different models and parameters.

B. Accessing a more diverse demographic of students

In the first research study, various demographic variables of participants were collected. This included age, gender, and education level – each of which (particularly gender) provided insights as it related to depression prevalence and chatbot effectiveness. Through the initial literature review and domain research, it was noted that socioeconomic status and race were factors in predicting depression rates. As these demographic variables were not collected in the pilot study, this is a point of improvement in our follow-up. Additionally, more robust research into the relationships between socioeconomic status and race with depression rates and depression severity has been conducted in the second study.

As considering these demographics will potentially provide more insight into the use of a virtual agent to screen for depression, we have the objective of accessing a wider, more diverse demographic of college students. Currently, this process is being carried out by reaching out to various student organizations at the University. In addition, interviews are being conducted with students to better understand different user stories to aid in the development of the new chatbot. The new chatbot implementation and its importance as it pertains to recruiting a more diverse sample are discussed in the succeeding section.

C. Developing a Customizable/Inclusive Virtual Agent

Another limitation discovered during the review process of feedback provided by participants was discomfort with the human representation of the Marcus chatbot. One participant expressed general discomfort with virtual chatbots, while another felt unease due to Marcus being represented as a white male. Recognizing this, our follow-up study aims to rectify the limitation of discomfort afforded by a human-like chatbot by seeking a customizability option on the frontend. As it currently stands, there are various options available to fulfill this which will most likely result in utilizing a different platform for the frontend of the new chatbot, different from the Kommunicate API that was used for Marcus.

Regarding customization of the chatbot, this will include altering the image (race and gender) as well as the name. While the exact details are contingent on the ongoing contextual inquiry, this will be a necessary improvement as we seek to access a more diverse sample of participants for the follow-up study. Ultimately, this implementation will aid in the creation of an inclusive technology for medical application.