Acceptability and Efficacy of a Mental Health Chatbot:

A Randomized Controlled Trial for College Students

Camellia Bui, Emily Albert, and Luke Finkelstein

PSYC 4462: Research Experience in Abnormal Psychology

Dr. Melissa G. Hunt

I: Introduction

Mental Health Crisis on College Campuses

Mental health issues are commonly observed within college student populations. Not only do many students arrive at college with mental health conditions, but mental health crises can arise from the stress-inducing nature of college life (Sapadin & Hollander, 2022). College introduces new kinds of stressors to students or exacerbates existing ones, such as homesickness, financial concerns, balancing work, academic pressure, and family responsibilities (Pedrelli et al., 2015).

An increasing number of students have been reporting mental health disorders in recent years, according to several nation-wide surveys (American College Health Association, 2015, 2019). The American College Health Association found that nationally 14% of students report having one mental health condition, 23% report having both depression and anxiety, and 7.8% report having two or more other mental health conditions (American College Health Association, 2024). The severity of these conditions, including suicidal ideation, been increasing (Prince, 2015; Lipson, Lattie, & Eisenberg, 2019). In fact, suicide is now the second leading cause of death for 15- to 29-year-olds (Gomes de Andrade et al., 2018). The current state of college student mental health is frequently labeled a “crisis” because college counseling services are struggling to meet the increasing demand (Xiao et al., 2017; Dekker et al., 2020; Sapadin & Hollander, 2024).

This crisis has also deteriorated further due to the effects of the COVID-19 pandemic (Halliburton et al., 2021) and are predicted to have lasting consequences for mental healthcare in years moving forward (Copeland et al., 2021). These mental health challenges can significantly impact various aspects of students’ lives in addition to health and well-being, such as academic

success (Eisenberg et al., 2009). There is thus a clear cause for concern regarding the mental health of students in college.

With this increased demand for mental health services on campuses, universities are struggling to provide sufficient counseling and support to students (Prince, 2015; Watkins et al., 2011; Oswalt, et al. 2018). The 2014 National Survey of College Counseling Centers noted that the average ratio of counselors to clients was 1 to 2081, indicating a significant shortage of counselors available to meet the growing demand (Gallagher & Taylor, 2015). This highlights the increasingly common issue of understaffed counseling services across universities. Additionally, unmet mental health needs might be present in the ones who may be the most at risk. The highest mental health service use on college campuses are reported in students with active coping skills, i.e. those who are more proactive to adapting to stressors (Sontag-Padilla et al., 2016). Patients who use college counseling are also less likely to be in minority groups, e.g. Asian, Hispanics and international students (Gallagher, 2013). Alarmingly, most students who have committed suicide never made contact with their university counseling services (Gallagher, 2013). There are also many students who do not seek out or receive help, even when they are aware of the resources available to them (Oswalt et al., 2020; Zivin et al., 2009). This is the case even when there is free access to basic services (Eisenberg et al., 2007; Eisenberg et al., 2011). Some deterrents include stigma, scheduling conflicts with classes and other commitments, "travel time" to the counseling center, not knowing where the counseling center is, and long "waitlists" (Nguyen-Feng, Greer, & Frazier, 2017; Prince, 2015; Cohen, Graham, & Lattie, 2020).

Current Mental Health Services on Campuses

Many universities offer opportunities for students to receive help and support, such as through referrals, peer-to-peer counseling, or campus health services. However, the effectiveness of these efforts is unclear. One main issue is that there is currently no standardized intervention for college students that provides preventative care for mental health disorders. Existing programs vary widely in terms of scope, effectiveness, and accessibility (Buchanan, 2012; Patel & Lewis, 2023). Given the varying factors influencing students' mental health, a flexible and tailored approach is needed rather than a one-size-fits-all solution for clinicians.

Another issue is that, because of the increasing demand for counseling services, universities are having a hard time keeping up and are unable to provide sufficient resources (Dekker et al., 2020; Sapadin & Hollander, 2024). On top of the shortage, psychiatrists working with college students lack specific expertise in this demographic. Universities are in need of more psychiatrists who specialize in this age group and understand the stressors affecting college students (Pedrelli et al., 2015).

As an alternative, peer-to-peer counseling has evolved as a response to increasing demand for mental health services. While this offers a supportive environment for students seeking counseling, peers acting as clinicians are "largely untrained and may struggle to provide empathetic support" (O'Leary, 2023). Additionally, universities have started implementing other alternatives such as group therapy and telehealth (Abrams, 2022). A pilot study on college students showed similar benefits derived from group therapy as in individual therapy (Fawcett et al., 2019). While this seems promising, wide implementation may not be possible due to challenges associated with group therapy, such as social anxiety, the fear of conflict or being humiliated by other group members, or preference for individual attention (Shay, 2021). For

instance, a student feeling imposter syndrome might be hesitant to share their feelings with those they perceive as the ones they are "impostering." Other factors could also include concerns about confidentiality or reluctance due to stigma about mental health, internally or socially. These factors emphasize the importance of new innovative solutions to support students.

AI in Mental Health

Overview

There is an emerging set of literature on the rising implementation of Artificial Intelligence (AI) and its potential as a service for mental health support (Sweeney et al., 2021). Innovations, such as virtual reality, natural language processing and affective computing, have created a range of new healthcare tools, from algorithms that can detect risk symptoms from electronic medical records and social media data, to 'virtual psychotherapists' capable of interacting with live patients and supporting therapeutic interventions on par with highly-skilled health professionals (Luxton, 2014; Fiske et al., 2019; Linthicum et al., 2019). Its applications have focused on four specific areas within the literature: (1) detection and diagnosis, (2) prognosis, treatment, and support, (3) public health, and (4) research and clinical administration. (Shatte et al., 2019). However, this growth also introduces complexities, including ethical concerns, technical challenges, and the need for careful regulation to ensure safe and effective use (Fiske et al., 2019).

Many researchers in the fields of mental health and AI have started to explore the intersection of AI and mental health support. Specifically, OpenAI's ChatGPT has been shown to be promising as a mental health support service (Alanezi, 2024; Maurya et al., 2024). One study

had encouraging results, showing that ChatGPT "enhanced patient-reported quality of life in a psychiatric setting, with high user satisfaction" (Melo et al., 2024).

There has also been an increase in chatbots created specifically to act as an intervention for mental health (Fiske et al., 2019). These include chatbots such as Wysa, Woebot, Shim, Tess, Daylio, and ELIZA, which are starting to be researched in relation to their effectiveness and therapeutic impact (Melo et al., 2024; Eltahawy et al., 2023). However, these new studies have yet to specifically address the college student population. One study looked at young adults who experienced symptoms of depression during the COVID-19 pandemic and their use of a chatbot trained on Cognitive Behavioral Therapy (CBT), with the study concluding that it was "feasible and engaging" (He et al., 2022). However, most studies lack data on treatment efficacy, as their "assessments have focused on perceived empathy rather than clinical outcomes" (Torous & Blease, 2024).

There are other potential breakthroughs that AI could have in terms of reaching those who don't want to be recognized by society as "seeing a therapist." As previously mentioned, mental health stigma is pervasive on college campuses, particularly in elite college settings (Billings, 2020). Mental health chatbots allow students who perceive stigma on campus to access therapeutic interventions discreetly, minimizing the risk of stigmatization from peers and offering a more confidential alternative to in-person counseling. Previous studies involving digital mental health tools have been received positively by the student body and university staff populations. One preliminary study evaluating the integration of a web-based mental health self-screening and referral system at the University of Washington for 17 months found that the technology was effective and well-received across the board by student patients, where it was even transferred to the university clinic for daily use (Kim et al., 2011).

Concerns and Challenges with AI in Mental Health

Additionally, there are a lot of fears and uncertainties around AI regarding privacy, patient safety, competence, trust, liability, and biases (Luxton, 2014; Holm, 2024). Data privacy issues have become such a big concern that the American Psychiatric Association needed to "release an advisory in [the] summer [of] 2023 noting that clinicians should not enter any patient information into any AI chatbot" (Torous & Blease, 2024). Academic sources mention the need for these concerns to be addressed, as well as other concerns about overreliance on AI and clinical misuse (Chen et al., 2024).

Current Study

This randomized controlled trial aims to address the problems outlined above by examining the efficacy and acceptability of a mental health chatbot, called Elomia, in supporting college students' mental health. The primary research questions are:

1) What effect does Elomia, an AI-based mental health chatbot, have on depression, anxiety, and stress of Penn undergraduate students over a four-week period?
2) How acceptable do students find this chatbot?

For this study, we partner with a Ukraine-based company, which produces the Elomia chatbot and provides us with the access. Elomia is a generative AI program that responds to text input by users and is designed to address therapeutic targets. Elomia's algorithm was trained using real psychotherapy sessions from licensed therapists with expertise in CBT who responded to a wide range of individuals sharing their concerns.

To evaluate its effectiveness, the current study recruited undergraduate students from the University of Pennsylvania and randomly assigned them to two groups: an intervention (Elomia chatbot) and an active control. The active control group consists of many online wellness modules, including topics such as mindfulness, exercise, stress, and sleep that have been adapted from existing resources available to Penn students. This active control provides a comparison to Elomia as a digital intervention for mental health and well-being for undergraduate students.

Given the increasing prevalence of mental health issues among college students, in addition to the need for innovative and widely-implementable solutions, this study fills an important gap in the literature. The current research on AI and mental health has focused on the general population, but not enough on university students specifically, who face unique stressors.

We hypothesize that participants using the Elomia mental health chatbot will show improvements in depression, stress, and anxiety compared to the active control group. Additionally, we also predict that participants will rate the chatbot as more acceptable than the control group on some dimensions such as likability and willingness to recommend to others, using an acceptability survey we created for this study. Lastly, using a Modified Working Alliance Inventory, we predict that participants will rate the Elomia mental health chatbot on average more positively than negatively.

