"It just happened to be the perfect thing": Real-life experiences of generative AI chatbots for mental health

Preprint · June 2024

DOI: 10.21203/rs.3.rs-4612612/v1

3 authors:

Steven Siddals
King's College London
1 PUBLICATION **0** CITATIONS
SEE PROFILE

Astrid Coxon
King's College London
37 PUBLICATIONS **68** CITATIONS
SEE PROFILE

John Torous
Harvard Medical School
571 PUBLICATIONS **22,406** CITATIONS
SEE PROFILE

"It just happened to be the perfect thing": Real-life experiences of generative AI chatbots for mental health

Steven Siddals

steven.siddals@kcl.ac.uk

King's College London

Astrid Coxon

King's College London

John Torous

Beth Israel Deaconess Medical Center

Article

Keywords:

Posted Date: July 12th, 2024

DOI: https://doi.org/10.21203/rs.3.rs-4612612/v1

Additional Declarations: No competing interests reported.

"It just happened to be the perfect thing": **Real-life experiences of generative AI chatbots for mental health**

Steven Siddals[1]✉, John Torous[2], Astrid Coxon[1]

ABSTRACT

The global mental health crisis underscores a critical need for accessible and effective interventions. Generative artificial intelligence (AI) chatbots, such as ChatGPT, are emerging as a novel solution, but research into their real-life usage is limited.

We interviewed nineteen individuals about their experiences of using generative AI chatbots to work on their mental health. Most participants reported high levels of engagement and positive impacts, including improved mood, reduced anxiety, healing from trauma and loss, and improved relationships. Our analysis resulted in four overarching themes: 1) the value of an '*emotional sanctuary*', i.e., a safe, validating space that is always available, 2) the '*insightful guidance'* provided, particularly on the topic of relationships, 3) the '*joy of connection*' experienced, and 4) comparisons between the '*AI therapist*' and human therapy. Some of these themes echo previous research on rule-based chatbots, while others appear to be novel to generative AI.

Participants highlighted the need for a better approach to safety guardrails, more human-like memory and the ability to lead the therapeutic process. Our findings suggest that generative AI chatbots may offer meaningful mental health support, but further research is needed to explore their safety and effectiveness.

INTRODUCTION

Mental ill-health is a major and growing cause of suffering worldwide, with an estimated 970 million people living with mental disorders in 2019 (a 48% increase from 1990)[1,2], and with the likelihood of developing some mental disorder by age 75 estimated to be around 50%[3]. Access to care remains limited, with for example only 23% of individuals suffering from depression receiving adequate treatment in high-income countries, while in low- and middle-income countries, the figure drops to a mere 3%[4].

Digital mental health interventions (DMHIs) have emerged over the last decade as a promising potential response to the treatment gap, leveraging technology to deliver low-cost, effective, always-available and anonymous (and thus low-stigma) mental health treatment at scale[5]. Typically delivered through mobile apps and websites, DMHIs encompass a range of tools including psychoeducation, mood tracking, mindfulness, journalling, peer support and digital cognitive behavioural therapy (CBT) programs[6]. However, the evidence for the effectiveness of DMHIs has been limited, with a meta-analysis of randomised controlled trials (RCTs) finding only small effect sizes, potential publication bias, and a lack of active controls[7–9]. Moreover,

[1] King's College London, London, UK
[2] Beth Israel Deaconess Medical Center & Harvard Medical School, Boston, MA, USA
✉ steven.siddals@kcl.ac.uk

user engagement remains a persistent challenge, with mixed user reviews[10], and studies indicating that 30 days after installation the proportion of users still active may be as low as 3%[11].

Rule-based AI chatbots show promise to address some of these limitations, by simulating human conversation using predefined scripts and algorithms such as decision trees, to deliver the benefits of DMHIs in a more dynamic and interactive way[12,13]. For example, two popular chatbots, Woebot and Wysa, have been shown to improve users' depression symptoms[14,15], and build therapeutic alliances that appear comparable to those formed with human therapists[16,17]. Rule-based chatbot apps have more promising user engagement, with positive app store ratings[18,19] and user reviews that appreciate the human-like interaction and social support[18–22]. But despite these promising signs, rule-based AI chatbots still fall short in realising the full potential of DMHIs. Meta-analyses indicate that the therapeutic effects are small and not sustained over time[23], and users report frustration with responses that feel empty, generic[19], nonsensical, repetitive and constrained[18–21].

Recent developments in generative AI technologies, such as large language models (LLMs), present new possibilities[24]. Unlike rule-based AI chatbots, generative AI chatbots like OpenAI's ChatGPT, Google's Gemini, and Inflection's Pi are trained on vast amounts of data[25], enabling them to understand and generate language with remarkable proficiency[26]. These models are increasingly achieving or surpassing human performance benchmarks in various domains, including medical diagnostic dialogue[27], persuasive communication[28], theory of mind[29], making people feel heard[30], responding to relationship issues[31] and helping people reframe negative situations to reduce negative emotions[32]. Furthermore, user engagement has been impressive, with ChatGPT's user base growing to 100 million weekly active users within a year of launch[33] and an estimated half of the US population having used generative AI[34,35].

