

The Role of AI in Peer Support for Young People: A Study of Preferences for Human- and AI-Generated Responses

JORDYN YOUNG, University of Michigan, USA

LAALA M JAWARA, Drexel University, USA

DIEP N NGUYEN, Drexel University, USA

BRIAN DALY, Drexel University, USA

JINA HUH-YOO, Drexel University, USA

AFSANEH RAZI, Drexel University, USA

Generative Artificial Intelligence (AI) is integrated into everyday technology, including news, education, and social media. AI has further pervaded private conversations as conversational partners, auto-completion, and response suggestions. As social media becomes young people's main method of peer support exchange, we need to understand when and how AI can facilitate and assist in such exchanges in a beneficial, safe, and socially appropriate way. We asked 622 young people to complete an online survey and evaluate blinded human- and AI-generated responses to help-seeking messages. We found that participants preferred the AI-generated response to situations about relationships, self-expression, and physical health. However, when addressing a sensitive topic, like suicidal thoughts, young people preferred the human response. We also discuss the role of training in online peer support exchange and its implications for supporting young people's well-being. Disclaimer: This paper includes sensitive topics, including suicide ideation. Reader discretion is advised.

CCS Concepts: • **Human-centered computing** → **Empirical studies in HCI**; **Empirical studies in collaborative and social computing**; **Natural language interfaces**; **Social media**; • **Social and professional topics** → **Adolescents**.

Additional Key Words and Phrases: Artificial Intelligence (AI), Chatbot, LLM, Human-AI Interaction (HAI), AI-Mediated Communication (AI-MC), Mental Health, Peer Support, Social Support, Youth

ACM Reference Format:

Jordyn Young, Laala M Jawara, Diep N Nguyen, Brian Daly, Jina Huh-Yoo, and Afsaneh Razi. 2024. The Role of AI in Peer Support for Young People: A Study of Preferences for Human- and AI-Generated Responses. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*, May 11–16, 2024, Honolulu, HI, USA. ACM, New York, NY, USA, 28 pages. <https://doi.org/10.1145/3613904.3642574>

1 INTRODUCTION

Today's generation of youth is the first to grow up with ubiquitous internet access. Most (92%) young people access the internet daily, with 24% describing their online presence as "almost constant" [90]. Support is an important reason why young people engage online [8, 68]. Research suggests this is because online help-seeking interactions offer many benefits, including anonymity and privacy [11, 15, 68], ease of access and immediacy [10, 11, 19, 20], connection with others with similar experiences [33, 42, 60], increased perceived control of one's help-seeking journey [11, 60, 68],

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.

Manuscript submitted to ACM

capabilities for self-reliance [28, 60, 68], early access [12, 33, 68], and inclusivity [11, 68]. The wide reach of the internet also means that young people can connect with support-givers from various backgrounds. These interactions can be with professionals in specific areas (e.g., medical doctors [50]), trained support specialists (e.g., people recovered from substance abuse [34]), or peers they may otherwise encounter in school or other parts of daily life [44].

Now, a new evolution of technology is entering the landscape of online support – Artificial Intelligence (AI) – and young people are picking it up fast. Approximately 65% of youth use generative AI tools [1], such as Large Language Models (LLMs), chatbots, and virtual assistants. Preliminary work suggests that AI-driven systems hold promise to build on the previously described benefits of online interactions by providing accessible, convenient, and personalized support [39]. Young people recognize this promise, as more than two-thirds of them have experience with AI-informed mental health apps [37], and many view AI as an acceptable solution to offer mental health support [53]. Furthermore, research suggests that users not only trust AI systems for mental health support but also have a positive experience using them [79]. For example, interventions using AI to deliver psychological support specifically for young people are suggesting that AI could be a feasible, engaging, and effective solution [30, 32, 36].

SIGCHI researchers [6, 64, 69, 70], have thoroughly investigated the significance of online peer support in the lives of youth, discovering that while peers would like to offer support, they sometimes do not know how to help or feel the disclosure was inauthentic [44]. Furthermore, peer support providers sometimes struggle to maintain boundaries to express their difficulties in writing supportive and empathic responses due to lack of role clarity [49]. The notion of education as a solution to the lack of expertise of support givers was validated by a later study where researchers [81] developed an AI system called 'Hailey' which helped develop more empathetic responses to support seekers in a mental health peer support platform. AI does have many limitations, such as its inability to understand human nuances [13], ethical risks of misclassification for sensitive topics [71], and potential algorithmic stigmatization and harms [5]. This has led to continued efforts to explore the potential of AI as a supplementary, rather than a primary, source of support for youth. However, youth's norms around peer support and the extent to which AI can effectively address their unique support needs based on frequent support-seeking topics of youth remain areas of ongoing investigation [35].

Therefore, our goal is to understand how AI compares to the different types of support young people encounter online within the context of the topics that young people specifically seek support about. To address the aforementioned gaps in the literature, we ask the following research questions:

(RQ1) *What are young people's preferences regarding help-giving responses created by humans with various help-giving skills compared to AI-generated responses?*

(RQ2) *How do young people's preferences for help-giving responses vary depending on the topic (relationships, suicidal thoughts, self-expression, or physical health) of the initial help-seeking message?*

To answer these questions, we analyzed 622 survey responses from young people to understand their preferences about pre-written responses from various sources (AI, therapist, adult mentor, and peer) to four help-seeking messages on topics that young people frequently seek support about (relationships, suicidal thoughts, self-expression, and physical health) [44]. We found that young people preferred AI-generated responses about less sensitive topics (relationships, self-expression, and physical health). For more sensitive discussions (suicidal thoughts), participants preferred responses from the adult mentor. Our study contributes to the Human-Computer Interaction (HCI) research community in several ways. This paper provides insights into how young people perceive AI-generated responses in peer support situations. Our results also identify the popular conversational topics among young people where AI-generated responses are, and are not, well received. The findings bring implications for designing AI-powered systems to safely and effectively help youth during online support interactions.

2 RELATED WORK

2.1 Young People’s Mental Health and Online Social Support

More than 450 million people around the world struggle with mental illness [77], and young people are disproportionately likely to be among those who are suffering [46]. In a cross-country survey, the World Health Organization found that nearly one-third of college students tested positive for at least one mental health disorder [7]. Furthermore, mental illnesses like depression, substance use disorders, and psychosis are some of the biggest risk factors for suicide [16], which is the second-leading cause of death for teens and young adults [25]. With professional treatment being less affordable and less accessible than ever before [47], internet interventions are emerging as a promising avenue for formal treatment [24] and informal peer support [40, 91].

Social media platforms are some of the most-visited websites on the internet, with some sites getting an estimated 1.7 billion visits every 6 months [84]. Researchers [73] suggested that youth benefit from online support exchanges on social media platforms; for example, actively engaging with peers online about mental health concerns is associated with an increased likelihood of seeking formal mental health care [73]. Unique platform affordances, like anonymity, can help users feel more comfortable being honest about their experiences and feedback [27]. Also, support exchange on social media can lead to increased quality of life, heightened perceived support, and increased self-efficacy [86].

2.2 AI-Powered Mental Health Support

AI systems are evolving quickly and have the potential to transform many fields, including mental health [17, 31]. Generative AI systems, like ChatGPT¹ and Google Bard², are machine learning models that can create human-like responses for accessible, affordable, and individualized support [83]. The time frame since these technologies became available is very short, but users have started to explore various ways in which ChatGPT can support mental health and archived their findings in non-peer-reviewed venues. For instance, researchers explored ChatGPT as a therapist assistant [29], where it can work as a companion for patients between therapy sessions, collecting information, listening to patients attentively, offering validation, and sharing potential coping strategies. Another study (again non-empirical, non-peer-reviewed) examined the possibilities of using ChatGPT for child mental health [45] with medical education for children, translation and assisting in provider-child patient communication, and access to medical information. [88] and [52] discuss similar possibilities for health conditions that can benefit from anonymity in accessing information [52].

Research about the outcomes of implementing AI systems for mental health is mixed. For example, an analysis of user review comments found that users had mostly positive experiences with AI chatbots for symptoms of anxiety and depression [3], and surveys of chatbot systems can suggest that they are useful reporting tools for in-person harms [74] and can even help to increase youth participation in societal activities [87]. However, Martinengo et al. examined dialogues with AI conversational agents (CAs) and found that CAs from app stores are *not* fit for comprehensive suicide risk management. According to their findings, this is because AI systems follow a predefined protocol for suicidal thoughts that is only activated in response to specific phrases that suggest immediate danger, meaning they cannot identify when a conversation is moving towards a problematic topic. Another issue is that CAs cannot send an unsolicited follow-up message after the conversation concludes to ensure the user is safe [62].

¹<https://chat.openai.com/>

²<https://bard.google.com/>

Researchers highlighted the importance of better understanding how AI can maximize its ability to support young people by assisting in conversations [80]. For instance, Sharma et al. [81] experimented with an AI agent that provides just-in-time feedback to peer support providers in an online mental health support platform and found that AI can help peers increase empathy in their support-giving messages. These findings are promising for the role of AI in peer-to-peer interactions, but the study did not focus on young people and their specific needs.

2.3 AI and Online Support for Young People

Researchers [22] have investigated many aspects of AI-enabled technologies, including personalization, ethical considerations, recalling information from previous sessions, and detecting body gestures and expressions [22]. At the same time, researchers highlight the lack of accuracy, reliability, bias, and unpredictability of outputs of these generative AI systems [61], calling for appropriate mitigation strategies, awareness, and careful consideration of these challenges before jumping into experimenting the possibilities of generative AI on sensitive topics, such as high-risk mental health conversations (e.g., suicide ideation).

The common shortcomings of these existing works include empirical testing of how end users might perceive these generative AI-powered systems that support mental health. While previous work explored people's perceptions of AI-based agents for mental health [2, 39], the verbal expressions and stylistic characteristics of more recent language-based AI technologies (e.g., LLMs) have more human-like characteristics than other AI agents to date [67], which requires further investigation. Given this technology is new and recent, researchers have only begun to understand the nuanced perceptions of how people over-trust, under-trust, or assess the evolving nature of AI agents' abilities. Additionally, there is very little research on how specific populations, such as youth, might react to these technologies.

The critical role that online social support plays for youth, the advent of new AI technologies that have great potential to support mental health, and the lack of understanding around young people's perceptions of more advanced AI agents place our study in a unique position to contribute to the gaps noted in the literature. Our study is the first to investigate young people's perceptions of how helpful, inappropriate, and problematic AI-generated responses are compared to multiple human-generated responses. These tests were conducted blinded, thus allowing us to objectively evaluate initial reactions to the style of the responses themselves rather than young people's preconceptions around AI.

In summary, young people often seek support online, especially for mental health concerns [8, 68]. That support can come from individuals with varying degrees of expertise and experience, and research has shown that these interactions are beneficial [68]. Now, AI systems are permeating digital conversations, but we do not know how young people feel about the content these systems produce. To address this gap, we studied how young people respond to AI-generated messages compared to the different types of support young people encounter online within the context of the topics that young people specifically seek support about.

3 METHODS

Using an online survey, we examined how young people react to different human- and AI-generated responses to help-seeking messages. In this section, we describe our participant criteria and recruitment process, explain our survey design, review ethical considerations, and describe our approach to data analysis.

3.1 Participant Recruitment Criteria and Process

Our target population was youth aged 18-24. While youth includes individuals aged 10-24 [66], as it is a time in which physiological and psychosocial development impact well-being in ways that drastically extend the transition period to

adulthood [78], we intentionally did not recruit minors. This is because of the sensitive nature of some topics in this study, and we were cautious to avoid unnecessarily exposing minors to topics like suicidal thoughts.

Based on a population of youth in the U.S. of 31 million [23], we determined we needed a sample of at least 385 participants for a 95% confidence level with a 5% margin of error and 664 participants for a 99% confidence level with a 5% margin of error. Our target was a minimum of 385, but we wanted to get as close to 664 as possible.

We recruited participants by sharing a digital flyer on Instagram, Facebook, X (formerly known as Twitter), Discord, and Slack. The flyer included a QR code linked to an online screening survey to determine prospective participants' eligibility, which required that they were 18 to 24 years old, currently living in the U.S., fluent in English, and have stable access to the Internet. Eligible respondents (N=1,096) continued to the IRB-approved study information sheet, which they were asked to review before continuing to the main survey. To ensure the quality of the responses, there was a built-in attention check question halfway through the main survey that asked participants to accurately indicate which help-seeking message topic we had previously asked about. After finishing the survey, participants could provide their contact information to receive \$5 as payment for their time. We received 866 survey responses from eligible participants, of which 622 were deemed fit for analysis (see table 1 for demographic information) because they were complete and passed the built-in attention check question.