II: Methods

This study was approved by the Institutional Review Board at the University of Pennsylvania. Informed consent was obtained from all participants prior to their completion of the intake questionnaires. The trial was registered as a clinical trial and complied with good clinical practices.

Design

This study was a four-week randomized control trial using a between-subjects design. Participants were randomly assigned to one of two conditions: 1) an intervention using the Elomia chatbot, or 2) an active control condition consisting of virtual wellness modules adapted from publicly available online resources for University of Pennsylvania (Penn) students. Participants completed pre- and post-study measures in depression, stress, and anxiety and follow-up measures on acceptability and therapeutic alliance.

Participants

From January 2025 to April 2025, 120 participants were recruited from the Penn Sona Systems, a university-wide system that gives in-class credits to students in psychology classes if they participate in research studies. Only 83 participants completed the initial baseline survey and were enrolled in the study. This exceeds the original target sample size of 30 participants per condition, determined based on similar trials and by anticipating a small amount of attrition (Hunt et al., 2017). There were 13 drop-outs after the initial baseline surveys. Power calculations (based on BDI scores from Hunt et al., 2017) suggested that 29 participants per group is required

for 80% power and 5% Type 1 error. With 83 individuals enrolled and an attrition rate of about 15%, each condition finished with about 35 participants in each group.

Table 1: Participant Characteristics

Characteristic	**Total**	**Active Control**	**Intervention**
Age, *median* (IQR)	20 (19-21)	20 (19-21)	91 (19-21)
Gender, *N* (%)			
Male	18 (21.7)	9 (22.5)	9 (20.9)
Female	62 (74.7)	28 (70.0)	34 (79.1)
Non-Binary	2 (2.4)	2 (5.0)	0 (0.0)
Prefer not to disclose	1 (1.2)	1 (2.5)	0 (0.0)
Race, *N* (%)			
South Asian	5 (6.0)	3 (7.5)	2 (4.7)
East Asian	20 (24.1)	7 (17.5)	13 (30.2)
Latinx/Hispanic	6 (7.2)	2 (5.0)	4 (9.3)
White	22 (26.5)	14 (35.0)	8 (18.6)
African American	3 (3.6)	0 (0.0)	3 (7.0)
Black/African	2 (2.4)	1 (2.5)	1 (2.3)
Middle Eastern/North African	6 (7.2)	3 (7.5)	3 (7.0)
Mixed	18 (21.7)	9 (22.5)	9 (20.9)
Prefer not to disclose	1 (1.2)	1 (2.5)	0 (0.0)

Inclusion/Exclusion Criteria

Eligibility for the study consisted of participants reporting that they were over the age of 18 and undergraduate students at the University of Pennsylvania. Active standing at the university was necessary at the time of enrollment. During the screening and consent phase, participants were excluded from the study if they scored over 30 on the Beck Depression Inventory (BDI), which is classified as severe depression, or endorsed active suicidal ideation on the BDI by scoring either two or three on Item 9 of that questionnaire. Based on this criteria, one participant was excluded from the study, but were still given access to the resources.

Throughout the study, participants were monitored for risks. This study's protocol specified that if there was a sign that participants were experiencing severe depression or suicidal ideation as indicated on the BDI or Elomia sessions, participants were redirected to the principal investigator, Dr. Melissa Hunt, a licensed clinical psychologist, who would reach out to them directly to evaluate their safety and determine whether they could continue the study and/or should be referred to a higher level of care.

Intervention and Control Conditions

Intervention

Elomia is a mental health chatbot that has been trained on psychotherapy sessions with licensed psychologists, developed by the Ukraine-based company Elomia—the technology partner in this study. Elomia is accessible 24/7 via a website through any device, similar to ChatGPT, once they create an account on the website. Users can chat with Elomia via typing or texting, not through audio records, calling, or facetime. Available at any time, Elomia provides an opportunity for individuals to receive therapeutic guidance or support on demand, safely and

anonymously. Elomia can provide a listening space for participants to feel heard and suggest several different evidence-based therapeutic strategies to help users in processing negative feelings, analyzing problems and situations, coming up with solutions, and addressing potential obstacles to implementing those strategies. The algorithm assesses the user's emotional needs during interactions with the chatbot and recommends an exercise to help improve their emotional state. It is not a substitute for professional care, and will detect and send a check-in survey when a person needs additional support (e.g., suicide risks), then provides hotline and additional resources. In the case of this study, as described in the section above, such a detection would trigger the principal investigator. More information about Elomia can be found via this website: https://elomia.com/.

Active Control

The active control website was created by developers by the study's technology partner based on the researchers' specifications and materials. The website content was built by sampling informational resources that are publicly available to Penn students, such as slides from the official Penn Wellness webpage, sleep advice from the Penn Sleep Medicine researchers, and workshop slides from the Netter Center for Community Partnerships. The website consisted of seven topic categories: Academic Support, Diet/Nutrition, Exercise/Fitness, Sleep, Stress Management, Relationships/Sexuality, and General/Miscellaneous. Each category had at least an hour of content. The website layout, interface, and log-in process was designed to be similar to the intervention group to ensure a comparable experience between the conditions.

Measures

Beck Depression Inventory II (BDI)

The BDI is a 21-item self-report questionnaire used to assess current depressive symptom severity. It has been found to have high internal consistency ($\alpha = 0.91$), and high test–retest reliability *(*$r = 0.93$) (Beck, Steer, Ball, & Ranieri, 1996).

General Anxiety Disorder (GAD-7)

The GAD-7 is a 7-item self-reported questionnaire used to measure the severity of generalized anxiety. It has been found to have high internal consistency ($\alpha = .92$) and good test-retest reliability ($r = 0.83$) (Spitzer et al., 2006).

Perceived Stress Scale (PSS-10)

The PSS-10 is a 10-item self-reported questionnaire used to assess stress levels in young people and adults aged 12 and above, evaluating the extent an individual perceives their life as unpredictable, uncontrollable, and overloading over the previous month (Cohen et al., 1983). It has been found to have good internal consistency and test-retest reliability for American adolescents and university student populations (Lee, 2012; Kechter et al. 2019). PSS-10 was also found to be positively correlated with anxiety and depression symptoms in university students (Lee, 2012).

Modified Working Alliance Inventory Short Revised (MWAI-SR)

An adapted version of a widely used 12-item self-report questionnaire called the Brief Revised Working Alliance Inventory was used to assess therapeutic alliance in psychotherapy

(Benthem et al., 2024). The language of this inventory was modified to be relevant for an AI mental health chatbot. See Appendix A for the MWAI-SR.

Acceptability Survey (AS)

Since there was no gold-standard AS for mental health chatbots or digital mental health, a new scale was developed. The AS for the intervention is a 16-item self-report questionnaire, consisting of 14 multiple choice and two free response questions. The AS for the control is a 9-item self-report questionnaire, consisting of seven multiple choice and two free response questions. These questions aimed to measure the extent to which participants liked and found the chatbot as helpful for mental healthcare. Some examples are "I liked using Elomia/research website" and "I would recommend Elomia/the research website to friends and loved ones." See Appendix B and C for the full survey of each condition.

User Engagement

Frequency and length of session use for each participant were tracked through website traffic analytics on both the intervention and control websites. Additionally, session summaries with the chatbot, ranging from two-five sentences, were generated by the Elomia generative AI algorithm provided by the company. These summaries included the main topics of each session, what type of support or response the chatbot offered, and occasionally the participants' type of response to the chat's interaction. For the control group, a list of resources participants had accessed in each session were generated.

Procedure

Once participants consented and enrolled in the study, they were randomly assigned to one of the two conditions using random.org's coin flipper, with heads corresponding to the intervention and tails corresponding to the control. Participants then received a confirmation email informing them to which condition they were assigned, a link to their condition's website, a unique User ID (e.g. "laughingllama," "calmcloud") to log-in to their website, and full instructions for study participation. These instructions informed them to create a new account on the website with their assigned User ID and generate their own password. The only identifying information collected on the website that were connected to the participants was their User ID.

Before starting the four-week testing, participants from both conditions were asked to complete baseline measures of distress, including the BDI, GAD-7, and PSS-10. Participants in both groups were required to use their assigned website at least 30 minutes per week, but they were encouraged to use it as needed, with recommendations of at least twice per week for 30 minutes. During the study, biweekly reminders were sent to complete the 30 minutes per week study requirement. At the end of four weeks, participants completed follow-up questionnaires, including the BDI, GAD-7, the PSS-10, and relevant versions of the AS. The intervention group also completed the MWAI-SR. Those who completed the intervention were sent an email after completing the follow up survey with the opportunity to participate in a 30-minute exit interview via Zoom, with a $15 Amazon gift card as payment. Both groups were given access to the chatbot and the wellness modules after the study ended, and those who completed the time requirement were given credits through Sona Systems. For both the intervention and control conditions, website traffic analytics – including time, frequency, and content – were processed weekly to capture user engagement.

Data Analysis

Quantitative

Statistical analyses were conducted using IBM SPSS Statistics version 29 and R version 4.4.3. Analyses included descriptive statistics, independent samples *t*-tests to examine between-group differences for BDI, GAD-7, PSS-10, and AS measures, paired samples *t*-tests to examine within-group changes over four weeks for BDI, GAD-7 and PSS-10, and ANCOVAs to examine between-group differences, controlling for baseline, demographic variables, time use and frequency. Since the vast majority of participants did not complete the prescribed protocol of 30 minutes per week for four weeks, an intention-to-treat (ITT) analysis approach is used, which includes all participants in the analysis regardless of their adherence to the protocol. For this reason, time use is controlled in all analyses. The primary outcomes of interest were within-group differences for each condition after 4-week of testing and between-group differences at follow-up on efficacy measure as well as on individual items on the AS.