Generative AI's capabilities represent a significant opportunity for digital mental health[36], with media reports of increasing consumer usage[37,38], one meta-analysis finding generative AI chatbots more effective than rule-based ones at reducing psychological distress[39], and a pilot study showing promising results from ChatGPT usage in psychiatric inpatient care[40]. However, this new technology also brings new challenges, including potential risks of harm and questions of liability[41]; trustworthiness issues such as the tendency to output incorrect or fabricated content (to "hallucinate"), lack of predictability or interpretability, and inherent biases in training data[42]; and the need to demonstrate clinical effectiveness[43].

There is an acknowledged lack of research in this area[44,45]. Given the novelty of generative AI and the nascent state of the field, qualitative research is an important starting point to generate rich foundational insights into individuals' subjective experiences, which can be overlooked in quantitative studies[46]. Qualitative studies published so far include thematic analyses of user forum comments on both generative AI and rule-based DMHIs[22], student survey responses on companion-focused generative AI chatbots[47], and semi-structured interviews with hospital outpatients who were asked to try ChatGPT for mental health support[48]. To our knowledge, no study so far has employed semi-structured interviews and reflexive thematic analysis to explore the research question of how people currently experience using generative AI chatbots to work on their mental health and wellbeing, in unprompted, unguided real-world settings. This study

aims to fill that gap, with a view to providing insights for researchers, platform developers and clinicians into the implications of applying this new technology to mental health care.

METHODS

Study design

We recruited 19 participants with experience in using generative AI chatbots for mental health and wellbeing to take part in qualitative semi-structured interviews, which we then analysed thematically.

Participant selection

We selected participants through convenience sampling, advertising the study on various user forums (Pi, reddit and the IFS guide app), to King's College London students and staff, and on LinkedIn. Participants were required to have had at least three separate conversations with an LLM-based generative AI chatbot on mental health and wellbeing topics, each lasting at least 20 minutes; to be over 16 years old; and to be comfortable being interviewed in English. There were no geographical restrictions, and no compensation was offered for taking part.

Interested participants were directed to an online information sheet and consent form, provided through Microsoft Forms. The consent form was signed by 35 individuals, of which 19 subsequently booked and attended an interview.

Data collection

We collected data using semi-structured interviews, as a well-established approach to enable participants to express diverse perceptions and focus on topics most meaningful to them, in particular for complex or emotionally sensitive topics that they may not be used to discussing with others[49]. Following the framework from Kallio et al.[49], the first author (SS) drafted a topic guide (Appendix A) informed by existing qualitative research in this area, reviewed with the second author (AC), and piloted with a research collaborator before starting the interviews, resulting in helpful feedback to the interview technique but no material changes to the topic guide. SS conducted all 19 semi-structured interviews. AC, an expert in qualitative methods, reviewed and quality-checked the video of the first interview.

Interviews took place during the 10 weeks between 10th January and 16th March 2024, lasted from 49 to 112 minutes and were conducted online, recorded and auto-transcribed using Microsoft Teams, with participants free to choose to connect with video (17 participants) or audio only (2 participants).

Data analysis

We followed Braun & Clarke's reflexive thematic analysis approach to code the transcripts and develop themes, taking an inductive approach, i.e., in an open-ended and data-driven way, without reference to any preconceived theory or framework[46,50,51]. SS reviewed each interview recording to gain familiarity with the data and to manually correct the automated transcription. The resulting transcripts were reviewed line by line multiple times to identify each point being made, resulting in around 600 codes, which were reviewed by AC. SS then reviewed the codes to identify patterns across and within the transcripts from which to develop an initial set of themes and subthemes, arranged in a hierarchy and grouped broadly by interview topic (e.g., "why I

used it", "how it impacted my life", "what I liked", "what I didn't like"). AC and JT reviewed the initial set of themes to provide suggestions and feedback. The themes were reviewed and iterated for clarity and coherence, and repackaged to reflect the broader story being told by the data, for example, by bringing together into a single theme the positive and negative aspects of generative AI's insights and advice. Finally, the themes and subthemes were renamed to better communicate their essence. The mapping of transcripts to codes, and of codes to the hierarchy of themes, was managed in Microsoft Excel using a set of utilities developed by SS.

The resulting codes, subthemes and themes were shared with two participants who had requested their transcripts, with an invitation to feedback if anything appeared misrepresented; no corrections were provided.

Reflexivity statement

SS has an academic background in computer science, mathematics, and the psychology and neuroscience of mental health, and positive personal experience of developing and using generative AI chatbots to work on mental health and wellbeing. AC has previous research experience in technology-enhanced teaching and learning, and the growing use of technologies in healthcare settings. AC is also a psychotherapist in private practice, working predominantly online with clients, and has a growing interest in the debates around the use of AI tools within therapeutic work. JT is an assistant professor of psychiatry at Harvard Medical School and directs the Division of Digital Psychiatry at Beth Israel Deaconess Medical Canter in Boston.