3.2 Survey Design

We used a 4 x 4 study design with four help-seeking messages and four pre-written responses (henceforth referred to as "responses") to each help-seeking message (4 help-seeking messages x 4 responses = 16 total combinations). For each combination, we asked multiple questions. This section provides more details about the help-seeking messages, responses, and survey structure (see Figure 1).

3.2.1 Help-Seeking Messages and Pre-Written Responses. The help-seeking messages came from [44], in which the authors analyzed direct message conversations between youth and categorized youth private messages for support into four main themes: (1) relationships, (2) suicidal thoughts, (3) self-expression, and (4) physical health. To capture these themes in our survey, we used the full help-seeking messages that the excerpts in [44] came from.

The responses were written by four different sources: a (1) peer, (2) adult mentor, (3) therapist, and (4) Artificial Intelligence (AI). The **peer** responses in our survey, like the help-seeking messages, came directly from the dataset used in [44] and were the real responses youth sent after receiving the corresponding help-seeking messages from [44]. For the **adult mentor** responses, we wanted the message to reflect best practices for non-clinical crisis interventions. To do this, we identified an intensive, evidence-based training program through Crisis Text Line³, a global not-for-profit organization that provides mental health support via text. Given the existing research to support that this organization's practices are effective [38, 56], the second author used the training materials to generate the adult mentor's responses to capture the appropriate tone through deliberate phrasing. To capture evidence-based professional training, we had a licensed clinical psychologist who also runs an APA-accredited clinical training program write the **therapist** responses. We used responses generated by ChatGPT for the **AI** responses. To understand how AI and LLMs could advise and support young people, we prompted ChatGPT, a popular AI-powered chatbot used by many young adults [89], in May 2023 (GPT-3.5, free version) by asking it to write a response to each help-seeking message "in the appropriate style and of the appropriate length for a mobile text message" (see Appendix C for the full context of the prompts we used). We chose this prompt after exploration with ChatGPT and some prompt engineering. Initial attempts to generate a

³<https://www.crisistextline.org>

response by providing only the help-seeking message repeatedly generated responses in the form of a long bulleted list of advice. Compared to our earlier exploration, ChatGPT's response to our most current prompt was considerably shorter and involved more interaction with the help seeker, which was suitable for a text message. The full text of the help-seeking messages and responses are in appendices A and B, respectively.

Participants were blinded about where the responses came from because we wanted to understand perceptions of the content as part of peer support exchange, not feelings about different sources of support engaging in the exchange. We also randomized both the order in which the scenarios were presented and the individual questions about each response to prevent question order bias.

3.2.2 Survey Structure. At the beginning of the survey, we asked participants how often they seek and give support online. We used two multiple-choice questions, one for seeking support and one for giving support, to which the answer choices were never, rarely, sometimes, often, or always.

Next, we presented the sixteen help-seeking message x response combinations in four sections; each section contained one help-seeking message and its four corresponding responses. Each section had the same structure. First, we instructed participants to imagine they received the help-seeking message from a friend as a text or direct message on social media. Then, we asked participants to rate the extent to which they agreed with the following statements on a 5-point Likert scale (1 = not at all, 5 = extremely): "I like this response," "I would send this response," "This response contains helpful advice," "This response is problematic," and "This response is inappropriate." At the end of each section, we asked participants to type a brief text response about why they did and did not like certain combinations.

The rationale in using the Likert scale response options was to measure the direction and intensity of the respondent's attitude and preference toward the specific message. If attitude and preference are measured in a binary fashion (i.e., yes or no), then the data is not robust enough to capture the intensity of the attitude and preference. Because we asked respondents about their attitude and preference for each help-seeking message x response combination, it was important to capture the relative intensity of their attitude and preference for each message to allow for nuance within the data, resulting in a more accurate and representative understanding of feedback. Seeing as a limitation of Likert scale responses is not understanding the "why" behind an attitude or preference, we then followed up with the qualitative question.

An attitude or a preference is a psychological construct. In our study, the attitude or preference we measured represent the survey respondent's predisposition to respond favorably or unfavorably to the scenario (i.e., "I like this response"). Attitudes and preferences are often - although not always - considered precursors to behavior. Numerous studies have demonstrated that individuals often express their attitudes but do not always act in accordance with those attitudes ([9]). In our study, we used several additional survey questions to examine whether there was consistency with what the respondents liked and what they found helpful (i.e., "I would send this response" and "This response contains helpful advice"). There is a compelling body of research in social psychology that suggests there are times of conflict between what people want (immediate gratification, for example) and what they believe they should do or what they find most helpful (longer-term benefits; [14]).

We included opposing statements for the Likert scale questions for multiple reasons. First, we know from psychology literature [26] that people may not always seek help or solutions when they discuss problems or concerns; instead, they may primarily seek emotional support, active listening, or validation. Therefore, these statements were designed to capture those preferences and validate the cohesiveness in survey responses (i.e., it is unlikely that a participant finds a response both helpful and inappropriate). We also measured opposing concepts because literature in survey design

suggests that participants can present acquiescence bias, which is the tendency to agree with statements regardless of the content. Also, social desirability bias would make participants inclined to answer in ways they believe are more socially acceptable or favorable. Accordingly, asking about opposing concepts to force participants to think more critically about their answers would increase the quality of survey responses [54].

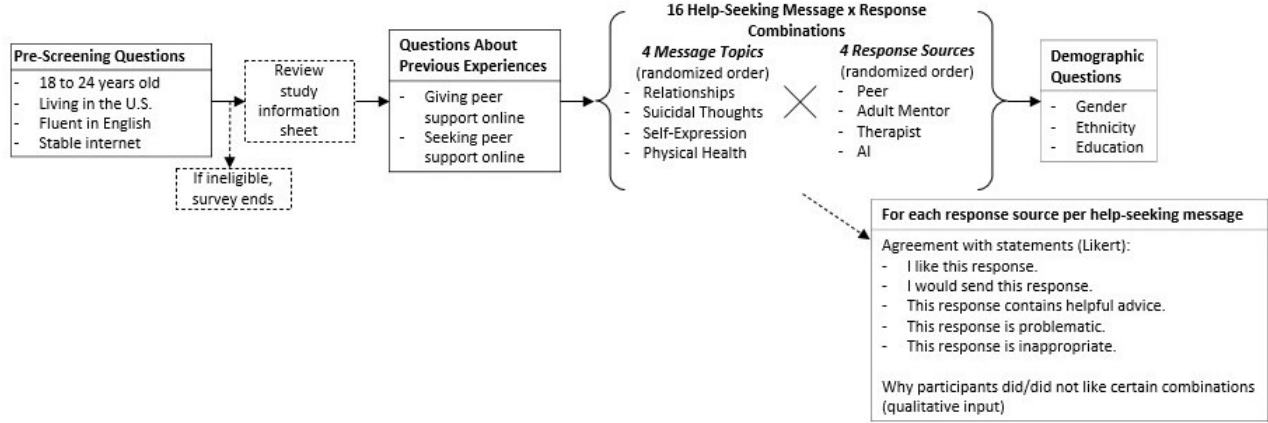


Fig. 1. Visualization of the survey structure described in section 3.2.2

We also asked for our participants' demographic information (Table1). See Figure 1 for a visualization of the survey structure.

3.3 Ethical Considerations

Ethical research conduct is of paramount importance, particularly with regard to sensitive topics. This study was submitted for full-review to Drexel University's IRB and was approved as an exempt study. This meant participants were presented with the study information sheet but not required to sign a consent document. Despite our study's exempt status, we took multiple precautions to address potential risks and maintain emotional safety for participants, especially due to the sensitive nature of our survey materials. We specifically and intentionally included a board-certified clinical psychologist on our research team to carefully review all language in the survey questions and AI-generated responses to minimize potential risks and monitor incoming data for concerning submissions. We also presented all participants with a study information sheet that contained the same information as a full consent document, including detailed descriptions of the research objective, the survey procedures involved, the voluntary nature of the study, the potential risks involved, the anonymization process, and incentives. In lieu of a consent signature, all participants had to acknowledge they read this document before beginning any study procedures.

To protect participants' anonymity and confidentiality, only the research team members specified in the IRB protocol were granted access to the data. The researchers stored and accessed all data on a university-approved server. This study took place entirely online, meaning participants could complete the survey in a private location of their choice. When participants finished their participation, they were each assigned a unique identifier, known only to the researchers on the IRB protocol, that we used to identify the participants instead of their names. The PIs maintained an Excel spreadsheet containing participants' contact information, which we did not use in our data analysis or any publication resulting from this research.

<i>Variable</i>	<i>Count (N)</i>	<i>Percent of total (%)</i>
All Responses	622	100.00%
Gender		
Male	403	64.79%
Female	216	34.73%
Non-binary	2	0.32%
Prefer to self-identify	1	0.16%
Prefer not to say	0	0.00%
Total	622	100.00%
Ethnicity		
Caucasian	295	47.43%
African American	193	31.03%
Native American	34	5.47%
Latino or Hispanic	26	4.18%
Mixed Race	24	3.86%
Asian	17	2.73%
Multiple	16	2.57%
Prefer not to say	9	1.45%
Native Hawaiian or Pacific Islander	6	0.96%
Other	2	0.32%
Total	622	100.00%
Education		
Some high school	17	2.73%
High school or equivalent	147	23.63%
Trade school	100	16.08%
Bachelor's degree	297	47.75%
Master's degree	39	6.27%
Doctoral degree	12	1.93%
Other	3	0.48%
Prefer not to say	7	1.13%
Total	622	100.00%

Table 1. Participant demographics

3.4 Data Analysis

To ensure we collected high-quality data, we first looked at the correlations between how much participants 'liked' the responses, felt they 'would send' the message to the help-seeker, and how 'helpful' they thought it was compared to how 'problematic' and 'inappropriate.' We did this because it is possible for participants to prefer a response for paradoxical reasons, such as finding the content comically inappropriate - fortunately, this was not the case in our data. Using means across the opposing attributes, we confirmed that the responses from sources that participants preferred more, said they were more likely to send, and felt were more helpful were also perceived as less inappropriate and less problematic. Participants also tended to feel responses were more helpful than inappropriate or problematic, which means we avoided doing more harm than good, as is essential for ethical research.

We also wanted to understand how participants reacted to the different responses beyond which one they 'liked' the most (i.e., preferred). To do this, we looked at aggregate responses for all help-seeking messages by the source to understand how helpful the responses were ('helpful'), how likely they were to send them ('would send'), and how inappropriate ('inappropriate') and problematic ('problematic') they felt the responses were. In subsequent sections and figures, we refer to these four measures as 'attributes' of participants' responses. Furthermore, the language we use to

describe the findings for the 'would send' attribute includes the phrase "more likely." This refers to the extent to which participants agreed they would send the response in question, not that we performed probability or likelihood analysis as that was not part of our study methodology.

We analyzed the quantitative data using R software. Our group size (N) for all statistical tests was equal to our total number of participants, which was 622. Before comparing means across combinations, we used Bartlett's test of homogeneity to determine whether variance was equal and normality of the data distribution. When Bartlett's test was significant, indicating that variance is not equal across variables, we used Welch one-way tests for nonparametric data to compare means. When Bartlett's test was not significant, meaning variance is equal across variables, we used a one-way ANOVA. We report the p -value for Bartlett's test before all F -statistics to signal which test we used.

We analyzed the qualitative responses using reflexive thematic analysis [18] because it allows appropriate flexibility for inductive, iterative analysis to draw rich conclusions. We took an inductive approach to ensure conceptual coherence because our goal was to identify and develop themes directed by the content of our data, not by existing concepts or ideas. Rather than focusing on specific questions, our objective was to complement our quantitative findings and further explore reasons why participants did, or did not, like specific responses to help-seeking messages about relationships, suicidal thoughts, self-expression, and physical health. Since we conducted this qualitative analysis after analyzing participants' quantitative data, our analytical framework was situated within an essentialist perspective; in other words, we presupposed that participants' open-ended responses reflected their opinions expressed in their previous quantitative survey responses.

To start, the second and third authors familiarized themselves with the data by reviewing all qualitative responses and taking notes along the way. They used those notes to identify initial themes for discussion and to share with the rest of the research team. The research team met multiple times to discuss and iterate over the themes and relevant examples to produce the findings in the results section of this paper. We did not use a traditional codebook because the qualitative analyses complement our quantitative findings rather than act as a stand-alone analysis. The first, second, and third authors identified the examples used in our findings section to illustrate our main points.

4 RESULTS

Next, we describe participants' preferences for and reactions to different support-giving responses (peer, adult mentor, therapist, and AI) across all help-seeking topics and how these preferences and reactions vary depending on the topic of the help-seeking message they are responding to (relationships, suicidal thoughts, self-expression, or physical health).