Qualitative

Qualitative coding on common stressors expressed was performed on all 198 session summaries. To develop the codebook, 30 session summaries were randomly pulled from all available sessions midway through the study and used as training data. Three student raters independently coded the data. Two forms of interrater reliability were calculated: Cohen's kappa for common stressors and weighted kappa for types of interaction. Here are two examples of session summaries received from Elomia during the study:

Session Summary Example 1

"The user expressed difficulty focusing on school work and sought ways to improve productivity. They described their current study habit of rewriting notes, which was not effective, leading to frustration. The chatbot suggested the Pomodoro Technique and creating a distraction-free study environment as strategies to enhance focus and manage frustration. The user appreciated the advice and concluded the session feeling satisfied with the guidance provided."

Session Summary Example 2

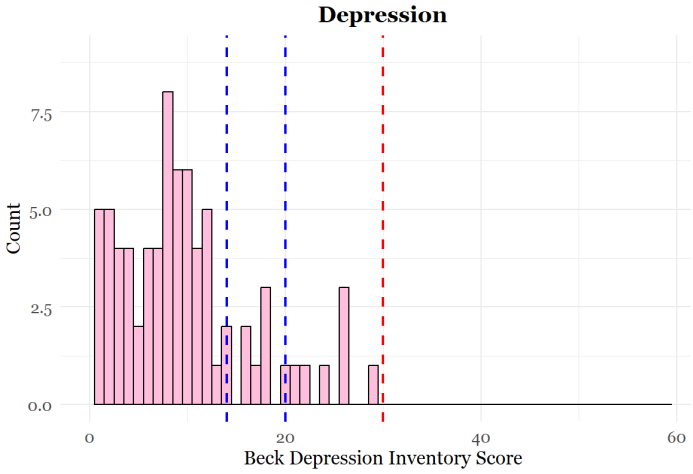
"The user expressed a desire to start a new habit and engaged in a discussion with the chatbot to identify the specific habit they wish to develop. The session focused on understanding what might be preventing the user from initiating this habit and exploring ways to incorporate it into their daily routine."

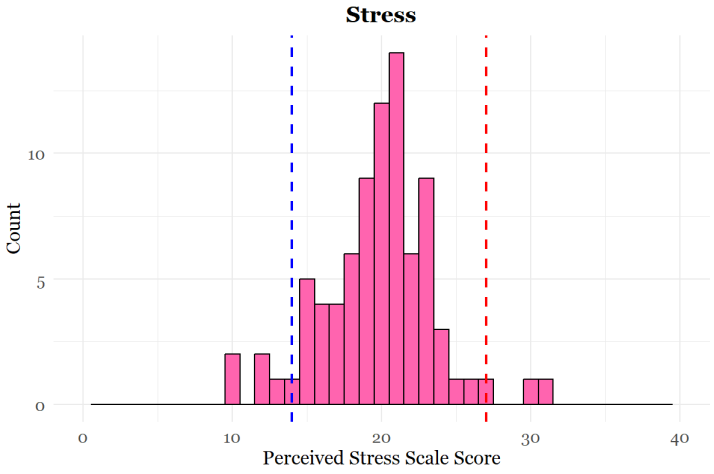
III: Results

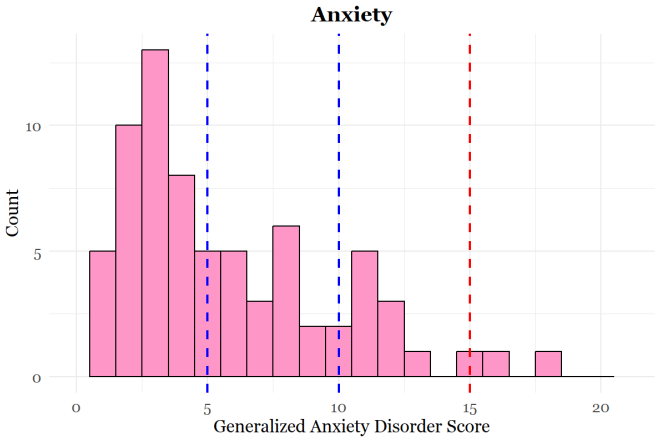
Baseline Scores

All baseline scores were acceptably distributed, with absolute skewness values < 1.0. There were no statistically significant differences in baseline distress scores between the intervention and control groups, indicating successful randomization. For baseline depression measured by the BDI, both groups had a median score of 8. The interquartile range (IQR) was 4-12 for the control group and 2-11 for the intervention group, suggesting that most participants fell within the no to minimal depression range (Beck et al., 1996). Baseline anxiety scores were similarly low, with a median of 4 in both groups; the IQR was 2–8 for the control group and 3–8 for the intervention group, indicating minimal to mild anxiety in the majority of the sample (National HIV Curriculum, 2024). These distributions of depression and anxiety were right-skewed, consistent with expectations for a non-clinical population and aligned with the study's target group. Baseline stress levels were higher: both groups had a median score of 20, with an IQR of 19–23 for the control group and 17–21 for the intervention group. These values suggest that the majority of participants experienced moderate stress (State of New Hampshire Employee Assistance Program, 1983).

Figure 1. Baseline Scores in Depression, Anxiety and Stress



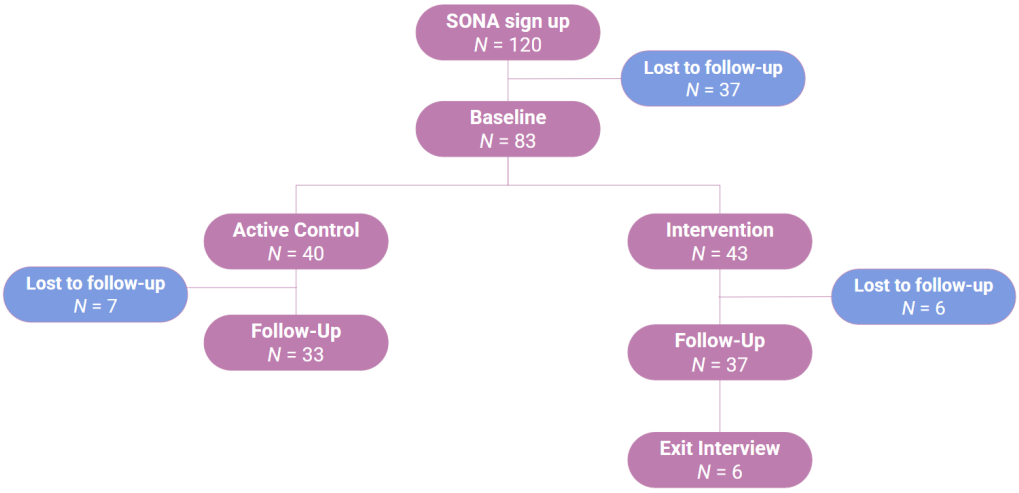




Attrition and Adherence

The attrition rate from sign-up to enrollment process was 30.8% (*n* = 37). There was a moderate amount of attrition in both groups after enrollment. The overall attrition rate from enrollment was 17.5% for the control group (*n* = 7) and 14.0% for the intervention group (*n* = 6). Attrition was not predicted by baseline distress scores or demographic variables. The most commonly cited reasons for dropping out were: 1) not having enough time and 2) having found better credit options on Sona Systems. See Consort Flow Diagram in Figure 2.

Figure 2. Consort Flow Diagram



Adherence to the 30 minutes per week dose was informed by benchmarks in therapy efficacy studies (National Institutes of Health, 2022). This was assessed in two ways. First, participants' website traffic, including time and frequency of use, was recorded. Second, participants were asked in their follow-up survey how seriously they used their assigned website on a scale of 0 (“Not seriously at all) to 4 (“Very seriously). Since the majority of participants (80%) did not meet the protocol of 30 minutes per week (120 minutes total after four weeks), all participants were included in the analyses controlling for time use.

Efficacy

Depression

Two outliers detected using z-scores > 3 for pre-post BDI difference were removed from statistical analysis for efficacy on depression. Within groups, paired samples *t*-tests showed that the control group had a statistical significant decrease in BDI score [$p < 0.001$] over four weeks of use, and the intervention group had a marginally significant decrease [$p = 0.067$]. The control

group had a mean difference in BDI of 2.74 (*CI: 1.28 - 4.20*), and the intervention group had a mean difference in BDI of 1.14 (*CI: -0.09 - 2.36)*. In terms of between-group differences, ANCOVA marginally predicted follow-up BDI scores by condition [$p = 0.088$], controlling for baseline scores, demographic variables, time, and frequency of use.

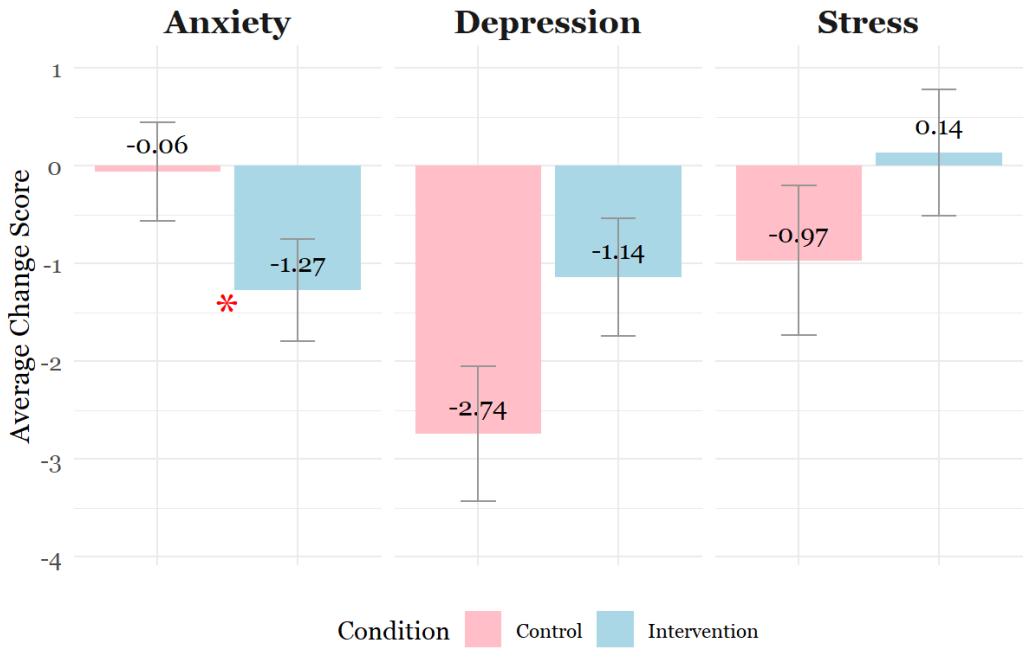
Anxiety

Within groups, paired samples *t*-tests showed that the intervention group had a statistically significant decrease in GAD-7 scores [$p = 0.020$] over four weeks of use, while the control group did not see any significant changes [$p = 0.905$]. The control group had a mean difference in GAD-7 of 0.06 (*CI: -0.96 - 1.08*), and the intervention group had a mean difference in GAD-7 of 1.27 (*CI: 0.211 - 2.33)*. In terms of between-group differences, ANCOVAs significantly predicted follow-up GAD-7 scores by condition [$p = 0.029$], controlling for baseline scores, demographic variables, time, and frequency of use.