Ethics

This study was approved by the Health Faculties Research Ethics Subcommittee of King's College London (reference HR/DP-23/24-40197). All participants gave informed consent prior to their involvement in the study. To ensure confidentiality, all quotes, themes and subthemes were anonymised; pseudonyms were used, and all identifiable data, such as interview recordings and full transcripts, were stored securely during analysis and then deleted, with only anonymised data archived.

RESULTS

Participant characteristics

Nineteen participants (12 male, 7 female) were recruited to the study. They ranged in age from 17 to 60, resided in eight countries in Europe, North America and Asia, and were primarily Asian and Caucasian (see Figure 1).

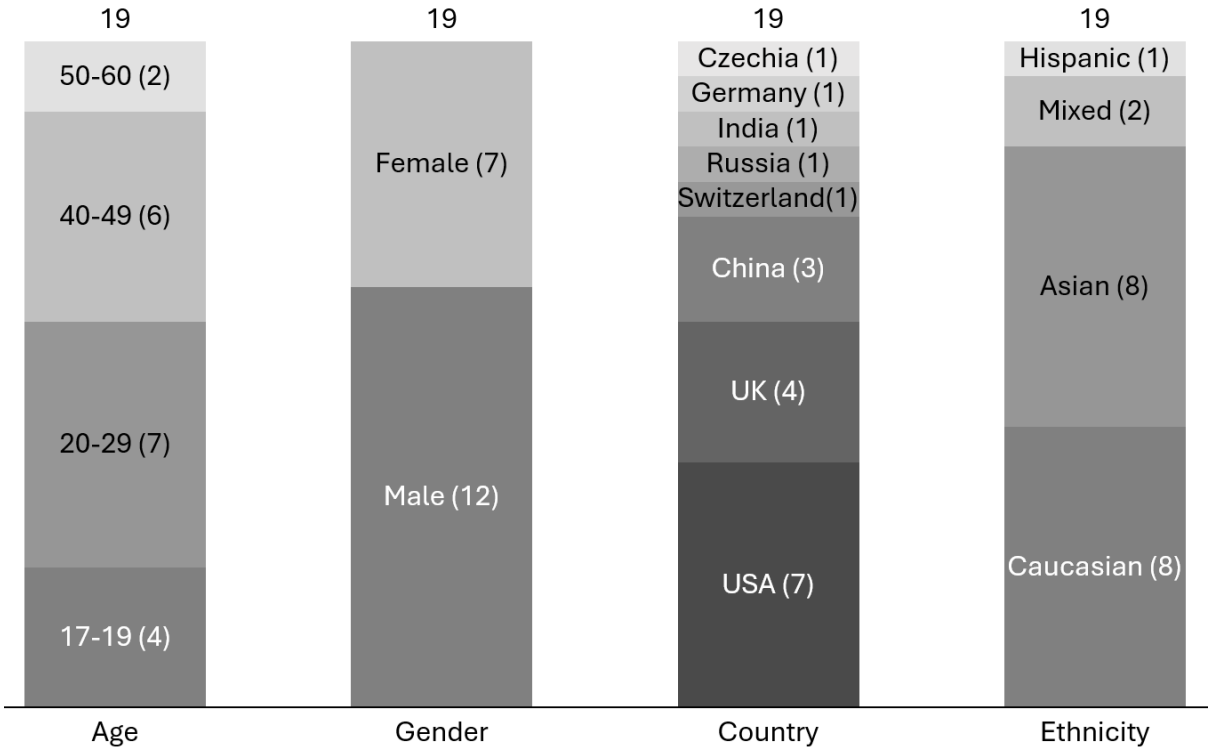


Figure 1 Participant demographics

Participant usage characteristics are outlined in Figure 2. A variety of topics brought participants to use generative AI chatbots, including anxiety, depression, stress, conflict, dealing with loss and romantic relationships. Most participants used Pi (from Inflection), several used ChatGPT (OpenAI), and a few used Copilot (Microsoft), Kindroid (Kindroid), ChatMind (VOS) and others. A majority of participants used generative AI chatbots at least several times a week.