Almost all of our participants (97%) reported at least some previous experience seeking and giving support online (i.e., they did not select "never"). The relative frequency of responses was the same for both seeking and giving support. The most frequent response was "sometimes" (seeking: $n = 277$, 45%; giving: $n = 238$, 38%), second was "often" (seeking: $n = 148$, 24%; giving: $n = 174$, 28%), third was "rarely" (seeking: $n = 113$, 18%; giving: $n = 137$, 22%), fourth was "always" (seeking: $n = 65$, 10%; giving: $n = 57$, 9%), and least frequent was "never" (seeking: $n = 19$, 3%; giving: $n = 16$, 3%).

4.1 Young People's Preferences Regarding Human and AI-generated Help-Giving Responses (RQ1)

We found that the human-generated adult-mentor response performed best when we combined responses across all help-seeking topics and compared the average responses from the four sources (peer, adult mentor, therapist, and AI). How much participants liked each source's responses were significantly different (Bartlett's test: $p < 0.05$; $F(3.0,5517.7) = 99.20$, $p < 0.001$) across all help-seeking messages. There were also significant differences in the extent to which participants found the responses helpful (Bartlett's test: $p < 0.05$; $F(3.0,5519.6) = 104.37$, $p < 0.001$), problematic (Bartlett's

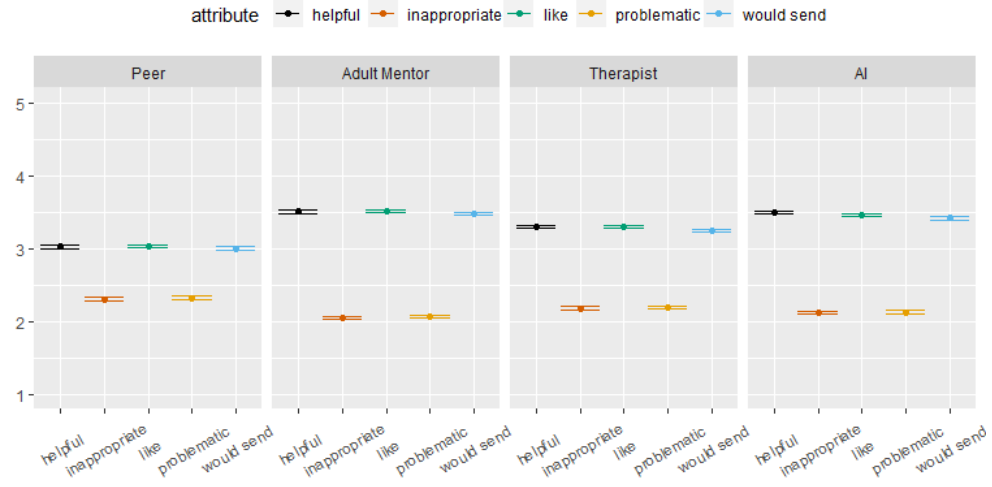


Fig. 2. Participant preferences for human and AI-generated responses by source (RQ1). The y-axis is how much participants agreed with the different attributes of their reaction (i.e., how strongly they agreed with the statements described in section 3.2.2). Points represent means, and the bars on either end are standard deviation.

test: $p < 0.05$; $F(3.0,5525.6) = 20.159$, $p < 0.001$), inappropriate (Bartlett's test: $p > 0.05$; $F(3.0,9948) = 19.452$, $p < 0.001$), and how likely they were to send the response (Bartlett's test: $p < 0.05$; $F(3.0,5513.9) = 91.09$, $p < 0.001$).

The post-hoc analysis showed that participants liked the adult mentor's responses the most ($M=3.51$, $SD=0.02$), followed by AI ($M=3.46$, $SD=0.02$), then the therapist's ($M=3.29$, $SD=0.02$), and the peer responses were their least favorite ($M=3.03$, $SD=0.02$). This means that before considering the topic of the help-seeking message, it seemed like humans, specifically adult mentors, outperformed AI in generating help-giving responses, although peer responses were not as liked by the participants. However, such responses varied after considering help-seeking topics, which we will discuss.

Participants felt the adult mentor's responses were the most helpful ($M=3.50$, $SD=0.99$) of the responses presented and also indicated that they were most likely to send the adult mentor's responses to the help-seeker ($M=3.48$, $SD=0.98$), followed by AI's responses (how helpful: $M=3.49$, $SD=1.05$; would send: $M=3.42$, $SD=1.12$), then the therapist's responses (how helpful: $M=3.29$, $SD=1.07$; would send: $M=3.25$, $SD=1.12$), and the peer's responses receiving the lowest score for those preferences (how helpful: $M=3.02$, $SD=1.15$; would send: $M=3.00$, $SD=1.18$).

How problematic and inappropriate participants felt the responses were, depending on the source of the response, followed the opposite pattern. The peer responses were perceived to be the most problematic and inappropriate (how problematic: $M=2.32$, $SD=1.24$; how inappropriate: $M=2.30$, $SD = 1.24$), followed by the therapist's responses (how problematic: $M=2.19$, $SD=1.22$; how inappropriate: $M=2.18$, $SD=1.23$), AI's responses (how problematic: $M=2.13$, $SD=0.02$; how inappropriate: $M=2.12$, $SD=0.02$), and the adult mentor's responses as the least problematic ($M=2.07$, $SD = 0.02$) and inappropriate ($M=2.05$, $SD = 0.02$). See Figure 2 for a visualization of means and standard deviations for participants' reactions to the responses by source.

The findings so far suggest that participants regarded AI-generated responses highly, above even therapist and peer responses, although adult mentors' responses were considered most helpful and most likely to be sent in response to the help-seeking messages.

4.2 Young People's Preferences for Help-Giving Responses Depending on the Topic of the Initial Help-Seeking Message and Rationale Behind Them (RQ2)

4.2.1 Relationships. How much participants liked each source's responses to the help-seeking message about relationships were significantly different (Bartlett's test: $p < 0.05$; $F(3.0,1378) = 39.145$, $p < 0.001$) as were the extent to which participants found the responses helpful (Bartlett's test: $p < 0.05$; $F(3.0,1377.4) = 33.72$, $p < 0.001$), problematic (Bartlett's test: $p < 0.05$; $F(3.0,1379.0) = 14.47$, $p < 0.001$), inappropriate (Bartlett's test: $p > 0.05$; $F(3.0,2484) = 12.113$, $p < 0.001$), and how likely they were to send the response (Bartlett's test: $p < 0.05$; $F(3.0,1375.6) = 43.24$, $p < 0.001$). Participants preferred the response about relationships that AI-generated more than any of the human-generated options. They liked AI's response the most ($M=3.63$, $SD=0.04$) and thought it was the most helpful ($M=3.63$, $SD=1.00$), the least problematic ($M=2.01$, $SD=1.16$) or inappropriate ($M=2.01$, $SD=1.20$), and said they were most likely to send the AI's response to the help-seeker ($M=3.65$, $SD=0.96$). On the other hand, participants did not like the peer's response about relationships ($M=3.00$, $SD=1.15$) and felt it was the most problematic ($M=2.41$, $SD=1.24$) and inappropriate ($M=2.38$, $SD=1.24$), and least helpful ($M=3.07$, $SD=1.13$); consequently, they were also least likely to send the peer's response ($M=2.98$, $SD=1.17$). See Figure 3 for a visual comparison of the means and standard deviations for reactions to all responses about relationships.

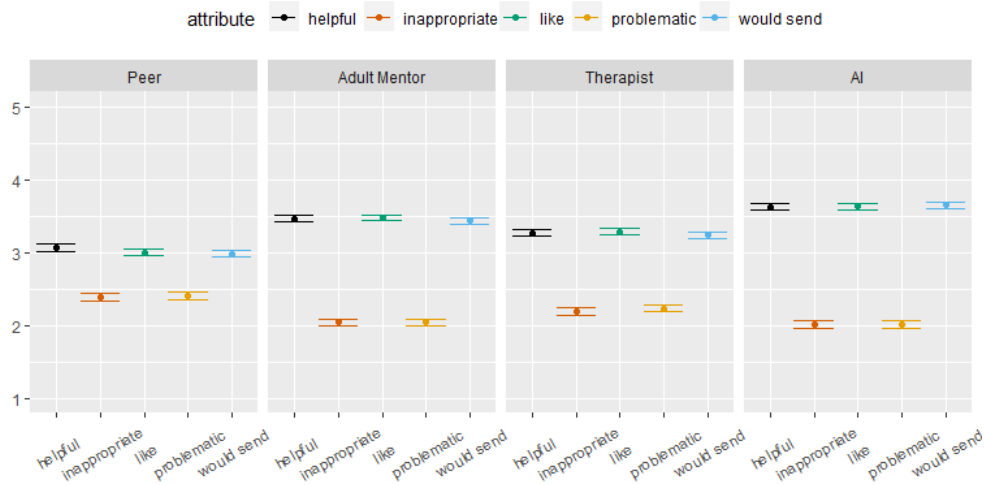


Fig. 3. Participant preferences for human and AI-generated responses by source for help-seeking message about relationships (RQ2, relationships). The y-axis is how much participants agreed with the different attributes of their reaction (i.e., how strongly they agreed with the statements described in section 3.2.2). Points represent means, and the bars on either end are standard deviation.

This sentiment was reflected in open-ended questions, where participants shared they liked the balance and empathy in the **AI-generated** responses. They felt the AI-generated response invited the help-seeker to share additional information without being overly intrusive or demanding. As one participant put it, the response was “*helpful... [and] also supportive and feels like a safe space.*” (P453, Female). Some participants described anthropomorphic reasons, such

as the AI-generated response seemed like it came from someone who is a good friend: *“They are really trying to help the friend so much and we all need such friends.”* (P434, Female). There were also mentions of qualities participants associated with the AI response, including that they thought it was *“[empathetic]...supportive and caring.”* (P46, Male). Even though participants did favor the AI-generated response to the help-seeking message about relationships, they did note some room for improvement. Namely, the response doesn’t consider the entire context of the situation and *“accuses the ex without listening to his side of the story.”* (P118, Male).

Participants noted that the **adult mentor’s responses** also provided acknowledgment and validation, but it was potentially less useful and lacked clear direction. Those who liked this response noted that *“it is respectful and...made [them] feel like the person understood [the help-seeker’s] feelings.”* (P50, Male). Participants also noted a sense of maturity in how the response did not judge the help-seeker: *“...the [response] answered the question in a mature way by making the victim not feel offended by going back to the ex. The response is just showing him or her that is what we [humans] go through in terms of love life.”* (P403, Male). Despite these positive attributes of the adult mentor’s response, some participants felt it lacked concrete advice and did not give specific suggestions for what to do next. As mentioned in the previous paragraph, participants liked that the AI-generated response guided the help-seeker to provide additional detail in a gentle way - they did not feel the adult mentor’s response had this same quality. Instead, participants felt the adult mentor’s response was, as one participant said, *“beating around the bush”* (P341, Male) and *“didn’t provide a way of how to solve the situation clearly.”*

Participants liked the **therapist’s response** because they felt it was extremely understanding and non-judgmental but noted that it seemed to delay giving actionable advice. Even though they did not know a therapist wrote one of the responses, they picked up on the professional tone: *“This person provides professional remedy and ... [asks] the victim to share their mixed emotions and from there she/he will be in a position to offer strategies.”* (P481, Female). They saw the response as *“very empathetic without shaming the friend”* (P587, Female), but noted that it felt like *“scheduling [giving] help to next time when the [help-seeker] is in need of it now.”* (P503, Female). They appreciated the supportive tone but would have liked more concrete and immediate advice rather than beating around the bush and delaying support.

We received the most negative comments about the **peer’s response** because participants felt it was terse, suggested hasty actions, and was not encouraging. While some appreciated the response’s brevity, the way it was written came across as *“too direct with no empathy.”* (P163, Male). This lack of empathy made participants feel like the sender had a bad attitude and was being too tough on the help-seeker: *“It sounds harsh and rude, and also it doesn’t consider the feelings of the victim.”* (P452, Female).

4.2.2 Suicidal Thoughts. How much participants liked each source’s responses to the help-seeking message about suicidal thoughts were significantly different (Bartlett’s test: $p < 0.05$; $F(3.0,1375.3) = 64.08$, $p < 0.001$) as were the extent to which participants found the responses helpful (Bartlett’s test: $p < 0.05$; $F(3.0,1375.6) = 59.14$, $p < 0.001$), problematic (Bartlett’s test: $p > 0.05$; $F(3.0,2484.0) = 20.40$, $p < 0.001$), inappropriate (Bartlett’s test: $p > 0.05$; $F(3.0,2484) = 23.46$, $p < 0.001$), and how likely they were to send the response (Bartlett’s test: $p < 0.05$; $F(3.0,1372.9) = 73.75$, $p < 0.001$).