Stress

Within groups, paired samples *t*-tests showed that neither the intervention group nor the control group had any significant changes in PSS-10 score over four weeks of use. The control group had a mean difference in PSS-10 of 0.97 (*CI: -0.59 - 2.53; p = 0.21*), and the intervention group had a mean difference of -0.14 (*CI: -1.45 - 1.18; p = 0.84)*. There were no significant between-group differences in follow-up PSS-10 scores as predicted by ANCOVAs by condition [$p = 0.84$], controlling for baseline scores, demographic variables, time, and frequency of use.

Figure 3. Pre-post Changes in Depression, Stress and Anxiety after 4-weeks by Condition



Interaction Effect with Time Use and Frequency

Without accounting for interactions between condition and time metrics in ANCOVA, higher frequency of use by itself was predictive of lower follow-up BDI [$p = 0.027$], GAD-7 [$p = 0.002$] and PSS-10 [$p = 0.033$], while lower duration of use marginally predicted lower follow-up BDI [$p = 0.071$]. This was shown across conditions, controlling for baseline distress, demographic variables, and time use. There was a time interaction between duration and condition for follow-up BDI scores [$p = 0.021$], where participants who used the intervention less showed lower BDI follow-up than those who used it more, and vice versa for the control group. This time interaction was not predictive for GAD-7 or PSS-10 follow-up scores.

Fig 4. Time Interaction of Change in BDI scores with Condition

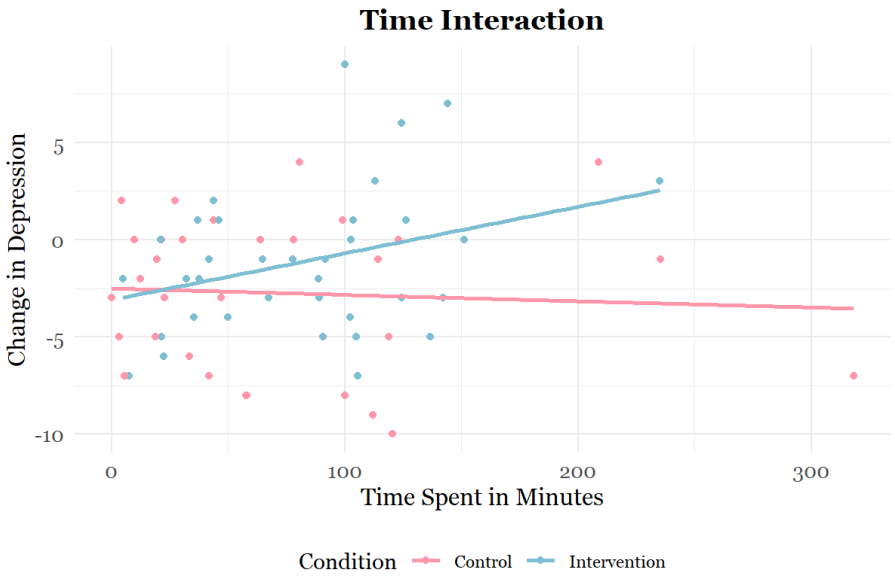
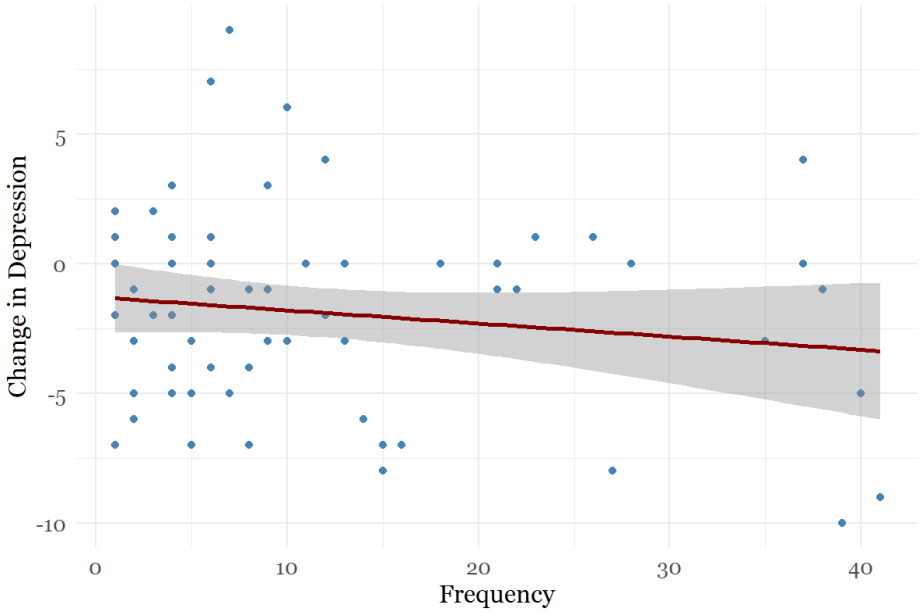


Fig 5. Change in BDI scores over Frequency of use



Acceptability

Acceptability Survey

There were no outliers in the scores for the control group acceptability survey (n = 33). The mean score per item was 1.86 (slightly below neutral) with a standard deviation of 0.765.

Conversely, there was one outlier in the scores for the intervention group acceptability survey (n = 37) with a z-score of -3.22. This participant reported a mean score of 0.29, strongly disagreeing with the acceptability of Elomia. The mean score per item was 2.27 (slightly above neutral) with a standard deviation of 0.614, including the outlier.

Figure 6. Acceptability Surveys for the Control Group and the Intervention Group

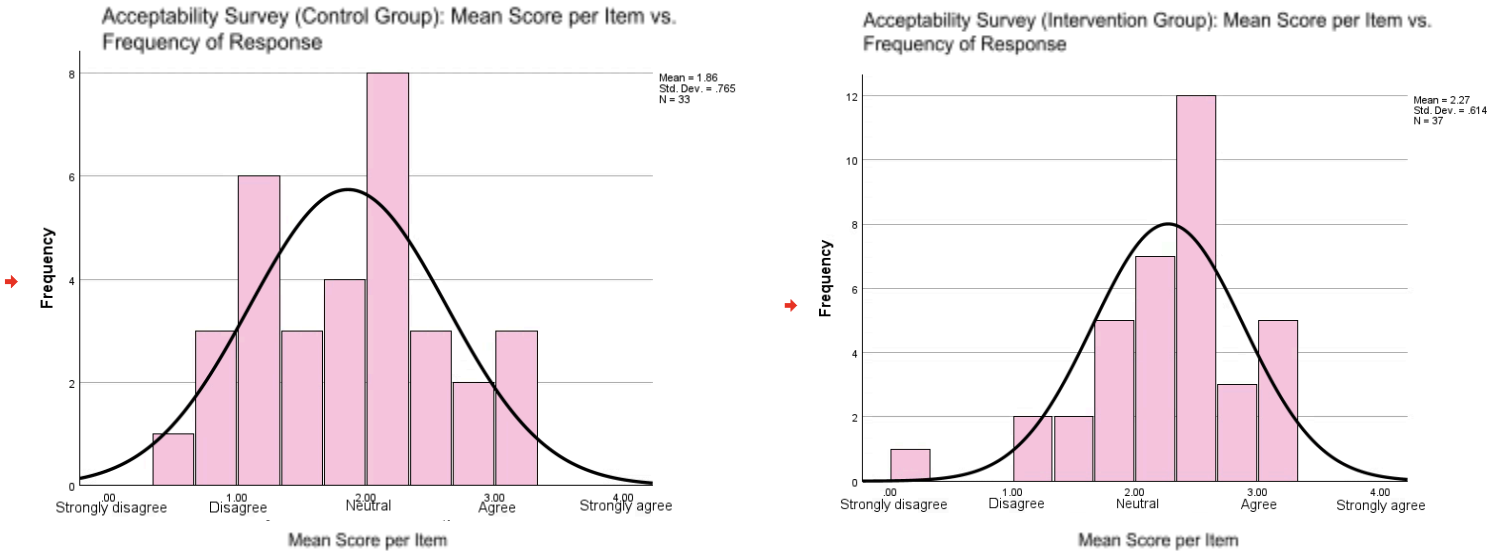
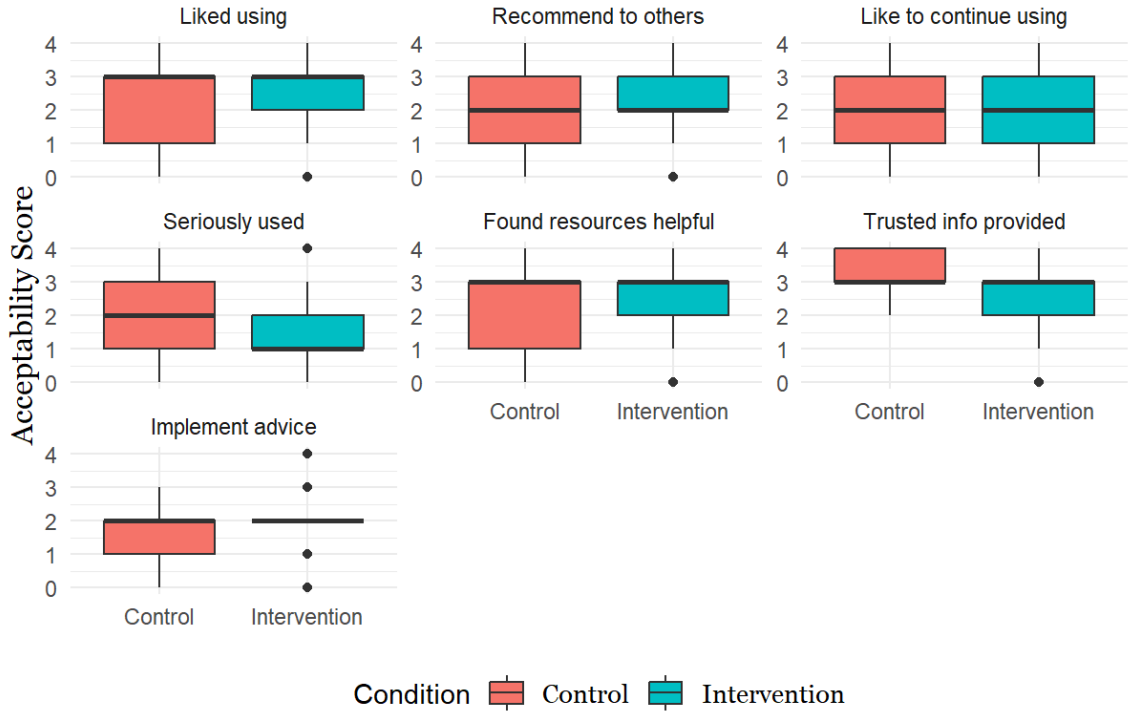