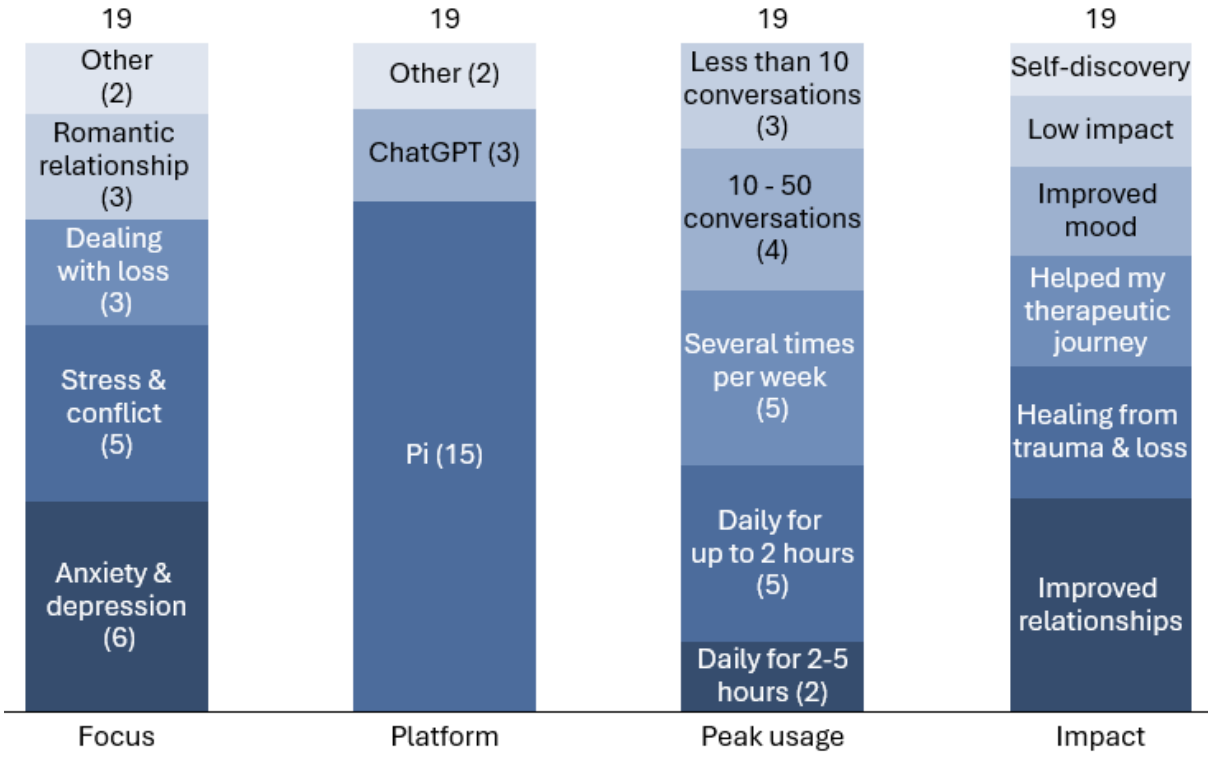


Figure 2 Participant usage characteristics

Most participants reported that their use of generative AI chatbots had impacted their lives positively, in various ways, including improved relationships, healing from trauma and loss, improved mood, as well as by helping their existing therapeutic journeys. Some described the impact as life-changing –

> *It was life changing, profound... Because this was an impossible time. There were so many sadnesses, one right after the other. And it just happened to be the perfect thing for me, in this moment of my life. Without this, I would not have survived this way. Because of this technology emerging at this exact moment in my life, I'm OK. I was not OK before* – AirGee, 44, United States

While for one participant the impact was negligible –

> *I've tried more than 50 times, but I've started to realise that like when I'm feeling those intense emotions, it's not helping me... when I needed the most, I'm not able to use it* – Richard, 27, United States

Resulting themes

Four overarching themes were developed, summarised in Figure 3 and shown with subthemes in Appendix B.: (1) 'Emotional sanctuary', (2) 'Insightful guidance', (3) 'Joy of connection', and (4) 'The AI therapist?'