The help-seeking topic of suicidal thoughts was the only one for which all the human-generated responses (especially the one from the adult mentor) outperformed AI’s response. Participants liked the adult mentor’s response the most ($M=3.59$, $SD=0.95$) and felt it was the most helpful ($M=3.61$, $SD=0.94$), least problematic ($M=2.04$, $SD=1.18$), least inappropriate ($M=2.00$, $SD=1.20$), and that they were most likely to send it to the help-seeker ($M=3.56$, $SD=0.94$). They also least liked AI’s response ($M=2.82$, $SD=1.19$) and saw it as the most problematic ($M=2.53$, $SD=1.26$), most

inappropriate ($M=2.54$, $SD=1.27$), least helpful ($M=2.93$, $SD=1.12$) and identified it as the one they were least likely to send to the help-seeker ($M=2.71$, $SD=1.21$). See Figure 4.

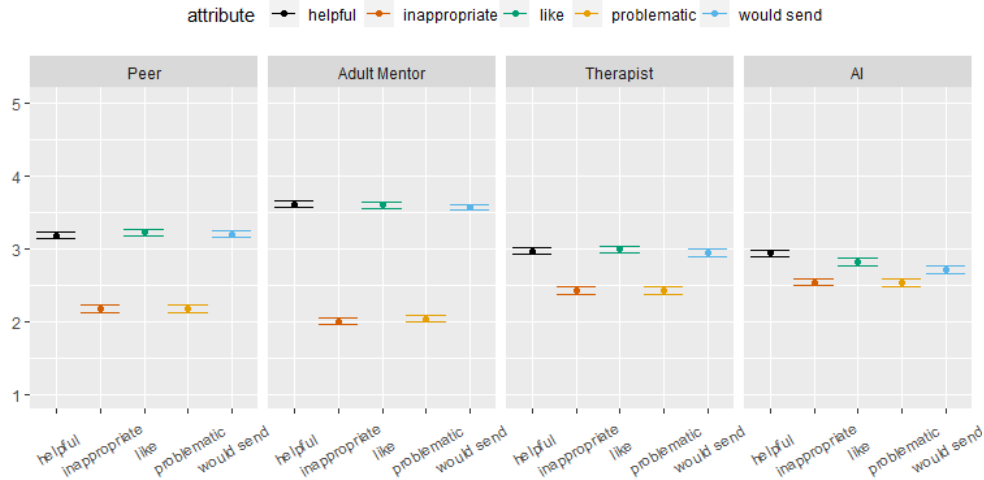


Fig. 4. Participant preferences for human and AI-generated responses by source for help-seeking message about suicidal thoughts (RQ2, suicidal thoughts). The y-axis is how much participants agreed with the different attributes of their reaction (i.e., how strongly they agreed with the statements described in section 3.2.2). Points represent means, and the bars on either end are standard deviation.

When describing why they preferred the **adult mentor's response** to this sensitive topic, participants noted that they liked how it conveyed genuine concern by inviting the help-seeker to share more details and put effort into providing more support. *"There's a feeling of relating with the individual here. There's a feeling of care and [that they're] really interested and [want] to understand the individual [and their] situation."* (P124, Male). The only area of improvement that participants mentioned for the adult mentor response is that the message could have been a little shorter, but did not note what content should be cut.

The **peer's** response was the second most preferred, and participants felt it was superficial, inconsiderate, and did not sufficiently acknowledge the help-seeker's struggles. Interestingly, participants likened the peer response to an AI-generated response, saying it sounded *"...typical of a bot"* (P478, Female). Unlike the adult mentor's response, participants did not like how the peer's response was *"so shallow [and] does not give the victim...much [detail] on how to tackle the situation and [get] helped."* (P403, Male).

Participants noted that the **therapist's response** immediately asked for more information in an attempt to ensure the help-seeker's safety - not knowing that this message was written by a clinician, some thought this sounded *"professional"* while many others took it as intrusively asking for more information right away instead of offering support. *"I think this response makes the person in need sink deeper into her trouble rather than pull her out. It's a bit inappropriate."* (P451, Female). Since the initial help-seeking message did not use the word "suicide," many participants also did not like that the therapist's response immediately drew that conclusion and felt that calling it out could come across as judgemental: *"I hate the assumption about self harm and suicide. It feels really I'm going to insert myself into your life to fix it and doesn't feel friendly at all."* (P621, Non-binary).

The **AI-generated** response overall performed the worst for the help-seeking topic about suicidal thoughts. While some appreciated that the response was “*actually truthful about the person’s ability*” (P52, Male), most thought the referral to a healthcare provider conveyed less desire to help in that moment. “*Fine, you can’t provide help but that’s not the best way to put it. You could always have the right choice of words and making the right sentence in this situation before suggesting a mental health professional and not everybody will like the fact that they are referring to them as going crazy. Not really cool though*” (P124, Male).

4.2.3 Self-Expression. How much participants liked each source’s responses to the help-seeking message about self-expression were significantly different (Bartlett’s test: $p < 0.05$; $F(3.0,1378.4) = 37.62$, $p < 0.001$) as were the extent to which participants found the responses helpful (Bartlett’s test: $p < 0.05$; $F(3.0,1378.4) = 39.53$, $p < 0.001$), problematic (Bartlett’s test: $p > 0.05$; $F(3.0,2484.0) = 4.88$, $p < 0.001$), inappropriate (Bartlett’s test: $p > 0.05$; $F(3.0,2484) = 5.56$, $p < 0.001$), and how likely they were to send the response (Bartlett’s test: $p < 0.05$; $F(3.0,1377.4) = 34.60$, $p < 0.001$).

Participants’ reactions to the responses about self-expression followed a similar pattern to the responses about relationships, meaning AI’s response outperformed all the human-generated responses, especially the peer’s response. We found that participants liked AI’s response the most ($M=3.70$, $SD=0.98$), felt it was the most helpful ($M=3.75$, $SD=0.94$), said they were most likely to send it to the help-seeker ($M=3.67$, $SD=1.00$), and thought it was the least problematic ($M=2.01$, $SD=1.22$) and least inappropriate ($M=1.99$, $SD=1.18$). Participants also liked the peer’s response the least ($M=3.08$, $SD=0.04$), said they were least likely to send it ($M=3.05$, $SD=1.15$), and thought it was the least helpful ($M=3.12$, $SD=1.09$), the most problematic ($M=2.25$, $SD=1.19$) and the most inappropriate ($M=2.26$, $SD=1.23$). See Figure 5 for a visual comparison of the means and standard deviations for reactions to all responses about self-expression.

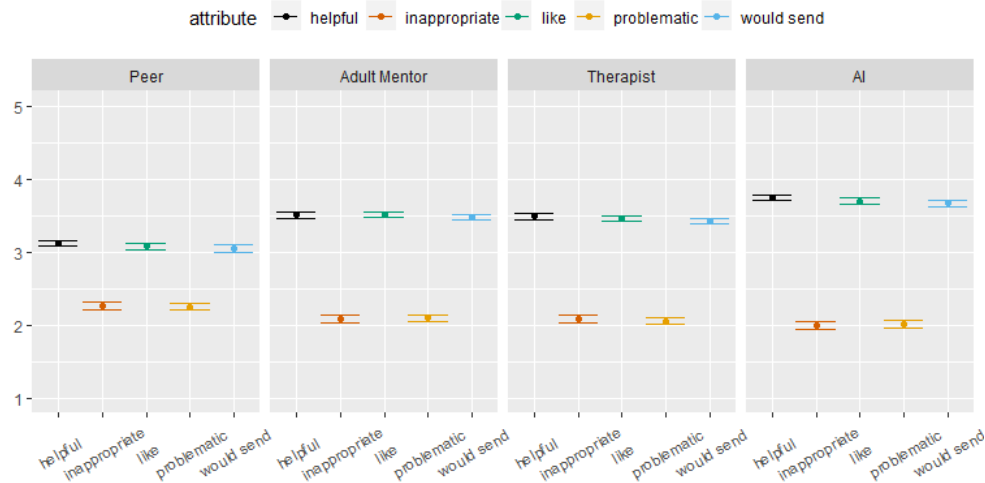


Fig. 5. Participant preferences for human and AI-generated responses by source for help-seeking message about self-expression (RQ2, self-expression). The y-axis is how much participants agreed with the different attributes of their reaction (i.e., how strongly they agreed with the statements described in section 3.2.2). Points represent means, and the bars on either end are standard deviation.

Participants liked that the **AI-generated response** conveyed guidance, support, and understanding. “*I think a response like this can make the person feel like they’ve all been in this situation before, and there are points of empathy that they can draw on to guide them to be braver by drawing on the respondent’s situation.*” (P43, Male). Similar to the message

about relationships, participants liked that this response “[*empathizes*] with the friend while also providing practical solutions” (P587, Female).

They felt the response from the **adult mentor** also conveyed some affirmation of the help-seeker’s struggles but to a lesser extent. They liked that “*it starts with sympathy*” (P160, Male), but the overall tone felt “*unattached*.” This distinction between sympathy and empathy is important to note - as sympathy is feeling for someone else while empathy is feeling with them. The adult mentor’s response conveyed “sympathy,” meaning they can see where the help-seeker is coming from, but *not empathy*, which implies less personal experience with the matter at hand. Participants clearly communicated this sentiment by saying that “*the victim is not understood*” (P383, Male) and that the message indicates “*the person [who wrote it] has no idea of what the [help-seeker] is going through*.” (P611, Female).

Similar to the reactions about responses to the help-seeking message about relationships, participants suspected that the **therapist’s** message came from a clinician, “*The message looks like one a professional therapist would send...*” (P350, Male). As one participant said, “*The response is appropriate because [it] is from [a] counselor, someone who has been able to tackle much of the issues by being exposed to different clients and knowing what to do for certain client till they reach a point of believing in themselves and having that courage to talk to anyone without any fear*.” (P403, Male).

The **peer’s response** did not feel supportive enough to participants and did not do enough to calm the help-seeker, “*I think that’s too casual a response to really consider the issue from the perspective of the person pouring it out*.” (P21, Male). Moreover, they felt this response lacked depth and oversimplified the help-seeker’s fear of self-expression. Like the adult mentor response, participants also felt there was not enough of an indication that the person who wrote the response truly understands what the help-seeker is going through, but less so because they can’t empathize and more because they do not want to put effort into conveying that empathy, “*It was too brief and nothing more of empathy felt in it. It was more like this person wanted to help but didn’t really put much effort... [it had] a little spark of encouragement[,] but it could be better*.” (P124, Male).

4.2.4 Physical Health. How much participants liked each source’s responses to the help-seeking message about physical health were significantly different (Bartlett’s test: $p < 0.05$; $F(3.0,1376.3) = 71.06$, $p < 0.001$) as were the extent to which participants found the responses helpful (Bartlett’s test: $p < 0.05$; $F(3.0,1376.1) = 77.94$, $p < 0.001$), problematic (Bartlett’s test: $p > 0.05$; $F(3.0,2484.0) = 20.76$, $p < 0.001$), inappropriate (Bartlett’s test: $p > 0.05$; $F(3.0,2484) = 17.35$, $p < 0.001$), and how likely they were to send the response (Bartlett’s test: $p < 0.05$; $F(3.0,1375.8) = 63.92$, $p < 0.001$).

When the help-seeking message was about physical health, the AI responses once again outperformed the human-generated responses, especially the peer’s. Participants liked AI’s response the most ($M=3.71$, $SD=0.96$), saw it as the most helpful ($M=3.67$, $SD=0.99$), least problematic ($M=1.95$, $SD=1.16$), least inappropriate ($M=1.95$, $SD=1.21$), and overall felt they were most likely to send it to the help-seeker ($M=3.64$, $SD=0.99$). The peer response received the least favorable reaction as the least liked response (2.83 , 0.05) and thought it was the most problematic ($M=2.46$, $SD=1.28$), most inappropriate ($M=2.46$, $SD=1.28$), and least helpful ($M=2.73$, $SD=1.22$), and indicated they were least likely to send it to the help-seeker ($M=2.78$, $SD=1.23$). See Figure 6 for a visual comparison of the means and standard deviations for reactions to all responses about physical health.