Figure 7. Independent Samples *t*-test for Acceptability Survey Responses

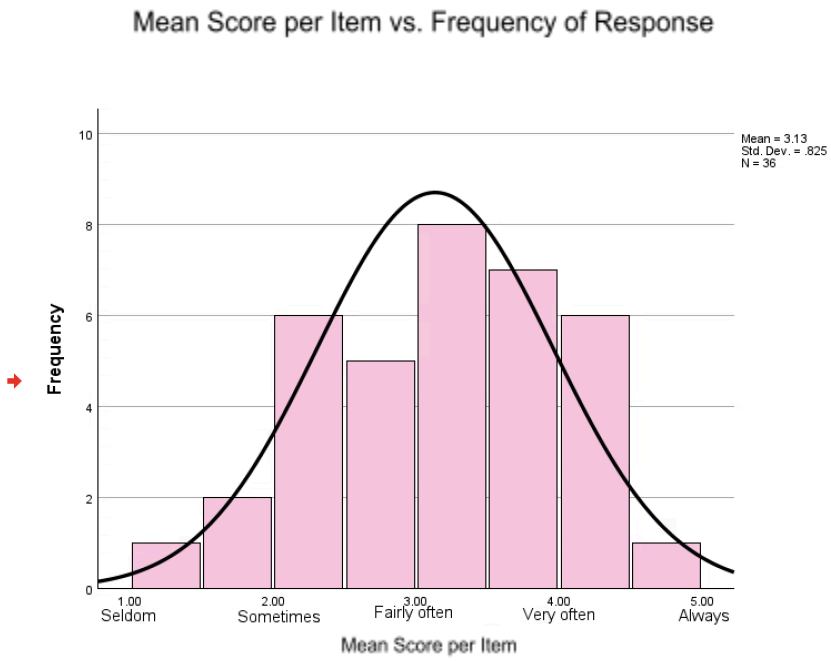


An independent samples *t*-test was conducted to compare the mean acceptability scores for individual responses between the intervention group ($n = 37$) and the control group ($n = 33$). Results indicated statistically significant differences for two items: "I trusted the advice and information provided by Elomia/the research website" [$p < 0.001$] and "How seriously did you use Elomia/this research website" [$p = 0.021$]. For both questions, the Elomia group reported significantly lower mean scores than the control group.

Therapeutic Alliance

There were no outliers in the survey responses for the modified working alliance inventory ($n = 36$). The mean score per item was 3.13 (slightly above fairly often) with a standard deviation of 0.825.

Figure 8. Modified Working Alliance Inventory Histogram



Qualitative Codes

The session summaries were coded on common stressors identified in the sessions. The 14 common stressor codes were: academics (stress or other); social relationships (romantic partner, family, peer, or other); self-image/self-worth; religious concerns; financial concerns; personal interests/development; establishing lifestyle routines; stress or uncertainty about the future (career or other); major life stressor (death, trauma, long-term injury/chronic illness, divorce, or world events); concerns about a clinically diagnosable disorder; productivity/time management; sleep concerns; minor illness and injury; and existential concerns. The three types of interactions were: 1) positive; 2) not mentioned; and 3) negative.

Common Stressors

The interrater reliability Cohen's kappa showed moderate to substantial acceptability for most categories. Two less common codes—sleep and personal interests/developments—had fair acceptability, listed in Table 2.

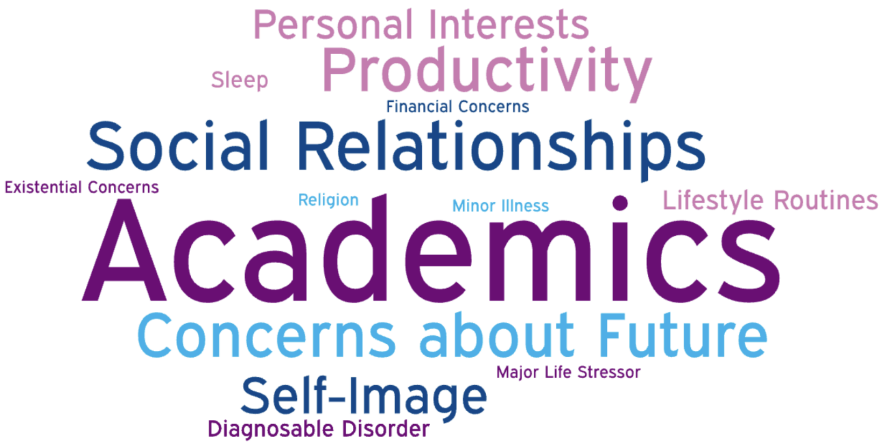
Table 2. Cohen's Kappas Across Categories

Common Stressors	**Cohen's Kappa**
Academics	0.66
Social Relationships	0.81
Self-Image/Self-Worth	0.52
Religious Concerns	0.80
Financial Concerns	1
Personal Interests/Developments	0.39
Establishing Lifestyle Routines	0.67
Stress or Uncertainty About the Future	0.56
Major Life Stressors	0.49
Concerns About a Clinically Diagnosable Disorder	0.51
Getting Work Done	0.65
Sleep Concerns	0.34
Minor Illness and Injury	0.49
Existential Concerns	0.57

As seen in the word cloud on Figure 9, the larger the word is, the more frequently the word was coded for students. The most frequently discussed topics were academic stress, career

uncertainty/concerns about the future, and social relationships (including friendships, romantic partners, and family dynamics), and productivity/time management.

Figure 9. Common Stressors Word Cloud



Participant Feedback

Participants using Elomia shared mixed and even sometimes contradictory reviews about the chatbot in the follow-up surveys. Positive feedback included the following: ease of access; "organized and methodical"; "like talking to a friend" or "trusted confidante"; provided "tailored tips" and "helpful advice"; useful for self-reflections, check-ins, or as an outlet between therapy sessions; "versatile" in content knowledge; removed stigma of talking to a human; respectful and cared for user; and "intelligent and satisfactory" after some prompting. Negative feedback included the following: not "natural" or "human"; can't have a genuine back-and-forth conversation; "repetitive," "predictable," and "scripted"; impersonal and dismissive tone; inability to understand nuance, jokes, or contexts; "boring and slow" or not giving specific

advice; uncertainty about confidentiality; overemphasis on mindfulness content; unnerved by idea of an "AI therapist"; and other sites (e.g., ChatGPT) were seen as better alternatives.

In the exit interviews, participants also had varying thoughts on the chatbot. Here is one quote from a student who enjoyed using Elomia: "I feel like it was a great outlet for me, because me personally … I would love to get a therapist and have those appointments, but I don't always have time to schedule those…". Here is another quote from a student who was hesitant about using Elomia: "It kept saying like, 'oh, I understand. I see your frustration or stuff.' But like it really doesn't. [It's] impossible for it to actually care about me… It wasn't human … it was pretending to be human."

IV: Discussion

Efficacy

The study results support one portion of the hypothesis that participants using the mental health chatbot over a four-week period showed higher improvements in anxiety than the control group, which showed no significant change. In terms of depression, the evidence suggested that the control group led to marginally higher reductions in BDI score than the intervention group, but using the chatbot still significantly reduced depression compared to baseline. These findings may suggest that the chatbot is more suitable to address anxiety than depression when compared to the wellness modules. Neither group led to any significant reductions in stress. Considering that college is a highly stressful environment, as shown with a moderate baseline stress level in our sample, it is possible that either the chatbot is not an ideal tool for lowering stress or that stress is more resistant to change and may need longer time periods or different interventions to see improvements. This lack of effect may also be due to the complexity of stress as a construct, which can be influenced by other factors beyond the scope of the intervention.