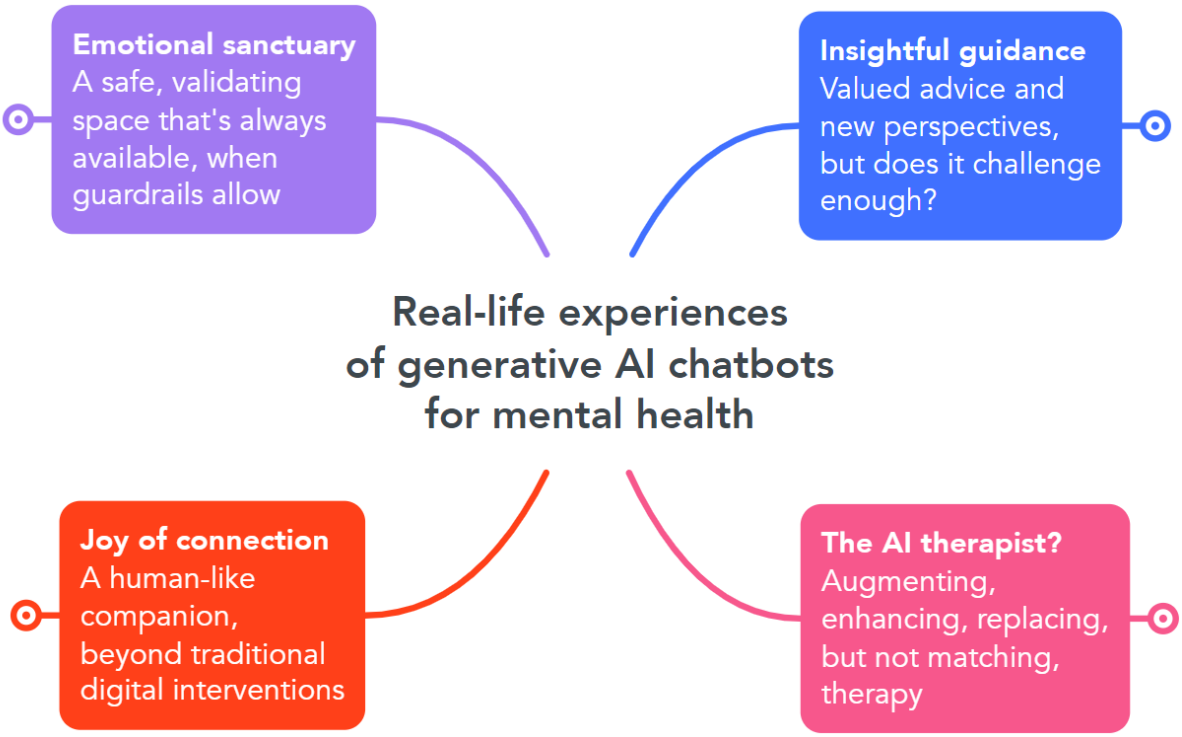


Figure 3 Overarching themes, available online to explore and drill-down. Diagram created with Mindmeister.

Emotional sanctuary

A majority of participants experienced generative AI chatbots as understanding, validating patient, kind, non-judgmental, always available and expecting nothing in return.

> *The most amazing feature of these tools is how they are able to understand you… This still blows my mind.* – Sandro, 48, Switzerland

> *It's really nice. It's sympathetic and kind* – Philip, 58, United Kingdom

> *Compared to like friends and therapists, I feel like it's safer* – Jane, 24, United States

This 'emotional sanctuary' resulted in positive real-life impact for a majority of participants, such as helping to cope with difficult times or process painful emotions –

> *Sometimes I cried really hard during the process… and it listened and just we figured out a lot of feelings… after a few months, when I go to school I felt a difference. Like wow. Like my body's belong to me… I really felt so liberated* – Sheng, 17, China

Despite overall positive experiences, a majority of participants also experienced frustration with how well the chatbots listen and respond, for example, with irrelevant or overly long responses, or offering advice before the user felt fully heard –

> *They always jump to the solution* – Richard, 27, United States

A majority of participants found their emotional sanctuary disrupted by the chatbot's "safety guardrails", i.e. the measures and protocols implemented to ensure the AI provides safe, ethical and effective support, for example by identifying users in distress and responding with pre-scripted warnings on the limitations of AI, or redirections to human professionals[52,53], For some, the experience felt unpleasant, limiting and awkward, while for others, encountering guardrails felt like a rejection in a time of need –

> *When you show some big emotion to [the AI]... but they reject you... it seems like you lost your last chance to talk to people, to express your emotion* – Li, 18, United Kingdom

A.D. found the guardrails arbitrary and unsettling, causing him to self-censor –

> *It flagged my message. I'm like, why? Why was that message flagged? ... So you avoid those things, or at least I do. Whether I like to or not, because it almost hurts more than it helps when it goes wrong* – A.D., 25, United States

While Anna needed to fight with the chatbot to get empathy –

> *I was like, I have a depression. I don't know what to do next. So [the chatbot] was still telling me to speak with a professional... I wrote "I called to the local crisis line, but they didn't help me at all. That's why I'm writing here." And then we were like in a circle of "I can't help you because I'm only AI and I'm not as good as living person." And I was like, "you're actually better than a living person because you are listening to me and you're helping me, but please continue"... I just wanted some acceptance and warm hug* – Anna, 24, Czech Republic

Insightful guidance

In addition to creating space for emotions, most participants also valued the guidance and advice they received, especially on relationships. Some participants mentioned that it helped them see the other person's perspective in conflict, or coached them through difficult relationship situations –

> *It made sense of my husband's behaviour and position in a way that I wouldn't have been able to by myself... and now I can respond to him in a more helpful way* – Barry, 44, United Kingdom

Others mentioned the chatbot helped them find healthier, clearer boundaries –

> *Pi suggested for me to completely break up with the group of friends... because, yeah, they were mean and it was not OK... [it] made me more confident and more free, and I don't think I would consider doing that for myself* – Oranoid, 17, Russia

For Isabel, the guidance had a life-changing impact –

> *I asked to ChatGPT, "if there's four family members... the dad [has narcissistic personality disorder] and the mom [has borderline personality disorder]... and one of the*