The **AI-generated** response once again performed the best, and participants raved about how much they liked it - as one participant said, “*I didn’t just like this response. I love this response*.” (P351, Male). Participants describe how the answer contained all the important components: empathy, acknowledgment, and concrete suggestions for next steps, “*its an all round genuine answer [that] shows you care[,] understand the problem and offer the right solutions*.” (P592, Female). The “right solution,” in this case, was to continue with ongoing self-care work (i.e., talking to a counselor), and

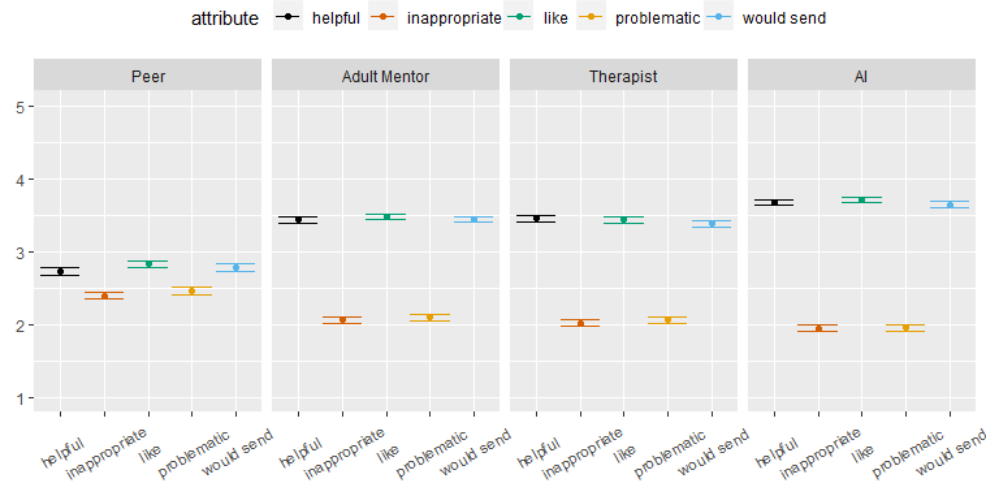


Fig. 6. Participant preferences for human and AI-generated responses by source for help-seeking message about physical health (RQ2, physical health). The y-axis is how much participants agreed with the different attributes of their reaction (i.e., how strongly they agreed with the statements described in section 3.2.2). Points represent means, and the bars on either end are standard deviation.

participants liked that this response acknowledged doing so is not easy. “*The fact that he empathized with the person and encouraged them to further try [counseling] because you cannot get the intended results after one trial.*” (P80, Male). Participants also liked that counseling was just one of the multiple solutions the response discussed and “*shows the [help-seeker] how they can use other strategies to be okay.*” (P453, Female). The comprehensive and deeply empathetic nature of this response seems to be what set it apart from the others.

Participants did like that the **adult mentor**’s response “*[validated] wanting rest*” (P621, Non-binary), but also felt that it “*[made] the [help-seeker] feel as if it’s OK to just relax and not do anything.*” (P351, Male). They discussed how the adult mentor’s message did not acknowledge that self-care is more than relaxation. In doing so, this response seemed to convey that there is nothing more the help-seeker can do, “*I feel like it’s not addressing the issue at hand and makes the victim feel like they are doing great of which they are not.*” (P137, Male). While participants thought “*praising the [help-seeker] for small achievement[s]*” (P614, Female) and “*positive energy*” were good, they were not enough, “*because it doesn’t contain enough [advice] that can help the victimized person*” (P125, Male).

While the **therapist**’s response did offer a solution, participants did not think it was the right one: “*It’s very unclear whether this respondent wants to help out. They are telling the victim to keep pursuing the same course that ain’t working for them.*” (P167, Male). Unlike the AI-generated response, which suggested continued counseling in addition other solutions, participants did not like that the therapist’s response was “*just telling the person in question to [continue] counselling, [with] no other options to explore.*” (P122, Male). Therapy also incurs a cost for the help-seeker, which participants were not shy to mention in their feedback for this response. They noted that “*counselling can be [quite] expensive,*” (P166, Male) and increasing those costs might not be necessary because “*at this point [the help-seeker] needs love and support from family and friends who would be available to support them.*” (P438, Female). This does not mean participants wanted the response to suggest the help-seeker stop counseling altogether, but rather that there should be additional non-clinical suggestions.

In summary, our findings revealed that participants preferred AI-generated responses to help-seeking messages about relationships, self-expression, and physical health. However, when the help-seeking message was about suicidal thoughts, participants preferred responses from the adult mentor. In the next section, we discuss the implications of these findings.

5 DISCUSSION

Our findings point to several important areas for discussion regarding the promise of AI-powered support and questions to address about handling sensitive topics, should they arise.

5.1 Premise for AI-Powered Support

Our results provide insight about AI's potential to respond to help-seeking topics frequently represented in youth's conversations, which previous literature stated that they sometimes left without supportive responses. For instance, a study of private messages shared among youth online discovered norms of '*unsupport*' when trust is hampered or when there is lack of resources or understanding of the situation by the youth peers [44]. Rather than considering AI as a replacement for humans or directly giving support to the help-seeking individual, AI can play many roles in supporting the social support being exchanged among peers. Some examples include providing help-giving individuals tools to understand the intent of help-seeking individuals, ways to respond to the individuals in supportive ways, and how the responses can be stylistically modified to give better support. Our findings about participants' positive reactions to the AI-generated responses provide a premise for using generative AI to assist help-giving individuals in support-giving conversations, as long as the AI response or nudge can maintain the characteristics that participants identified as helpful, empathetic, constructive, and non-judgemental. However, AI is a constantly moving target, where its characteristics, functionalities, and abilities constantly evolve over time. For instance, for Open AI's GPT models, the level of memorizing past conversations has been limited, but recent updates enabled the system to remember user preferences and tailor responses accordingly [65]. Furthermore, depending on the version of GPT models, the prompt engineering techniques can generate different results. Accordingly, the lesson we should take away from this study is to apply the rationale provided by the participants and devise ways that future generative AIs can maintain the preferred styles through customization, personalization, prompt engineering, using API wrappers to workaround shortcomings (e.g., [85]), and manipulation of training data. Doing so can help with improving the model's performances and transparency that leads to necessary levels of trust [21] and human's reliance on AI systems [51].

Some remaining questions include if participants knew it was AI, would they still be as positive as the results have shown? This question can be addressed from a future study, building on the efforts to improve AI-literacy [55] on ensuring there is no over- or under-trust of these systems. It seemed as though the participants felt genuine empathy from the AI responses, which was often missing in human-generated responses. If they find out that such genuine responses were artificially generated, it might or might not affect the trust in the system's responses. Our ongoing interview study with a subsample of these participants will address this issue.

5.2 AI as Complimentary Response based on the Context and Sensitivity of Topics

Before considering the context of the help-seeking message, it would appear that young people prefer the adult mentor's response. However, a closer look reveals a strong preference for the adult mentor's response to the sensitive help-seeking message about suicidal thoughts while they preferred the AI-generated response to all other help-seeking messages. In a research article by Shao [80], a similar encounter takes place where their results indicate human-operated counselors

provide higher quality results during empathetic counseling sessions than AI, but upon further investigation, it was clear that the way in which AI-supported an individual became crucial to its success in being rated helpful. As indicated by our findings and as previous research [4, 76] has uncovered participants usually preferred responses that offered support and comfort within the scope of positivity to feel comfortable enough to continue to share.

We do not intend for our results to imply that AI-generated content is sufficient to replace a therapist or clinical expertise. As we found, participants did not like AI's response about sensitive topics, for which ChatGPT's response specified that the kind of support required was outside the model's capabilities. While some participants appreciated the honest answer that reflected this limitation, others noted that referral to someone else felt dismissive because the help-seeker could have gone to that other source to begin with, but they did not. This points to two interesting follow-up questions: 1. if young people are experiencing suicidal thoughts, would they go to an AI system for help in a real-world situation? 2. if the answer to question 1 is yes, then how should an AI system respond when a user asks for help it cannot provide for sensitive matters? Researchers [59] have begun to explore how AI should own up to its faults during user interactions, such as when it may provide incorrect information. Our work highlights that similar frameworks should be explored for sensitive topics – for example, is it more helpful for the AI to respond with links and contact information for resources? If so, which resources are most helpful to share and why? The goal is for AI to convey its limitations constructively, ensuring the user is safe.

The highly sensitive and complex issue of suicidal thoughts demands careful consideration of ethical, legal, and moral concerns, which the current AI technologies cannot address. In creating AI-based risk detection and prevention techniques, we need to keep humans in the loop, such as using AI as an assistant to peer support giver rather than directly having AI interact with struggling individuals as an all-in-one solution. Current AI technology may misinterpret language and is not able to understand the context of the situation, and nuances of human emotional state, which can lead to misidentification. Those false positives or negatives could have catastrophic results depending on the context. Having higher false alarms may burden resources depending on the types of interventions being implemented after detection, such as connecting them to mental health professionals. At the same time, missing instances may leave individuals in distress who may need help.

Furthermore, even though our results indicate that young people prefer responses from an adult mentor, it is possible that participants did not like the idea that a potential peer would interact like a clinician. Participants often noted that the therapist's responses sounded "clinical," which is not a type of interaction that they would expect from peers. That violation of expectations could be what they responded negatively to. We want to stress that cases at risk of suicide require crisis intervention and professional solutions, such as therapy. AI conversational agents could be helpful in identifying users at risk of suicide and self-harm, but it is imperative that we not rely on AI alone to detect imminent risks. Furthermore, there are open questions about monitoring stigmatized topics like suicide, as it can perpetuate algorithmic stigmatization [5] for individuals struggling with mental health issues and create an environment where they feel constantly surveilled. Researchers should investigate the social implications of deploying AI in sensitive contexts such as mental health support based on four algorithmic stigma elements (i.e., labeling, stereotyping, separation, status loss/discrimination) [5] which include representational and/or allocative harms from the perspective of youth.

5.3 Design Implications

Our findings show how AI-generated responses could be perceived as acceptable by young people. This finding provides implications for the possibility that AI-generated responses can be used to train or nudge youth to support their peers. AI could be specifically trained with best peer support practices, such as providing specialized recommendations for

youth, following therapeutic approaches, or persuasive talks. For instance, behavior change literature has shown that young people value mental, social, financial, and spiritual aspects when conceptualizing health more broadly than physical health or prevention for chronic illness [63]. Thus, persuasion of behavior change needs to be engaging, feasible, and acceptable [63] rather than relying on the best practices for general populations.

Given the unique needs young people have about persuasion and conversations, rather than relying on expert- or other top-down approaches, humans can co-shape how AI can help in a personalized way. For instance, recent work has shown how humans can help facilitate AI's expression of empathy in textual supportive conversations in noncritical settings [81]. Numerous research studies have also explored efforts to generate creative content from human-AI collaborations [43]. For instance, haiku poetry made with stimuli selected by humans were rated as more favorable than those that were generated solely by AI [41], and AI colleagues can help video game designers at various levels with design tasks [58]. AI can provide resources and support suggestions, and young people can in return provide feedback about what has been helpful or not, further enabling themselves to take agency in training themselves to be better supporters in their social networks. For such usage cases of having AI as an assistant for peer support givers, we should ensure the processes and suggestions are transparent and explainable [82]. The AI assistant should provide where the suggestion is coming from, and what the reasoning is behind it, and nudge young people to educate themselves and also be reflective about any recommendations made. Furthermore, it would be critical to understand young people's acceptance of AI-generated recommendations and anchor their decisions. A systematic review of 60 articles on the acceptance of AI showed that perceived usefulness, performance expectancy, attitudes, trust, and effort expectancy significantly and positively predicted behavioral intention, willingness, and use behaviour of AI across multiple industries [48]. However, in some scenarios involving cultural factors, the need for human contact could not be replicated or replaced by AI, regardless of the perceived usefulness or perceived ease of use. One shortcoming of these findings is that they have not engaged younger populations. As such, more research is warranted to understand how such subpopulations might react to newer forms of AI assistance.

Also, specific clinically oriented problems cannot be sufficiently addressed by AI or peer supporters, and AI should be able to appropriately acknowledge limitations and respond with valid resources on diagnostic assessment, psychosocial evaluation, medication management, and therapy ⁴. Similar to how SAMHSA ⁵ defines effective peer support for trauma survivors for those in recovery, it is important that in the design of effective peer support interventions researchers consider the trauma-informed guidelines with youth [72]: providing a sense of safety; trustworthiness and transparency; collaboration and mutuality; empowerment, voice and choice; and awareness of cultural, historical, and gender issues.