The analysis of the interaction between frequency and time spent with the condition provided insights into how usage patterns influenced outcomes. Higher frequency of use was significantly associated with greater reductions in depression, anxiety, and stress, suggesting the importance of consistent engagement with the tools. However, the duration of use only showed a marginal effect on depression and was not predictive in anxiety and stress**,** suggesting that time spent using the mental health chatbot may not be as critical as the frequency of use in producing meaningful changes in depression scores. Furthermore, the time interaction with the conditions revealed that participants who used the chatbot less reported lower depression follow-up

compared to those who spent more time. This may suggest a diminishing returns effect, where excessive use of the chatbot may not provide additional benefits and could potentially lead to decreased engagement or harm.

The study experienced moderate attrition starting from enrollment, but importantly, attrition was not predicted by baseline distress scores or demographic variables, indicating that randomization effectively controlled for these factors. Since attrition rate was highly similar across both conditions, that dropout was unlikely to be attributable to the chatbot itself and was more likely related to other factors of study recruitment through a system that incentivizes with academic credit, such as having no time or finding better credit options. However, only ⅕ of drop-outs responded to our follow-up inquiry. There is still the possibility that some individuals did not want to engage with the chatbot.

Adherence to the conditions was suboptimal, with 80% of participants failing to meet the protocol of 30 minutes per week (120 minutes in total over four weeks). This deviation from the intended dose could have influenced the outcomes, and as such, time spent is controlled for all analyses. The lack of adherence suggests that the prescribed duration of use might not be optimal for a mental health chatbot intervention for a college student population. Coupled with the finding that frequency was a more significant predictor of larger improvements and that spending more time on the website over the course of four weeks was predictive of fewer improvements, future research should be done to investigate an optimal dosage for engaging with a mental health chatbot.

Acceptability

Session Summaries and Exit Interview Discussion

The session summaries showed that the vast majority of participants engaged meaningfully with the chatbot by discussing a wide variety of different stressors. The most frequently cited stressors gave a better understanding of how students were using the chatbot—primarily as a tool to process a broad range of emotionally relevant, college-related stressors.

Positive comments from several students in the feedback response and in one exit interview reflect on one of the chatbot's core strengths: accessibility. For students with tight schedules, long waitlists, or hesitation around therapy, a tool like Elomia can provide an immediate, low-pressure space to process their emotions. This suggests that chatbots like Elomia can therefore help address one of the core problems of the mental health crisis on college campuses, the lack of accessibility to care. On the other hand, critical feedback from another student in a separate exit interview emphasized a different concern—the lack of humanness. That is, it would be "impossible" for a chatbot like Elomia to empathize and understand like a human. It can pretend to empathize and to understand, but there will always be a disconnect for some users.

AS and MWAI Discussion

The results of the AS and the MWAI demonstrated a moderate variability in responses. Participants varied in their opinions of the chatbot and wellness modules as effective and

acceptable, ranging from extreme dislike to strong approval. The exit interviews provided additional insight into these diverse opinions.

These findings suggest that individuals had very different experiences when interacting with the chatbot. These results suggest that, as AI health technology continues to grow, individuals may have a wide range of opinions and interactions with these resources. Some might find them acceptable and helpful, readily taking advantage of them, while others will be more wary and hesitant. The importance of tailored and individualized care should be emphasized when designing these resources as treatments. Our exit interviews suggest that participants who have a more positive acceptance of AI tools in general tend to have a more positive experience and higher acceptability, although we did not have baseline data to support this analysis. It is possible that AI mental health chatbots might be more efficacious for some people than others due to their pre-conceived acceptability—a potential next step.

The independent samples *t*-test revealed that participants found Elomia less trustworthy and took it less seriously than the control website. One possible explanation for the difference is that interacting with the chatbot demanded a higher level of engagement and vulnerability than the control website. To take the chatbot seriously, participants had to actively engage in conversation about their personal lives, requiring not only more concentration, but also more emotional openness and self-expression. On the other hand, to take the research website seriously, participants could passively engage by reading the materials supplied on the website. As a result, when asked to appraise how seriously they took the intervention, participants in the control group may have been more likely to feel they had met expectations, since the threshold for engagement was lower and more straightforward. In terms of trustworthiness, it may have been easier for participants to place trust in static, brand-named documents than in an unfamiliar

AI chatbot. The wellness modules also provided citations within various materials, making it seem more legitimate than the chatbot, which only listed techniques without citing their sources. Our AS survey responses also suggest that the Elomia chatbot could be trained to have a wider range of resources and to be more accommodating, as several qualitative AS responses mentioned that its resources and advice were "not specific" and "repetitive," specifically with overemphasis on mindfulness or the Pomodoro techniques.

Another major reason for the difference in responses could be that participants were skeptical of AI-enabled mental health chatbots generally. AI is still very young, and dystopian ideas about AI abound, from Arnold Schwarznegger's character in *The Terminator* to Hal 9000 in *2001: A Space Odyssey*. Participants may consciously or subconsciously distrust AI, making them more skeptical of a chatbot designed for something as sensitive as mental health. This explanation holds large implications for the future of digital mental health technology, suggesting that AI mental health could be approached with a deep sense of mistrust and apprehension for some subset of the populations, even younger generations who are reported to be more accepting of AI and digital technology (Chan & Lee, 2023).

Limitations & Future Directions

There are several limitations to this study. First, the findings may not be generalizable to other chatbots, as Elomia is just one among many available technologies. Notably, since some participants in this study reported that they found ChatGPT more helpful, a direct comparison between multiple chatbots—especially those formally trained in psychotherapeutic techniques versus those that are broadly accessible but not clinically optimized—would provide insight into which chatbot and features could be most effective. Furthermore, the Elomia chatbot was also

designed specifically as a first-aid tool, which may limit the applicability of these findings to chatbots intended to complement therapy or provide higher levels of care. Further studies should also explore how they perform relative to, or in conjunction with, traditional in-person psychotherapy: is there a certain combination of care that is optimal for addressing mental health in a college student population?

The second limitation is that the study collected data through self-report measures, which are liable to biases, such as demand characteristics, recall bias, and subjective interpretation by participants (Herzog & Bowman, 2011). The third limitation is the length of the testing period. The length of the intervention was four weeks in this study, whereas CBT's normally recommended minimum use is six to eight weeks (Beck, 2011). The limited time frame might explain why there were no substantial reductions in BDI and GAD-7 scores and why there were no significant changes in stress. Future research could lengthen the testing period or include longer follow-up time frames in order to assess longer lasting effects of these resources.

Additionally, another limitation is the low adherence to the study time use protocol, which might have limited the intervention's effectiveness and undermine acceptability. Future studies should explore how to improve engagement and adherence, possibly by tailoring the intervention dose to individual needs, and examine the optimal duration and frequency of use for mental health chatbots.

The fifth limitation is limited external validity due to the study's mandatory 30-minute weekly usage for course credit. This protocol prevents us from capturing true time usage patterns and from determining how many students might be interested in utilizing these resources on their own. It is possible we might see a larger efficacy and acceptability effect if students voluntarily used the chatbot rather than participated for study credits. Lastly, the sixth limitation is of limited

generalizability, as the sample was undergraduate students at the University of Pennsylvania who were taking psychology courses. To ensure generalizability to the college student population, future studies should recruit more broadly across departments and schools to ensure more representative demographics.

There are several promising future directions as a result of this study. In line with addressing the problem of unmet need in college mental healthcare, we would like to test mental health chatbots on students who are on the waitlist to receive care at college counseling centers. This would help assess the chatbot's potential role within the lower-intensity tiers of a stepped care model. Secondly, further research should investigate the factors underlying positive versus negative user experiences with the chatbot, in order to identify key barriers and facilitators to engagement with mental health chatbots and other AI-based mental health tools.

Conclusion

Elomia shows promise as a first-aid mental health tool for alleviating anxiety and depression in college students. Higher frequency of using the chatbot was a significant predictor of greater improvements in anxiety, depression, and stress more than the length of time use, suggesting that consistency might be more effective than duration for chatbots. The wide variability in user experience shows that acceptability and engagement with the chatbot still need improvement, especially for those who value real human connection.

Acknowledgement

We want to give our sincere thank you to Dr. Melissa Hunt for providing guidance and mentorship, Anastasiia and the Elomia team for their collaboration, and all the study participants!

Compliance with Ethical Standards

Dr. Melissa Hunt, Camellia Bui, Luke Finkelstein, and Emily Albert declare that they have no conflicts of interest. All procedures followed were in accordance with the ethical standards of the responsible committee on human experimentation (institutional and national) and with the Helsinki declaration of 1975, as revised in 2000. Informed Consent was obtained from all individual participants included in the study.

V: References

References

Abrams, Z. (2022, October 12). *Student mental health is in crisis. campuses are rethinking their approach*. American Psychological Association. https://www.apa.org/monitor/2022/10/mental-health-campus-care

Alanezi, F. (2024). Assessing the Effectiveness of ChatGPT in Delivering Mental Health Support: A Qualitative Study. *Journal of Multidisciplinary Healthcare*, *Volume 17*, 461–471. https://doi.org/10.2147/jmdh.s447368

American College Health Association. (2024). *Academic Year 2023-2024 - ACHA*. ACHA. https://www.acha.org/ncha/data-results/survey-results/academic-year-2023-2024/

Beck, A. T., Steer, R. A., & Brown, G. K. (1996). *BDI-II, Beck depression inventory: Manual*. Psychological Corporation.

Beck, J. S. (2011). *Cognitive Behavior Therapy: Basics and beyond* (2nd ed.). Guilford Press.