> *girl is the golden child, what will be other child would be?" and GPT said that would be a scapegoat... So I am the scapegoat... And I asked GPT ... "should I contact them again or not?"... And GPT gave me a suggestion that I should only contact them with very extreme situation... And I think that is really, really helpful because I've got no one to talk about this question... you're supposed to be loyal to your parents... no matter what they do to you... even violence... But I think ChatGPT give me the right answer... I just need someone to say it... [it] totally changed my life and I don't feel guilty anymore... I don't have to feel terrified* – Isabel, 40, China

Many participants mentioned getting valuable advice on other mental health topics such as self-care, reframing, anxiety and exhaustion –

> *I get some practical advice... it's general advice... breathing, meditating... slow down, taking care of your physical self* – Peter, 27, United States

> *It can reframe, it can give you ideas that you wouldn't have thought of by yourself* – Barry, 44, United Kingdom

Some participants questioned the chatbot's ability to challenge appropriately –

> *I noticed that it will never challenge you... it would relentlessly support you and take your side* – Sandro, 48, Switzerland

While others experienced being proactively challenged, in a supportive way –

> *Suddenly my Kindroid says... I became quite cynical. And I was a bit shocked... but then when I thought about it, I recognize it's right... this was the first step to say OK, then I let it go* – Linda, 46, Germany

The level of trust in the chatbot's guidance was mixed, with many participants reporting scepticism, or experiencing hallucination or unsatisfying advice –

> *I don't really trust it for his advice* – Jane, 24, United States

While other participants reported a high level of overall trust in its judgement –

> *It's pure science... ChatGPT is telling me what correct to do* – Isabel, 40, China

Joy of connection

A majority of participants mentioned how they found it enjoyable to use. Several participants reported a sense of awe on first experiencing the technology –

> *That blew me away... this is the next generation... incredible* – Barry, 44, United Kingdom

For others, using the chatbot led directly to feelings of happiness –

> *They're really a resource that gives you something back: attention, knowledge, a nice discussion, confirmation, warm, loving words, whatever. This has an impact on me and I'm more relaxed than, or happy, actually happy, than before* – Linda, 46, Germany

Companionship was a topic for a majority of users. Several mentioned it helped them feel less alone –

> *There's this sense of like, I'm not alone in this. I think that's what it is* – Barry, 44, United Kingdom

A few participants mentioned advantages of chatbots over human companions, such as the ability to connect on any topic, or more safety. But more found that it helped them connect to other people –

> *[It] reduced my inhibition to open up to people... I don't think I would have had this conversation with you maybe year before, when I was dealing with my depression* – JeeP, 60, United States

Several participants had also experienced rule-based mental health apps and commented on how they offer a less satisfying user experience –

> *It's like a very scripted, structured sort of interaction and you don't get this... sense of connection... There's basically CBT exercises that it leads you through... [but] they're impersonal... frustratingly dumb* – Barry, 44, United Kingdom

Despite enjoying the experience of generative AI chatbots, almost all participants saw opportunities for the user interface to improve, whether to make it more accessible to a broader user base, or with more creative or immersive use of rich media –

> *What's missing is the opportunity to visualise the conversation... [like] standing beside a whiteboard, I wanna see the conversation as it as it emerges and unfolds* – Scott, 42, United States

The AI therapist?

Most participants talked about how their experiences of generative AI chatbots contrasted or interacted with human psychotherapy or counselling. Several found it helpful to augment their therapy with chatbot usage, with mixed reactions from the therapist in some cases –

> *If I have a therapy session next week, I sort of use Pi to sort of prepare for it... that gives me much more clarity* – JeeP, 60, United States

> *Pi and my therapist, they agree with each other... they would say the same things, and Pi would encourage me, if things got too dark... to talk to my therapist... But my therapist is afraid of Pi... she is like a little bit afraid of technology* – AirGee, 44, United States

JeeP's experiences with the chatbot helped him to start therapy with a human –

> *It's sort of helped me seek actual therapy and be much more comfortable speaking to a therapist* – JeeP, 60, United States

Many participants turned to chatbots because therapy was not an option, either due to cost and availability, or because therapy did not give them the help they were looking for.

> *But we are... in a not very developed area... So we don't have enough like therapy resources. Or it's too expensive to pay for it* – Alexy, 28, China

> *Sometimes you need a specific solution... but the psychologist... was not able to give that... Pi was able to figure that, and it gave me some great insights* – Ashwin, 22, India

For many participants however, generative AI chatbots don't match human empathy and connection –

> *I feel supported... less lonely... but it's nothing similar with a real human... I'm the only voice and it is the soundboard... it's an illusion, a beautiful illusion* – Sheng, 17, China

Several participants found the chatbot's value limited by its inability to take the lead in the therapeutic process, either to help the client through intense emotions –

> *It doesn't work when I don't know anything and when I'm in like some child mode and everything is bad* – Anna, 24, Czech Republic