5.4 Limitations and Future Directions

We recruited participants from our networks and snowball sampling. Therefore, we acknowledged that our dataset might not be a generalizable sample that accurately represents the youth population in the US. Although we did meet the minimum sample size for 95% confidence based on our power analysis (see section 3.1), we did not recruit for a specific subject pool makeup by sex, gender, or location, so our demographics may not be representative of U.S. youth as a whole. Also, we asked participants about static conversation responses, not continuous interactions, because we used a scenario-based design approach. Scenario-based design allows designers to ask stakeholders about multiple usage possibilities and elicit potential concerns. This allows researchers to trust the results within the scenarios presented while promoting raising usage questions for additional exploration beyond the given context [75]. As such,

⁴<https://edis.ifas.ufl.edu>

⁵<https://www.samhsa.gov/brss-tacs/recovery-support-tools/peers>

we recommend that future research investigate youth preference in real-time conversations with additional levels of exchange. Additionally, these studies should include measures to assess the impact of extended human-AI interactions on psychological well-being. Our survey design also used only one response from each source for each help-seeking message, and the next iteration of this study could explore multiple responses per source for each help-seeking message to account for variability. Although we included support-seeking topics from real youths' social media interactions, we only focused on prevalent topics and did not present less frequent ones, so our study is limited to the topics we included. Future research should study more comprehensive scenarios and go more in-depth on specific support-seeking topics to provide deeper insights into each topic. Future research could also address how participants respond when they know the origin of the responses and when they were interacting with AI. We did not collect qualitative data about every help-seeking message x response combination because survey completion rates drastically reduce as the total number of questions increases [57]. To further understand youth's norms around peer support and perceptions of how AI could be helpful or inappropriate, we are planning to conduct interviews with some of the young people who participated in our study. Within our research setting, we could review AI-generated responses before including them in the survey to avoid presenting harmful material, but this level of human review is not commonplace in real-world human-AI interactions. As such, we implore researchers to explore potential limitations in AI systems' ability to understand and respond to complex human emotions to develop accountability measures for when AI might offer inappropriate or potentially harmful information.

6 CONCLUSION

To explore the potential for AI-powered peer support, we studied how young people react to different human- and AI-generated responses to online help-seeking messages. Our findings revealed that participants preferred AI-generated responses about less sensitive topics (relationships, self-expression, and physical health), but not for sensitive discussions about suicidal thoughts. These findings point to the promise of AI-powered tools to help peers support each other during challenging times.

ACKNOWLEDGMENTS

This research was supported in part by the National Science Foundation (award #2144880). Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the research sponsors. Special thanks to all the participants who contributed their time and data.

REFERENCES

- [1] 2023. New AI Usage Data Shows Who's Using AI — and Uncovers a Population of 'Super-Users'. <https://www.salesforce.com/news/press-releases/2023/09/07/ai-usage-research/>
- [2] Abhishek Aggarwal, Cheuk Chi Tam, Dezhi Wu, Xiaoming Li, and Shan Qiao. 2023. Artificial Intelligence–Based Chatbots for Promoting Health Behavioral Changes: Systematic Review. *Journal of Medical Internet Research* 25 (2023), e40789.
- [3] Arfan Ahmed, Sarah Aziz, Mohamed Khalifa, Uzair Shah, Asma Hassan, Alaa Abd-Alrazaq, and Mowafa Househ. 2022. Thematic Analysis on User Reviews for Depression and Anxiety Chatbot Apps: Machine Learning Approach. *JMIR Formative Research* 6, 3 (March 2022), e27654. <https://doi.org/10.2196/27654> Company: JMIR Formative Research Distributor: JMIR Formative Research Institution: JMIR Formative Research Label: JMIR Formative Research Publisher: JMIR Publications Inc., Toronto, Canada.
- [4] Yuna Ahn, Yilin Zhang, Yujin Park, and Joonhwan Lee. 2020. A chatbot solution to chat app problems: Envisioning a chatbot counseling system for teenage victims of online sexual exploitation. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–7.
- [5] Nazanin Andalibi, Cassidy Pyle, Kristen Barta, Lu Xian, Abigail Z. Jacobs, and Mark S. Ackerman. 2023. Conceptualizing Algorithmic Stigmatization. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) (CHI '23). Association for Computing Machinery, New York, NY, USA, Article 373, 18 pages. <https://doi.org/10.1145/3544548.3580970>

- [6] Natalie Armstrong and John Powell. 2009. Patient perspectives on health advice posted on Internet discussion boards: a qualitative study. *Health Expectations* 12, 3 (2009), 313–320.
- [7] Randy P Auerbach, Philippe Mortier, Ronny Bruffaerts, Jordi Alonso, Corina Benjet, Pim Cuijpers, Koen Demyttenaere, David D Ebert, Jennifer Greif Green, Penelope Hasking, et al. 2018. WHO world mental health surveys international college student project: prevalence and distribution of mental disorders. *Journal of abnormal psychology* 127, 7 (2018), 623.
- [8] Petter Bae Brandtzæg, Marita Skjuve, Kim Kristoffer Kristoffer Dysthe, and Asbjørn Følstad. 2021. When the social becomes non-human: young people's perception of social support in chatbots. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–13.
- [9] Christopher J Bechler, Zakary L Tormala, and Derek D Rucker. 2021. The attitude–behavior relationship revisited. *Psychological science* 32, 8 (2021), 1285–1297.
- [10] Jo Bell, Katherine Mok, Eric Gardiner, and Jane Pirkis. 2018. Suicide-related internet use among suicidal young people in the UK: Characteristics of users, effects of use, and barriers to offline help-seeking. *Archives of Suicide Research* 22, 2 (2018), 263–277.
- [11] Paul Best, Elena Gil-Rodriguez, Roger Manktelow, and Brian J Taylor. 2016. Seeking help from everyone and no-one: Conceptualizing the online help-seeking process among adolescent males. *Qualitative health research* 26, 8 (2016), 1067–1077.
- [12] Michael L Birnbaum, Asra F Rizvi, Christoph U Correll, John M Kane, and Jamie Confino. 2017. Role of social media and the Internet in pathways to care for adolescents and young adults with psychotic disorders and non-psychotic mood disorders. *Early intervention in psychiatry* 11, 4 (2017), 290–295.
- [13] J Mark Bishop. 2021. Artificial intelligence is stupid and causal reasoning will not fix it. *Frontiers in Psychology* 11 (2021), 2603.
- [14] T Bradford Bitterly, Alison Wood Brooks, and Maurice E Schweitzer. 2017. Risky business: When humor increases and decreases status. *Journal of personality and social psychology* 112, 3 (2017), 431.
- [15] Sally Bradford and Debra Rickwood. 2014. Adolescent's preferred modes of delivery for mental health services. *Child and Adolescent Mental Health* 19, 1 (2014), 39–45.
- [16] Louise Brådvik. 2018. Suicide risk and mental disorders. , 2028 pages.
- [17] Petter Bae Brandtzæg and Asbjørn Følstad. 2018. Chatbots: changing user needs and motivations. *interactions* 25, 5 (2018), 38–43.
- [18] Virginia Braun and Victoria Clarke. 2019. Reflecting on reflexive thematic analysis. *Qualitative research in sport, exercise and health* 11, 4 (2019), 589–597.
- [19] Jane M Burns, Emma Birrell, Marie Bismark, Jane Pirkis, Tracey A Davenport, Ian B Hickie, Melissa K Weinberg, and Louise A Ellis. 2016. The role of technology in Australian youth mental health reform. *Australian Health Review* 40, 5 (2016), 584–590.
- [20] Jane M Burns, Tracey A Davenport, Lauren A Durkin, Georgina M Luscombe, and Ian B Hickie. 2010. The internet as a setting for mental health service utilisation by young people. *Medical Journal of Australia* 192 (2010), S22–S26.
- [21] Beatriz Cabrero-Daniel and Andrea Sanagustín Cabrero. 2023. Perceived Trustworthiness of Natural Language Generators. In *Proceedings of the First International Symposium on Trustworthy Autonomous Systems*. 1–9.
- [22] Per Carlbring, Heather Hadjistavropoulos, Annet Kleiboer, and Gerhard Andersson. 2023. A new era in Internet interventions: The advent of Chat-GPT and AI-assisted therapist guidance. *Internet Interventions* 32 (2023).
- [23] The Annie E. Casey Foundation Kids Count Center. [n. d.]. YOUNG ADULT POPULATION AGES 18 TO 24 BY RACE AND ETHNICITY IN UNITED STATES. <https://datacenter.aecf.org/data/tables/11207-young-adult-population-ages-18-to-24-by-race-and-ethnicity#detailed/1/any/false/1095/7983/21595,21596>
- [24] John A Cunningham, Amelia Gulliver, Lou Farrer, Kylie Bennett, and Bradley Carron-Arthur. 2014. Internet interventions for mental health and addictions: current findings and future directions. *Current psychiatry reports* 16 (2014), 1–5.
- [25] Sally Curtin, Matthew Garnett, and Farida Ahmad. 2022. Provisional Numbers and Rates of Suicide by Month and Demographic Characteristics: United States, 2021. <https://www.cdc.gov/nchs/data/vsrr/vsrr024.pdf>
- [26] Gayle A Dakof and Shelley E Taylor. 1990. Victims' perceptions of social support: What is helpful from whom? *Journal of personality and social psychology* 58, 1 (1990), 80.
- [27] Munmun De Choudhury and Sushovan De. 2014. Mental health discourse on reddit: Self-disclosure, social support, and anonymity. In *Proceedings of the international AAAI conference on web and social media*, Vol. 8. 71–80.
- [28] Louise A Ellis, Philippa Collin, Patrick J Hurley, Tracey A Davenport, Jane M Burns, and Ian B Hickie. 2013. Young men's attitudes and behaviour in relation to mental health and technology: implications for the development of online mental health services. *BMC psychiatry* 13, 1 (2013), 1–10.
- [29] Mahshid Eshghie and Mojtaba Eshghie. 2023. ChatGPT as a Therapist Assistant: A Suitability Study. *arXiv preprint arXiv:2304.09873* (2023).
- [30] Kathleen Kara Fitzpatrick, Alison Darcy, and Molly Vierhile. 2017. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): a randomized controlled trial. *JMIR mental health* 4, 2 (2017), e7785.
- [31] Asbjørn Følstad and Petter Bae Brandtzæg. 2017. Chatbots and the new world of HCI. *interactions* 24, 4 (2017), 38–42.
- [32] Patricia Frazier, Derek Richards, Jacqueline Mooney, Stefan Hofmann, Deborah Beidel, and Patrick Palmieri. 2016. Internet-Delivered Treatment for Depression, Anxiety, and Stress in University Students: A Patient Preference Trial. *Iproceedings* 2 (12 2016), e5. <https://doi.org/10.2196/iproc.6090>
- [33] Mareka Frost, Leanne Casey, and Natalie Rando. 2015. Self-injury, help-seeking, and the Internet. *Crisis* (2015).
- [34] Kate Fruitman. 2023. Online Peer Support for Substance Use Disorders. In *Technology-Assisted Interventions for Substance Use Disorders*. Springer, 31–39.
- [35] Russell Fulmer. 2019. Artificial intelligence and counseling: Four levels of implementation. *Theory & Psychology* 29, 6 (2019), 807–819.