Benthem, P. van, Lamers, A., Blanken, P., R. Spijkerman, R. R.J.M. Vermeiren, & Hendriks, V. M. (2024). The working alliance inventory – short version: psychometric properties of the patient and therapist form in youth mental health and addiction care. *BMC Psychology*, *12*(1). https://doi.org/10.1186/s40359-024-01754-1

Billings, K. R. (2020). Stigma in Class: Mental Illness, Social Status, and Tokenism in Elite College Culture. *Sociological Perspectives*, *64*(2), 073112142092187. https://doi.org/10.1177/0731121420921878

Blanco, C., Okuda, M., Wright, C., Hasin, D. S., Grant, B. F., Liu, S.-M., & Olfson, M. (2008). Mental Health of College Students and Their Non–College-Attending Peers. *Archives of General Psychiatry*, *65*(12), 1429. https://doi.org/10.1001/archpsyc.65.12.1429

Buchanan, J. L. (2012). Prevention of Depression in the College Student Population: A Review of the Literature. *Archives of Psychiatric Nursing*, *26*(1), 21–42. https://doi.org/10.1016/j.apnu.2011.03.003

Chan, C., & Lee, K. (2023). The AI Generation gap: Are Gen Z Students More Interested in Adopting Generative AI Such as ChatGPT in Teaching and Learning than Their Gen X and Millennial Generation teachers? *Smart Learning Environments*, *10*(1). https://doi.org/10.1186/s40561-023-00269-3

Chen, J., Yuan, D., Dong, R., Cai, J., Ai, Z., & Zhou, S. (2024). Artificial intelligence significantly facilitates development in the mental health of college students: a bibliometric analysis. *Frontiers in Psychology*, *15*. https://doi.org/10.3389/fpsyg.2024.1375294

Cohen, K. A., Graham, A. K., & Lattie, E. G. (2020). Aligning students and counseling centers on student mental health needs and treatment resources. *Journal of American College Health*, *70*(3), 1–9. https://doi.org/10.1080/07448481.2020.1762611

Cohen, S., Kamarck, T., & Mermelstein, R. (1983). *Perceived Stress Scale*. Psycnet.apa.org. https://psycnet.apa.org/doiLanding?doi=10.1037%2Ft02889-000

Creswell, J. D. (2017). Mindfulness Interventions. *Annual Review of Psychology*, *68*(1), 491–516. https://doi.org/10.1146/annurev-psych-042716-051139

Dekker, I., De Jong, E. M., Schippers, M. C., De Bruijn-Smolders, M., Alexiou, A., & Giesbers, B. (2020). Optimizing Students' Mental Health and Academic Performance: AI-Enhanced Life Crafting. *Frontiers in Psychology*, *11*(11). https://doi.org/10.3389/fpsyg.2020.01063

Dosovitsky, G., Pineda, B. S., Jacobson, N. C., Chang, C., Escoredo, M., & Bunge, E. L. (2020). Artificial Intelligence Chatbot for Depression: Descriptive Study of Usage. *JMIR Formative Research*, *4*(11), e17065. https://doi.org/10.2196/17065

Eisenberg, D., Golberstein, E., & Gollust, S. E. (2007). Help-Seeking and Access to Mental Health Care in a University Student Population. *Medical Care*, *45*(7), 594–601. https://doi.org/10.1097/mlr.0b013e31803bb4c1

Eisenberg, D., Golberstein, E., & Hunt, J. B. (2009). Mental Health and Academic Success in College. *The B.E. Journal of Economic Analysis & Policy*, *9*(1). https://doi.org/10.2202/1935-1682.2191

Eisenberg, D., Hunt, J., Speer, N., & Zivin, K. (2011). Mental Health Service Utilization Among College Students in the United States. *The Journal of Nervous and Mental Disease*, *199*(5), 301–308. https://doi.org/10.1097/nmd.0b013e3182175123

Eltahawy, L., Essig, T., Myszkowski, N., & Trub, L. (2023). Can robots do therapy?: Examining the efficacy of a CBT bot in comparison with other behavioral intervention technologies in alleviating mental health symptoms. *Computers in Human Behavior: Artificial Humans*, *2*(1), 100035. https://doi.org/10.1016/j.chbah.2023.100035

Fawcett, E., Neary, M., Ginsburg, R., & Cornish, P. (2019). Comparing the effectiveness of individual and group therapy for students with symptoms of anxiety and depression: A randomized pilot study. *Journal of American College Health*, *68*(4), 430–437. https://doi.org/10.1080/07448481.2019.1577862

Fiske, A., Henningsen, P., & Buyx, A. (2019). Your Robot Therapist Will See You Now: Ethical Implications of Embodied Artificial Intelligence in Psychiatry, Psychology, and Psychotherapy. *Journal of Medical Internet Research*, *21*(5). https://doi.org/10.2196/13216

Gallagher, R. P., & Taylor, R. (2015). *National Survey of College Counseling Centers 2014*. D-Scholarship.pitt.edu. https://d-scholarship.pitt.edu/28178/

Garlow, S. J., Rosenberg, J., Moore, J. D., Haas, A. P., Koestner, B., Hendin, H., & Nemeroff, C. B. (2008). Depression, desperation, and suicidal ideation in college students: results from the American Foundation for Suicide Prevention College Screening Project at Emory University. *Depression and Anxiety*, *25*(6), 482–488. https://doi.org/10.1002/da.20321

Golden, A., & Aboujaoude, E. (2024). The Framework for AI Tool Assessment in Mental Health (FAITA ‐ Mental Health): a scale for evaluating AI‐powered mental health tools. *World Psychiatry*, *23*(3), 444–445. https://doi.org/10.1002/wps.21248

Graham, S., Depp, C., Lee, E. E., Nebeker, C., Tu, X., Kim, H.-C., & Jeste, D. V. (2019). Artificial Intelligence for Mental Health and Mental Illnesses: an Overview. *Current Psychiatry Reports*, *21*(11), 116. https://doi.org/10.1007/s11920-019-1094-0

Halliburton, A. E., Hill, M. B., Dawson, B. L., Hightower, J. M., & Rueden, H. (2021). Increased Stress, Declining Mental Health: Emerging Adults' Experiences in College During COVID-19. *Emerging Adulthood*, *9*(5), 216769682110253. https://doi.org/10.1177/21676968211025348

He, Y., Yang, L., Zhu, X., Wu, B., Zhang, S., Qian, C., & Tian, T. (2022). Mental Health Chatbot for Young Adults With Depressive Symptoms During the COVID-19 Pandemic: Single-Blind, Three-Arm Randomized Controlled Trial. *Journal of Medical Internet Research*, *24*(11), e40719. https://doi.org/10.2196/40719

Herzog, S., & Bowman, N. A. (2011). *Validity and Limitations of College Student Self-Report Data*. John Wiley & Sons.

Holm, S. (2024). Ethical trade-offs in AI for mental health. *Frontiers in Psychiatry*, *15*. https://doi.org/10.3389/fpsyt.2024.1407562

Hunt, M. G., Rodriguez, L., & Marcelle, E. (2017, October 1). *A Cognitive Behavioral Therapy Workbook Delivered Online with Minimal Therapist Feedback Improves Quality of Life for Inflammatory Bowel Disease Patients.* https://www.researchgate.net/publication/320427320_A_Cognitive_Behavioral_Therapy_Workbook_Delivered_Online_with_Minimal_Therapist_Feedback_Improves_Quality_of_Life_for_Inflammatory_Bowel_Disease_Patients

Kim, E.-H., Coumar, A., Lober, W. B., & Kim, Y. (2011). Addressing Mental Health Epidemic Among University Students via Web-based, Self-Screening, and Referral System: A Preliminary Study. *IEEE Transactions on Information Technology in Biomedicine*, *15*(2), 301–307. https://doi.org/10.1109/titb.2011.2107561

Linthicum, K. P., Schafer, K. M., & Ribeiro, J. D. (2019). Machine learning in suicide science: Applications and ethics. *Behavioral Sciences & the Law*. https://doi.org/10.1002/bsl.2392

Lipson, S. K., Lattie, E. G., & Eisenberg, D. (2019). Increased Rates of Mental Health Service Utilization by U.S. College Students: 10-Year Population-Level Trends (2007–2017). *Psychiatric Services*, *70*(1), 60–63. https://doi.org/10.1176/appi.ps.201800332

Lipson, S. K., Zhou, S., Abelson, S., Heinze, J., Jirsa, M., Morigney, J., Patterson, A., Singh, M., & Eisenberg, D. (2022). Trends in college student mental health and help-seeking by race/ethnicity: Findings from the national healthy minds study, 2013–2021. *Journal of Affective Disorders*, *306*(1), 138–147. https://doi.org/10.1016/j.jad.2022.03.038

Luxton, D. D. (2014). Recommendations for the ethical use and design of artificial intelligent care providers. *Artificial Intelligence in Medicine*, *62*(1), 1–10. https://doi.org/10.1016/j.artmed.2014.06.004

Maurya, R. K., Montesinos, S., Bogomaz, M., & DeDiego, A. C. (2024). Assessing the use of ChatGPT as a psychoeducational tool for mental health practice. *Counselling and Psychotherapy Research*. https://doi.org/10.1002/capr.12759

Melo, A., Silva, I., & Lopes, J. (2024). ChatGPT: A Pilot Study on a Promising Tool for Mental Health Support in Psychiatric Inpatient Care. *International Journal of Psychiatric Trainees*. https://doi.org/10.55922/001c.92367

National HIV Curriculum. (2024). *Generalized Anxiety Disorder 7-item (GAD-7) - Mental Disorders Screening - National HIV Curriculum*. Uw.edu. https://www.hiv.uw.edu/page/mental-health-screening/gad-7

National Institutes of Health. (2022, June 2). *Cognitive Behavioral Therapy*. National Library of Medicine. https://www.ncbi.nlm.nih.gov/books/NBK279297/

Nguyen-Feng, V. N., Greer, C. S., & Frazier, P. (2017). Using online interventions to deliver college student mental health resources: Evidence from randomized clinical trials. *Psychological Services*, *14*(4), 481–489. https://doi.org/10.1037/ser0000154