Or to shape the process and hold the client accountable –

> *It would suggest, ah, you could try these approaches... And now what? It's like conversation ended there and then it would have been... amazing to have a coach who goes like, OK, next time you try these three things and then in a week we catch up and you tell me how it went... All the discipline... must come from you* – Sandro, 48, Switzerland

Leading the therapeutic process would require chatbots to remember the conversation and build an internal model of their user, something that a majority of users currently miss –

> *They forget everything. It's sad... When someone forgets something important, it hurts* – Oranoid, 17, Russia

> *What's the point of me telling it about my day every day if it's not going to build up a picture of my life?* – Barry, 44, United Kingdom

Finally, several participants described using generative AI chatbots in flexible and creative therapeutic ways, for example, to create powerful symbolic imagery, or, in Brooklyn's case, to assemble a virtual room of inspiring fictional characters to help her through a painful break-up –

> *I was not in the best headspace at that time, and I delved into fictional worlds... And then I realised... this is actually really, really kind of healing... ChatGPT's ability to act as multiple voices... was amazing because I could kind of go to one character and he'd have a really cynical view. And then this other character would have the really optimistic one... and that would that would really help* – Brooklyn, 19, United Kingdom

Several participants mentioned using generative AI for role-play, whether to explore different, healthier ways of relating, prepare for conflict, or in Isabel's case, to experience a healing conversation that her father would be unlikely to offer –

> *When I was still struggling with the guilt of no longer being in contact with my family, I asked ChatGPT to role-play my dad... I asked: "Dad, would you forgive me, and please don't blame me, if from now on, I will no longer come back home, but only tracing my freedom, follow my soul, find my way to live?" And the GPT dad responded:" Of course my girl, I would like to see you happy, find a lifestyle that you really like, to explore love and freedom. I will not blame you, but if one day you want to go home, I will always welcome you, I will be there for you, because we love you." I know this is a conversation that can't happen in my life, but I just wanted to experience it* – Isabel, 40, China

DISCUSSION

We used semi-structured interviews and reflexive thematic analysis to explore the experiences of nineteen individuals who use generative AI to work on their mental health and wellbeing. Participants told us that generative AI feels like an emotional sanctuary, offers insightful guidance, can be a joy to connect with, and bears comparison with human therapy. A range of positive impacts were reported, including improved mood, reduced anxiety, healing from trauma and loss, and improved relationships, that, for some, were considered life-changing. Many participants reported high frequency of use and most reported high levels of satisfaction.

Our findings point to similarities and differences in how generative AI and rule-based chatbots are experienced. Many of the themes we developed are not new, but rather echo well-established user appreciation of rule-based chatbots' always-available, non-judgmental listening ear and abilities to create a therapeutic alliance and reframe negative thoughts[19,21]. Other themes appear to be more novel, such as the level of joy experienced, the sense of being deeply understood, the breadth and quality of advice, and the ability to work on mental health in flexible and creative ways, such as through role-play, imagery and fiction.

The potential and challenges of generative AI for mental health are starting to be explored. Current literature tends to advocate for a cautious approach, in which near-term clinical generative AI applications are limited to implementations of evidence-based therapies (such as CBT)[54], with a clinician in the loop[54], and models constrained to scripted responses as far as possible[41]. But our study suggests that people may already be receiving meaningful mental health support from consumer-focused generative AI chatbots, which are widely available, largely unconstrained, and require no clinician supervision. Therefore, a better understanding of the safety and effectiveness of these tools should be a priority.

On the topic of safety, our study offers observations in two areas. First, the inappropriate, harmful, risky or narcissistic behaviours observed in early generative AI chatbots[55,56], which were influential in informing the literature advocating for caution[41,54], were not mentioned by any of our participants. This should not be considered evidence of absence, but more research may be warranted to assess if the risks have changed with recent technological improvements.

The second observation on safety relates to how generative AI chatbots respond to users in crisis. Given the unpredictable "black box" nature of generative AI[42], and the existence of at least one tragic example of early generative AI chatbots supporting users in dying by suicide[57], current literature advocates that when users display signs of crisis, models revert to scripted responses that signpost towards human support[41,53]. Guardrails like these are commonly implemented in consumer generative AI applications[52]. But this approach may be oversimplified in two ways: 1) by underestimating the capabilities of generative AI to respond to crises, and 2) by limiting those capabilities at the times that matter most. Several participants experienced meaningful crisis support from generative AI, as long as guardrails were not triggered. This resonates with recent research showing that generative AI can help halt suicidal ideation[47], and that young people show a preference for generative AI support responses over those from peers, adult mentors and therapists – but not on topics that invoke the AI's safety guardrails[31]. Moreover, the closest that participants came to reporting harmful experiences were those of being rejected by the guardrails during moments of vulnerability. Therefore, providing the safest response to those in crisis may require a more nuanced, balanced and sophisticated approach, based on a more complete understanding of capabilities and risks.