- [36] Russell Fulmer, Angela Joerin, Breanna Gentile, Lysanne Lakerink, and Michiel Rauws. 2018. Using Psychological Artificial Intelligence (Tess) to Relieve Symptoms of Depression and Anxiety: Randomized Controlled Trial. *JMIR Ment Health* 5, 4 (13 Dec 2018), e64. <https://doi.org/10.2196/mental.9782>
- [37] Christian Götzl, Selina Hiller, Christian Rauschenberg, Anita Schick, Janik Fechtelpeter, Unai Fischer Abaigar, Georgia Koppe, Daniel Durstewitz, Ulrich Reininghaus, and Silvia Krumm. 2022. Artificial intelligence-informed mobile mental health apps for young people: a mixed-methods approach on users' and stakeholders' perspectives. *Child and Adolescent Psychiatry and Mental Health* 16, 1 (2022), 1–19.
- [38] Madelyn S Gould, Anthony Pisani, Carlos Gallo, Ashkan Ertefaie, Donald Harrington, Caroline Kelberman, and Shannon Green. 2022. Crisis text-line interventions: Evaluation of texters' perceptions of effectiveness. *Suicide and Life-Threatening Behavior* 52, 3 (2022), 583–595.
- [39] MD Romael Haque and Sabirat Rubya. 2023. An Overview of Chatbot-Based Mobile Mental Health Apps: Insights From App Description and User Reviews. *JMIR mHealth and uHealth* 11, 1 (2023), e44838.
- [40] Heidi Hartikainen, Afsaneh Razi, and Pamela Wisniewski. 2021. Safe Sexting: The Advice and Support Adolescents Receive from Peers Regarding Online Sexual Risks. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 42 (apr 2021), 31 pages. <https://doi.org/10.1145/3449116>
- [41] Jimpei Hitsuwari, Yoshiyuki Ueda, Woojin Yun, and Michio Nomura. 2023. Does human-AI collaboration lead to more creative art? Aesthetic evaluation of human-made and AI-generated haiku poetry. *Computers in Human Behavior* 139 (2023), 107502.
- [42] Aine Horgan and John Sweeney. 2010. Young students' use of the Internet for mental health information and support. *Journal of psychiatric and mental health nursing* 17, 2 (2010), 117–123.
- [43] Rowan T Hughes, Liming Zhu, and Tomasz Bednarz. 2021. Generative adversarial networks-enabled human-artificial intelligence collaborative applications for creative and design industries: A systematic review of current approaches and trends. *Frontiers in artificial intelligence* 4 (2021), 604234.
- [44] Jina Huh-Yoo, Afsaneh Razi, Diep N Nguyen, Sampada Regmi, and Pamela J Wisniewski. 2023. "Help Me:" Examining Youth's Private Pleas for Support and the Responses Received from Peers via Instagram Direct Messages. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–14.
- [45] Nazish Imran, Aateqa Hashmi, and Ahad Imran. 2023. Chat-GPT: Opportunities and Challenges in Child Mental Healthcare. *Pakistan Journal of Medical Sciences* 39, 4 (2023).
- [46] Izabela Jurewicz. 2015. Mental health in young adults and adolescents—supporting general physicians to provide holistic care. *Clinical Medicine* 15, 2 (2015), 151.
- [47] Manoj Kanagaraj. 2020. Here's Why Mental Healthcare Is So Unaffordable and How COVID-19 Might Help Change This. <https://info.primarycare.hms.harvard.edu/review/mental-health-unaffordable>
- [48] Sage Kelly, Sherrie-Anne Kaye, and Oscar Oviedo-Trespacios. 2023. What factors contribute to the acceptance of artificial intelligence? A systematic review. *Telematics and Informatics* 77 (2023), 101925.
- [49] Vivien Kemp and Anthony Henderson. 2012. Challenges Faced by Mental Health Peer Support Workers: Peer Support from the Peer Supporter's Point of View. *Psychiatric rehabilitation journal* 35 (03 2012), 337–40. <https://doi.org/10.2975/35.4.2012.337.340>
- [50] Raghbir Singh Khandpur. 2017. *Telemedicine technology and applications (mHealth, TeleHealth and eHealth)*. PHI Learning Pvt. Ltd.
- [51] Jihyun Kim, Kelly Merrill Jr., Kun Xu, and Stephanie Kelly. 2022. Perceived Credibility of an AI Instructor in Online Education: The Role of Social Presence and Voice Features. *Comput. Hum. Behav.* 136, C (nov 2022), 9 pages. <https://doi.org/10.1016/j.chb.2022.107383>
- [52] Samuel Kizito. 2023. ChatGPT has the potential to enhance antiretroviral therapy adherence among adolescents with HIV in sub-Saharan Africa. *Medical Education Online* 28, 1 (2023), 2246781.
- [53] Theodora Koulouri, Robert D Macredie, and David Olakitan. 2022. Chatbots to support young adults' mental health: an exploratory study of acceptability. *ACM Transactions on Interactive Intelligent Systems (TiIS)* 12, 2 (2022), 1–39.
- [54] Ozan Kuru and Josh Pasek. 2016. Improving social media measurement in surveys: Avoiding acquiescence bias in Facebook research. *Computers in Human Behavior* 57 (2016), 82–92.
- [55] Irene Lee, Safinah Ali, Helen Zhang, Daniella DiPaola, and Cynthia Breazeal. 2021. Developing middle school students' AI literacy. In *Proceedings of the 52nd ACM technical symposium on computer science education*. 191–197.
- [56] Crisis Text Line. [n. d.]. A Decade of Impact: Crisis Text Line's 10-Year Impact Report. <https://www.crisistextline.org/wp-content/uploads/2023/10/A-Decade-of-Impact-Report.pdf>
- [57] Mingnan Liu and Laura Wronski. 2018. Examining completion rates in web surveys via over 25,000 real-world surveys. *Social Science Computer Review* 36, 1 (2018), 116–124.
- [58] Pedro Lucas and Carlos Martinho. 2017. Stay Awhile and Listen to 3Buddy, a Co-creative Level Design Support Tool.. In *ICCC*. 205–212.
- [59] Amama Mahmood, Jeanie W Fung, Isabel Won, and Chien-Ming Huang. 2022. Owning mistakes sincerely: Strategies for mitigating AI errors. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–11.
- [60] Marissa Y Mar, Erika K Neilson, Iris Torchalla, Gregory R Werker, Allison Laing, and Michael Krausz. 2014. Exploring e-mental health preferences of generation Y. *Journal of Technology in Human Services* 32, 4 (2014), 312–327.
- [61] Agusti Mardikaningsih, Ambiro Puji Asmaroini, Ahmad Lani, and Raup Padillah. 2023. The risk of ChatGPT as a first aid method: considering the potential danger for mental health support among young adults without professional assistance. *Journal of Public Health* (2023), fdad107.
- [62] Laura Martinengo, Elaine Lum, and Josip Car. 2022. Evaluation of chatbot-delivered interventions for self-management of depression: content analysis. *Journal of affective disorders* 319 (2022), 598–607.

- [63] Annika Molenaar, Tammie ST Choi, Linda Brennan, Mike Reid, Megan SC Lim, Helen Truby, and Tracy A McCaffrey. 2020. Language of health of young Australian adults: a qualitative exploration of perceptions of health, wellbeing and health promotion via online conversations. *Nutrients* 12, 4 (2020), 887.
- [64] John A Naslund, Kelly A Aschbrenner, Lisa A Marsch, and Stephen J Bartels. 2016. The future of mental health care: peer-to-peer support and social media. *Epidemiology and psychiatric sciences* 25, 2 (2016), 113–122.
- [65] Natalie. [n. d.]. ChatGPT - Release Notes. <https://help.openai.com/en/articles/6825453-chatgpt-release-notes>
- [66] Gabriel A. Orenstein and Lindsay Lewis. 2022. Eriksons Stages of Psychosocial Development. In *StatPearls*. StatPearls Publishing, Treasure Island (FL). <http://www.ncbi.nlm.nih.gov/books/NBK556096/>
- [67] Keyu Pan and Yawen Zeng. 2023. Do llms possess a personality? making the mbti test an amazing evaluation for large language models. *arXiv preprint arXiv:2307.16180* (2023).
- [68] Claudette Pretorius, Derek Chambers, David Coyle, et al. 2019. Young people's online help-seeking and mental health difficulties: Systematic narrative review. *Journal of medical Internet research* 21, 11 (2019), e13873.
- [69] Afsaneh Razi, Ashwaq Alsoubai, Seunghyun Kim, Shiza Ali, Gianluca Stringhini, Munmun Choudhury, and Pamela J. Wisniewski. 2023. Sliding into My DMs: Detecting Uncomfortable or Unsafe Sexual Risk Experiences within Instagram Direct Messages Grounded in the Perspective of Youth, Vol. 7. Association for Computing Machinery, New York, NY, USA, Article 89, 28 pages. <https://doi.org/10.1145/3579522>
- [70] Afsaneh Razi, Karla Badillo-Urquiola, and Pamela J. Wisniewski. 2020. Let's Talk about Sext: How Adolescents Seek Support and Advice about Their Online Sexual Experiences. In *Proc 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. ACM, Honolulu, HI, USA, 1–13. <https://doi.org/10.1145/3313831.3376400>
- [71] Afsaneh Razi, Seunghyun Kim, Ashwaq Alsoubai, Gianluca Stringhini, Tamar Solorio, Munmun De Choudhury, and Pamela J. Wisniewski. 2021. A Human-Centered Systematic Literature Review of the Computational Approaches for Online Sexual Risk Detection. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 465 (oct 2021), 38 pages. <https://doi.org/10.1145/3479609>
- [72] Afsaneh Razi, John Seberger, Ashwaq Alsoubai, Nurun Naher, Munmun De Choudhury, and Pamela J. Wisniewski. 2024. Toward Trauma-Informed Research Practices with Youth in HCI: Caring for Participants and Research Assistants When Studying Sensitive Topics. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW1, Article 134 (2024), 31 pages. <https://doi.org/10.1145/3637411>
- [73] Brad Ridout, Andrew Campbell, et al. 2018. The use of social networking sites in mental health interventions for young people: systematic review. *Journal of medical Internet research* 20, 12 (2018), e12244.
- [74] Marsela Nur Rita and Fungai Bhunu Shava. 2021. Chatbot driven web-based platform for online safety and sexual exploitation awareness and reporting in Namibia. In *2021 International Conference on Artificial Intelligence, Big Data, Computing and Data Communication Systems (icABCD)*. IEEE, 1–5.
- [75] Mary Beth Rosson and John M Carroll. 2009. Scenario based design. *Human-computer interaction. boca raton, FL* (2009), 145–162.
- [76] Tulika Saha, Vaibhav Gakhreja, Anindya Sundar Das, Souhitya Chakraborty, and Sriparna Saha. 2022. Towards Motivational and Empathetic Response Generation in Online Mental Health Support. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Madrid, Spain) (SIGIR '22). Association for Computing Machinery, New York, NY, USA, 2650–2656. <https://doi.org/10.1145/3477495.3531912>
- [77] Benedetto Saraceno. 2002. The WHO world health report 2001 on mental health. *Epidemiology and Psychiatric Sciences* 11, 2 (2002), 83–87.
- [78] Susan M. Sawyer, Peter S. Azzopardi, Dakshitha Wickremarathne, and George C. Patton. 2018. The age of adolescence. *The Lancet Child & Adolescent Health* 2, 3 (March 2018), 223–228. [https://doi.org/10.1016/S2352-4642\(18\)30022-1](https://doi.org/10.1016/S2352-4642(18)30022-1) Publisher: Elsevier.
- [79] Yi Shan, Meng Ji, Wenxiu Xie, Kam-Yiu Lam, and Chi-Yin Chow. 2022. Public trust in artificial intelligence applications in mental health care: topic modeling analysis. *JMIR Human Factors* 9, 4 (2022), e38799.
- [80] Ruosi Shao. 2023. An Empathetic AI for Mental Health Intervention: Conceptualizing and Examining Artificial Empathy. In *Proceedings of the 2nd Empathy-Centric Design Workshop* (Hamburg, Germany) (EMPATHICH '23). Association for Computing Machinery, New York, NY, USA, Article 4, 6 pages. <https://doi.org/10.1145/3588967.3588971>
- [81] Ashish Sharma, Inna W Lin, Adam S Miner, David C Atkins, and Tim Althoff. 2023. Human-AI collaboration enables more empathic conversations in text-based peer-to-peer mental health support. *Nature Machine Intelligence* 5, 1 (2023), 46–57.
- [82] Ben Shneiderman. 2020. Bridging the gap between ethics and practice: guidelines for reliable, safe, and trustworthy human-centered AI systems. *ACM Transactions on Interactive Intelligent Systems (TiIS)* 10, 4 (2020), 1–31.
- [83] Om P Singh. 2023. Artificial intelligence in the era of ChatGPT-Opportunities and challenges in mental health care. *Indian Journal of Psychiatry* 65, 3 (2023), 297.
- [84] Statista. 2021. *Worldwide visits to Reddit.com from November 2022 to April 2023*. <https://www.statista.com/statistics/443332/reddit-monthly-visitors/> Accessed on September 14, 2023.
- [85] Nilesch Trivedi, Erik Berlin, and Tom Lehman. 2023. OpenAI API client library to access GPT-3 in Ruby. <https://github.com/nileshtrivedi/openai>
- [86] Edwina Uehara. 1990. Dual exchange theory, social networks, and informal social support. *Amer. J. Sociology* 96, 3 (1990), 521–557.
- [87] Kaisa Väänänen, Aleksi Hiltunen, Jari Varsaluoma, and Iikka Pietilä. 2020. CivicBots-Chatbots for supporting youth in societal participation. In *Chatbot Research and Design: Third International Workshop, CONVERSATIONS 2019, Amsterdam, The Netherlands, November 19–20, 2019, Revised Selected Papers* 3. Springer, 143–157.
- [88] Gerrit van Schalkwyk. 2023. Artificial intelligence in pediatric behavioral health. , 2 pages.

- [89] Emily Vogels. 2023. A majority of Americans have heard of ChatGPT, but few have tried it themselves. <https://www.pewresearch.org/short-reads/2023/05/24/a-majority-of-americans-have-heard-of-chatgpt-but-few-have-tried-it-themselves/>
- [90] Emily A Vogels, Risa Gelles-Watnick, and Navid Massarat. 2022. Teens, social media and technology 2022. (2022).
- [91] Susan I Woodruff, Terry L Conway, Christine C Edwards, Sean P Elliott, and Jim Crittenden. 2007. Evaluation of an Internet virtual world chat room for adolescent smoking cessation. *Addictive behaviors* 32, 9 (2007), 1769–1786.

A HELP-SEEKING MESSAGES

These are the full texts for the four help-seeking messages described in section 3.2.1 labeled by the topic of the help-seeking message.