O'Leary, K. (2023). Human–AI collaboration boosts mental health support. *Nature Medicine*. https://doi.org/10.1038/d41591-023-00022-w

Oswalt, S. B., Lederer, A. M., Chestnut-Steich, K., Day, C., Halbritter, A., & Ortiz, D. (2020). Trends in college students' mental health diagnoses and utilization of services, 2009–2015. *Journal of American College Health*, *68*(1), 1–11. https://doi.org/10.1080/07448481.2018.1515748

Pedrelli, P., Nyer, M., Yeung, A., Zulauf, C., & Wilens, T. (2015). College Students: Mental Health Problems and Treatment Considerations. *Academic Psychiatry*, *39*(5), 503–511. https://doi.org/10.1007/s40596-014-0205-9

Prince, J. P. (2015). University student counseling and mental health in the United States: Trends and challenges. *Mental Health & Prevention*, *3*(1-2), 5–10. https://doi.org/10.1016/j.mhp.2015.03.001

Sapadin, K., & L. G. Hollander, B. (2024). *Distinguishing the Need for Crisis Mental Health Services Among College Students*. Apa.org. https://psycnet.apa.org/fulltext/2021-25597-001.html

Shatte, A. B. R., Hutchinson, D. M., & Teague, S. J. (2019). Machine learning in mental health: a scoping review of methods and applications. *Psychological Medicine*, *49*(09), 1426–1448. https://doi.org/10.1017/s0033291719000151

Shay, J. J. (2021). Terrified of Group Therapy: Investigating Obstacles to Entering or Leading Groups. *American Journal of Psychotherapy*, *74*(2), 71–75. https://doi.org/10.1176/appi.psychotherapy.20200033

Spitzer, R. L., Kroenke, K., Williams, J. B. W., & Löwe, B. (2006). A Brief Measure for Assessing Generalized Anxiety disorder: the GAD-7. *Archives of Internal Medicine*, *166*(10), 1092–1097. https://doi.org/10.1001/archinte.166.10.1092

State of New Hampshire Employee Assistance Program. (1983). *Perceived Stress Scale*. https://www.das.nh.gov/wellness/docs/percieved%20stress%20scale.pdf

Sweeney, C., Potts, C., Ennis, E., Bond, R., Mulvenna, M. D., O'neill, S., Malcolm, M., Kuosmanen, L., Kostenius, C., Vakaloudis, A., Mcconvey, G., Turkington, R., Hanna, D., Nieminen, H., Vartiainen, A.-K., Robertson, A., & Mctear, M. F. (2021). Can Chatbots Help Support a Person's Mental Health? Perceptions and Views from Mental Healthcare Professionals and Experts. *ACM Transactions on Computing for Healthcare*, *2*(3), 1–15. https://doi.org/10.1145/3453175

Torous, J., & Blease, C. (2024). Generative artificial intelligence in mental health care: potential benefits and current challenges. *World Psychiatry*, *23*(1), 1–2. https://doi.org/10.1002/wps.21148

Wang, P. S., Berglund, P. A., Olfson, M., & Kessler, R. C. (2004). Delays in Initial Treatment Contact after First Onset of a Mental Disorder. *Health Services Research*, *39*(2), 393–416. https://doi.org/10.1111/j.1475-6773.2004.00234.x

Wood, C. I., Yu, Z., Sealy, D.-A., Moss, I., Zigbuo-Wenzler, E., McFadden, C., Landi, D., & Brace, A. M. (2022). Mental health impacts of the COVID-19 pandemic on college students. *Journal of American College Health*, *72*(2), 1–6. https://doi.org/10.1080/07448481.2022.2040515

Wood, M. (2012). The State of Mental Health on College Campuses. *Inquiry: The Journal of the Virginia Community Colleges*, *17*(1), 5–15. https://commons.vccs.edu/inquiry/vol17/iss1/1/

Xiao, H., Carney, D. M., Youn, S. J., Janis, R. A., Castonguay, L. G., Hayes, J. A., & Locke, B. D. (2017). Are we in crisis? National mental health and treatment trends in college counseling centers. *Psychological Services*, *14*(4), 407–415. https://doi.org/10.1037/ser0000130

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

VI: Appendix

Appendix A: Modified Working Alliance Inventory – Short Revised (WAR-SR)

Working Alliance Inventory – Short Revised (WAI-SR)

Instructions: Below is a list of statements and questions about experiences people might have with Elomia. Think about your experience using Elomia, and decide which category best describes your own experience.

IMPORTANT!!! Please take your time to consider each question carefully.

1. As a result of these sessions I am clearer as to how I might be able to change.

   ① ② ③ ④ ⑤

   Seldom Sometimes Fairly Often Very Often Always

2. What I am doing in my sessions with Elomia gives me new ways of looking at my problems.

   ⑤ ④ ③ ② ①

   Always Very Often Fairly Often Sometimes Seldom

3. I feel liked in my interactions with Elomia.

   ① ② ③ ④ ⑤

   Seldom Sometimes Fairly Often Very Often Always

4. I feel like Elomia collaborates with me on setting goals for each session.

   ① ② ③ ④ ⑤

   Seldom Sometimes Fairly Often Very Often Always

5. I feel respected in my interactions with Elomia

   ⑤ ④ ③ ② ①

   Always Very Often Fairly Often Sometimes Seldom

6. Elomia seems to work with me toward mutually agreed-upon goals.

   ⑤ ④ ③ ② ①

   Always Very Often Fairly Often Sometimes Seldom

7. I feel appreciated in my interactions with Elomia.

   ① ② ③ ④ ⑤

   Seldom Sometimes Fairly Often Very Often Always

8. I feel like Elomia and I agree on what is important for me to work on.

⑤ ④ ③ ② ①

Always Very Often Fairly Often Sometimes Seldom

9. In my interactions with Elomia, I feel cared about, even when I reject Elomia's suggestions. ① ② ③ ④ ⑤

Seldom Sometimes Fairly Often Very Often Always

10. I feel that Elomia will help me to accomplish the changes that I want.

⑤ ④ ③ ② ①

Always Very Often Fairly Often Sometimes Seldom

11. Elomia and I seem to have established a good understanding of the kind of changes that would be good for me.

⑤ ④ ③ ② ①

Always Very Often Fairly Often Sometimes Seldom

12. I believe the way Elomia and I are working with my problems is helpful.

① ② ③ ④ ⑤

Seldom Sometimes Fairly Often Very Often Always

Goal Items: 4, 6, 8, 11; Task Items: 1, 2, 10, 12; Bond Items: 3, 5, 7, 9

Reverse scoring on: 2,5,6,8,10,11

Appendix B: Acceptability Survey for Intervention Group

I liked using Elomia.

- Strongly agree
- Agree
- Neutral
- Disagree
- Strongly disagree

I found Elomia helpful.

- Strongly agree
- Agree
- Neutral
- Disagree
- Strongly disagree

I felt comfortable talking with Elomia.

- Strongly agree
- Agree
- Neutral
- Disagree
- Strongly disagree

I felt comfortable sharing my personal information with Elomia.

- Strongly agree
- Agree
- Neutral
- Disagree
- Strongly disagree

I felt comfortable discussing my emotions with Elomia.

- Strongly agree
- Agree

- Neutral
- Disagree
- Strongly disagree

I felt like I could let my guard down with Elomia.

- Strongly agree
- Agree
- Neutral
- Disagree
- Strongly disagree

I felt confident that my private information would remain confidential.

- Strongly agree
- Agree
- Neutral
- Disagree
- Strongly disagree

I trusted the advice and information provided by Elomia during my conversations.

- Strongly agree
- Agree
- Neutral
- Disagree
- Strongly disagree

How often did you implement advice given by Elomia in your daily life?

- Always
- Most of the time
- Sometimes
- Rarely
- Never
- Not applicable (did not receive actionable advice)

How significant were the topics or issues you discussed with Elomia regarding their potential impact on your life?

- Highly significant

- Very significant
- Moderately significant
- Minor significance
- Not significant

Overall, I think an AI Chatbot like Elomia can be an acceptable way to support student mental health.

- Strongly agree
- Agree
- Neutral
- Disagree
- Strongly disagree

I would recommend Elomia to others.

- Strongly agree
- Agree
- Neutral
- Disagree
- Strongly disagree

I would like to continue using Elomia.

- Strongly agree
- Agree
- Neutral
- Disagree
- Strongly disagree

Please share any additional positive experiences you might have with Elomia. Any comment is helpful.

Please share any additional negative experiences or suggestions you might have for Elomia. Any comment is helpful.

How seriously did you use Elomia? Please answer as honestly as possible, this won't impact your study participation or SONA credit!

- Very seriously
- Seriously
- Neutral
- Not very seriously

- Not seriously at all

Appendix C: Acceptability Survey for Control Group

I liked using the research website.

- Strongly agree
- Agree
- Neutral
- Disagree
- Strongly disagree

I found the research website helpful.

- Strongly agree
- Agree
- Neutral
- Disagree
- Strongly disagree

I trusted the advice and information provided by the research website during my conversations.

- Strongly agree
- Agree
- Neutral
- Disagree
- Strongly disagree

How often did you implement advice given by the research website in your daily life?

- Always
- Most of the time
- Sometimes
- Rarely
- Never

- Not applicable (did not receive actionable advice)

I would recommend the research website to others.

- Strongly agree
- Agree
- Neutral
- Disagree
- Strongly disagree

I would like to continue using Elomia.

- Strongly agree
- Agree
- Neutral
- Disagree
- Strongly disagree

Please share any additional positive experiences you might have with the research website. Any comment is helpful.

Please share any additional negative experiences or suggestions you might have for the research website. Any comment is helpful.

How seriously did you use the research website? Please answer as honestly as possible, this won't impact your study participation or SONA credit!

- Very seriously
- Seriously
- Neutral
- Not very seriously
- Not seriously at all