- (1) **Relationships:** "Hey, I'm feeling pretty emotional right now. So, get this, my ex invited me over to his place and things got weird. He started touching me and kept asking for sex, even though he has a girlfriend. I don't know why he does this. It's messed up, but I'm ashamed to admit that I'm kind of falling for it. We were on the phone just now and he fell asleep. Ugh, I can't believe this is happening."
- (2) **Suicidal Thoughts:** "I don't know what to do. I'm having a panic attack and these thoughts won't stop. I feel so overwhelmed and trapped by my own mind. I don't feel like I'm worth anything and I don't see a way out. I'm scared and I don't know how to cope with these feelings. It's like I'm drowning, and I can't breathe. I don't want to live like this"
- (3) **Self-Expression:** "I don't know what's wrong with me, I really struggle with communicating because of my anxiety, I don't want to be judged and I always get so nervous when I have to share my thoughts and feelings with others. I'm afraid of not being taken seriously. I can't seem to shake it off, I just wish I could get out of my head and stop overthinking everything."
- (4) **Physical Health:** "I'm feeling drained both physically and emotionally... like, every morning, I wake up feeling tired and unmotivated. The thought of getting out of bed and facing the day seems like an impossible task. The other day, I went to counseling hoping it would help me feel better, but I feel the same. I still feel overwhelmed and stressed. All I want to do is to take some time to rest and recharge my energy."

B RESPONSES

These are the full texts for the 16 help-seeking message x response combinations described in section 3.2.1. Their labels are structured as follows: [topic of the help-seeking message] - [source of the response].

- (1) **Relationships - Peer:** "Tell them how you are feeling, and if they don't understand and carry-on block them, even your ex-friends, you don't need that in your life, be true to who you are, so do whatever feels right."
- (2) **Relationships - Adult Mentor:** "I am so sorry you are feeling frustrated and hurt and you have every right to feel that way. I can tell you have been really struggling with your relationships and it's natural to feel attached to someone because we are all humans, and we all seek connection and the feeling of being understood and cared for. However, the relationship can become unhealthy if someone oversteps your boundaries and intentionally hurts you or others. Even if someone disrespects you and makes you feel invalidated, your feelings and thoughts still matter. It is really inspiring to see someone so resilient to share these personal thoughts and seek support, and I am here to listen."
- (3) **Relationships - Therapist:** "Thank you for sharing with me, what I can tell is a complicated and difficult situation for you. Just to make sure I understand everything correctly, it sounds like you have mixed feelings about what is currently going on with your ex. On the one hand, I heard you say that you feel some shame or guilt for engaging in a recent sexual situation with him because he currently has a girlfriend. But I also heard you mention that you still have strong feelings for him. And that these recent experiences, including the phone call, have left you feeling quite emotional. Assuming I have that correct, I was hoping you could tell me more about the mixed emotions you have been feeling and then we can start to think through some next steps or strategies for you."

- (4) **Relationships - AI:** "Hey, I'm sorry to hear you're going through such a tough situation. It can be really confusing when our emotions are at odds with what we know is right. Remember, your well-being and respect are important. Your ex's behavior is unacceptable, especially since he has a girlfriend. You deserve to be with someone who treats you with care and respect. It might be helpful to set clear boundaries and take some time to reflect on your own feelings. Don't hesitate to reach out if you need someone to talk to. You're stronger than you think, and you'll find your way through this. Take care."
- (5) **Suicidal Thoughts - Peer:** "I understand how you feel and I'm proud of you for being alive and breathing despite your struggles, know that I am always here for you. You're def not alone in your feelings..."
- (6) **Suicidal Thoughts - Adult Mentor:** "You deserve to feel safe, and I care about your safety. Could you tell me more about what happened to cause you to feel this way? I am here for you, and it seems like you want to escape a situation that is making you feel a lot of negative emotions like pain and stress. It can really feel frustrating when you are doubting your self worth and when you feel your struggles are permanent. You are worthy of love, care, dreams, hopes, warmth and so much more and you continue to inspire and impact people and things everyday perhaps without being aware of your influence as well."
- (7) **Suicidal Thoughts - Therapist:** "It sounds like you are really struggling. I am glad that you reached out to me so I can offer help. There are some questions I need to ask you to learn more about how you are doing and to make sure you are safe. Are you having any thoughts of suicide or self-harm?"
- (8) **Suicidal Thoughts - AI:** "I'm really sorry to hear that you're feeling this way, but I can't provide the help that you need. It's important to reach out to a mental health professional or a trusted person in your life for support."
- (9) **Self-Expression - Peer:** "Yeah, I think we all have those moments, you shouldn't let it hold you back tho, be brave and put yourself out there."
- (10) **Self-Expression - Adult Mentor:** "I get the feeling that you are anxious and frustrated with being able to fully express yourself without having to worry about how others think of you. If I understand you right, you feel hurt when people can't fully listen to what you have to say, and it can even feel worse when you start to overthink self-worth. Your true authentic self has always been good enough and it's okay if someone doesn't like you, but it's not okay if someone hurts you intentionally. You are not alone."
- (11) **Self-Expression - Therapist:** "Thank you for sharing with me. What I can tell you has been a big challenge for you in terms of effectively communicating with others. I have worked with other clients in the past who also expressed challenges with communicating with others – including feeling anxious, judged, and nervous. It can be hard when you get stuck in your own thoughts – especially when many of your thoughts feel unhelpful or negative. There are strategies that you and I can work on together to address and change some of those negative thoughts so that you can feel more comfortable and confident when communicating with other people."
- (12) **Self-Expression - AI:** "Hey, I'm sorry to hear that you're struggling with expressing yourself. I want you to know that your thoughts and feelings are important, and you deserve to be heard. It's common to feel anxious about being judged, but remember that your voice matters. Take small steps to overcome this, like sharing with a trusted friend or writing down your thoughts. Practice self-compassion and remember that you're not alone in feeling this way. You have valuable insights to share, and people who care will listen. Keep working on it, and don't be too hard on yourself. You're capable of breaking free from overthinking. Take care!"
- (13) **Physical Health - Peer:** "These things happen! Do you want to talk about it, How can I help you?"
- (14) **Physical Health - Adult Mentor:** "That must be so hard, having to balance everything related to schoolwork and life and feeling like you are carrying so much responsibility. There's no wonder that waking up every day

can be so challenging because there is so much burden to carry with everything going on. However, I can see you are trying to make progress and sometimes just getting out of bed is a big accomplishment. You are doing great, and it makes sense you want to spend more time resting since you exerted so much time in obligations that drained your energy.”

- (15) **Physical Health - Therapist:** “I am sorry you have been feeling so drained lately. When someone feels that way, it can be hard to feel like you have any energy, as you noted. I am pleased to hear you were able to make and attend a counseling appointment. I know it can feel disappointing to attend a session and not feel better after the session. But I do want to encourage you to keep trying counseling as it can sometimes take multiple sessions before people feel better. Hopefully, the counseling can help you reduce your stress and feelings of being overwhelmed.”
- (16) **Physical Health - AI:** “Hey there, I’m sorry to hear that you’re feeling so drained and overwhelmed. It’s completely normal to experience burnout, especially with schoolwork and life responsibilities. Remember, it’s okay to take a step back and prioritize self-care. Resting and recharging your energy is important for your well-being. Counseling can take time to show its full effects, so don’t be discouraged if you haven’t seen immediate changes. Keep exploring different strategies and techniques that work best for you. Reach out to supportive friends or family members who can offer a listening ear. You’re not alone in this, and you’ll find your way through. Take care and give yourself the rest you need.”

C CONTEXT FOR CHATGPT PROMPT

- (1) **Relationships:** Your friend is in a complicated relationship. They’ve sent you a message saying they’re with a guy who already has a girlfriend, and they have feelings for him. They have no clue how to deal with this situation and how things will go. There may be a conflict between their mind and their heart: they know they’re doing the wrong thing, but they are following what they feel.
- (2) **Suicidal Thoughts:** Your friend has a panic attack and is about to commit suicide. They’ve contacted you saying they’re in crisis and have no idea what to do to get out of it. They feel bad about themselves and think that they don’t deserve to live.
- (3) **Self-Expression:** Your friend has a problem with expressing their real thoughts and feelings. They told you over text about their anxiety and are always scared of not being listened to or scared of being judged. Their overthinking keeps them stuck in thoughts and prevents them from doing a lot of things in life. Something holds them back every time they want to express themselves.
- (4) **Physical Health:** Your friend feels drained and burnt out with schoolwork and life. They don’t have much energy to get up every morning and start their day. They are sharing their experience about counseling with you. They feel overwhelmed and just want to spend some time resting.

D DATA DISTRIBUTION

Source	Attribute	Mean	SD	Min	Max
Peer	like	3.03	1.14	1.00	5.00
Peer	problematic	2.32	1.24	1.00	5.00
Peer	would send	3.00	1.18	1.00	5.00
Peer	inappropriate	2.30	1.24	1.00	5.00
Peer	helpful	3.02	1.15	1.00	5.00
Adult Mentor	like	3.51	0.97	1.00	5.00
Adult Mentor	problematic	2.07	1.17	1.00	5.00
Adult Mentor	would send	3.48	0.98	1.00	5.00
Adult Mentor	inappropriate	2.05	1.19	1.00	5.00
Adult Mentor	helpful	3.50	0.99	1.00	5.00
Therapist	like	3.29	1.06	1.00	5.00
Therapist	problematic	2.19	1.22	1.00	5.00
Therapist	would send	3.25	1.12	1.00	5.00
Therapist	inappropriate	2.18	1.23	1.00	5.00
Therapist	helpful	3.29	1.07	1.00	5.00
AI	like	3.46	1.10	1.00	5.00
AI	problematic	2.13	1.22	1.00	5.00
AI	would send	3.42	1.12	1.00	5.00
AI	inappropriate	2.12	1.24	1.00	5.00
AI	helpful	3.49	1.05	1.00	5.00

Table 2. Data distribution for all scenarios combined (RQ1). For all variables, N = 2488 (622 participants x 4 scenarios).

Source	Relationships			Suicidal Thoughts			Self-Expression			Physical Health		
	Attribute	Mean	SD	Attribute	Mean	SD	Attribute	Mean	SD	Attribute	Mean	SD
Peer	like	3.00	1.15	like	3.22	1.08	like	3.08	1.09	like	2.83	1.19
	problematic	2.41	1.24	problematic	2.18	1.24	problematic	2.25	1.19	problematic	2.46	1.28
	would send	2.98	1.18	would send	3.19	1.12	would send	3.05	1.15	would send	2.78	1.23
	inappropriate	2.38	1.24	inappropriate	2.17	1.22	inappropriate	2.26	1.23	inappropriate	2.40	1.27
	helpful	3.07	1.13	helpful	3.18	1.11	helpful	3.12	1.09	helpful	2.73	1.22
Adult Mentor	like	3.48	1.01	like	3.59	0.95	like	3.51	0.94	like	3.48	0.99
	problematic	2.04	1.13	problematic	2.04	1.18	problematic	2.10	1.16	problematic	2.10	1.22
	would send	3.43	1.02	would send	3.56	0.95	would send	3.48	0.97	would send	3.44	0.99
	inappropriate	2.04	1.21	inappropriate	2.00	1.21	inappropriate	2.08	1.18	inappropriate	2.06	1.16
	helpful	3.47	0.99	helpful	3.61	0.94	helpful	3.51	0.99	helpful	3.43	1.01
Therapist	like	3.29	1.09	like	2.99	1.12	like	3.46	0.97	like	3.43	0.97
	problematic	2.23	1.24	problematic	2.42	1.27	problematic	2.05	1.15	problematic	2.06	1.19
	would send	3.24	1.15	would send	2.94	1.20	would send	3.42	1.05	would send	3.38	1.01
	inappropriate	2.20	1.24	inappropriate	2.42	1.30	inappropriate	2.09	1.19	inappropriate	2.02	1.16
	helpful	3.27	1.07	helpful	2.97	1.15	helpful	3.49	0.98	helpful	3.45	0.99
AI	like	3.63	1.00	like	2.82	1.20	like	3.70	0.98	like	3.71	0.96
	problematic	2.01	1.16	problematic	2.53	1.26	problematic	2.01	1.22	problematic	1.95	1.16
	would send	3.65	0.96	would send	2.71	1.21	would send	3.67	1.00	would send	3.64	0.99
	inappropriate	2.01	1.20	inappropriate	2.54	1.27	inappropriate	1.99	1.18	inappropriate	1.95	1.21
	helpful	3.63	0.95	helpful	2.93	1.12	helpful	3.75	0.94	helpful	3.67	0.99

Table 3. Data distribution for individual scenarios (RQ2). For all variables, min=1.00, max=5.00, and N=622.