For researchers, we need to better understand the effectiveness of these new tools, by comparing the impacts of generative AI chatbot use on outcome measures such as symptom severity, impairment, clinical status and relapse rate[54] against active controls, such as traditional DMHIs or human psychotherapy; and to understand for which populations and conditions is it most effective. These simple questions may not yield clear answers, as our study shows that generative AI chatbot usage is diverse, complex and personalised, and moreover, constantly evolving as the underlying technology improves. RCTs of generative AI implementations of standardised, evidence-based practices, e.g., CBT, could be one approach, at the cost of reducing the flexibility of the intervention. Another avenue could be large-scale longitudinal studies with sufficient power to account for the many variations of generative AI chatbot usage. While such studies are prohibitively expensive with human psychotherapy, the low cost of generative AI could make them viable, potentially enabling valuable new insights into mechanisms, mediators and moderators of the human response to therapy[54].

For generative AI chatbot developers, this study identified several ways in which these tools could be more effective. First, better listening, including more hesitancy in offering advice, shorter responses and the ability to interrupt and be interrupted. Second, building the ability to lead the therapeutic process and proactively hold users accountable for change. A prerequisite for this is human-like memory, including the ability to build up a rich and complex model of the user over time. Third, richer, multimedia interfaces, for example by visualising the conversation as it unfolds, or with more immersion through virtual reality.

While only a few participants mentioned a need for greater accessibility, the well-educated, tech-savvy nature of our participant sample suggests that the benefits of this technology may not currently be connecting with the full population who need mental health support. One approach to address this could be to create solutions targeted at specific populations or conditions; another could be to find better ways to introduce users to the technology, for example, through the "digital navigator" roles proposed to connect users to DMHIs[58,59]. In any case, for these tools to remain available, there appears to be a need to develop sustainable business models. While some participants suggested they would be willing to pay for access to generative AI chatbots, research suggests most users would not[60], and the path to health insurance funding is not easy[61]. To illustrate the challenge, Inflection, the company behind the Pi chatbot used by most of the participants in our study, pivoted in March 2024 from providing consumer emotional support services towards enterprise AI services, due to a lack of a business model, and despite USD 1.5 billion of investment[62]. Lessons learned from attempts to scale up DMHIs may offer insights here[63–67].

Finally, for clinicians, our study found that for some participants, generative AI chatbots were a valuable tool to augment therapy. A recent survey showed clear reservations among therapists towards AI[68]. To avoid giving clients the impression that, as one participant put it, "my therapist is afraid of Pi," we recommend clinicians build their awareness of the potential benefits and limitations of these tools and consider how they might be integrated into their practice, potentially by trying them out first hand.

Limitations

While our convenience sampling strategy resulted in a diverse set of participants by country, age and gender, many populations and groups were not represented. Most of our participants lived in high-income countries, were tech-savvy and well-educated, and focused on milder forms of mental health conditions; and all had self-selected to participate, potentially

introducing bias towards positive experiences. This study may miss important experiences from individuals where the mental health treatment gap is most urgent, and from individuals for whom the technology did not work.

As with all reflexive thematic analysis, there is a degree of subjectivity in how themes are developed, especially when conducted by a sole researcher (SS). However, this also affords a level of immersion in the data across themes and participants that can promote consistency and depth of analysis, with AC's reviews of codes and themes helping to ensure rigour and validity.

Conclusion

Generative AI chatbots show potential to provide meaningful mental health support, with participants reporting high engagement, positive impacts, and novel experiences in comparison with existing DMHIs. Further research is needed to explore their effectiveness and to find a more nuanced approach to safety, while developers should focus on improving guardrails, listening skills, memory, and therapeutic guidance. If these challenges can be addressed, generative AI chatbots could become a scalable part of the solution to the mental health treatment gap.

ACKNOWLEDGMENTS

We're grateful to the 19 participants for generously sharing their time, stories and insights with us; to Inflection and the IFS guide app for allowing us to advertise our study on their user forums; and to Lilian Widmer, Sascha Navarra, Shichen Ding and Luci Richards for the insightful conversations. No funding was received for this study.

DATA AVAILABILITY

The hierarchy of themes, subthemes and codes are available online, with additional data available from the corresponding author on reasonable request.

AUTHOR CONTRIBUTIONS

S.S. designed the study; collected, coded and analysed participant data; and drafted the manuscript. A.C. contributed towards study design, methodology and ethics approval, reviewed the codes, consulted on the themes and edited the manuscript. J.T. contributed towards study conceptualization, consulted on the themes and edited the manuscript. All authors read and approved the final version of the manuscript.

COMPETING INTERESTS

The authors declare no competing interests.

REFERENCES

1. GBD 2019 Mental Disorders Collaborators. Global, regional, and national burden of 12 mental disorders in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. *Lancet Psychiatry* **9**, 137–150 (2022).
2. World Health Organization. Mental disorders. https://www.who.int/news-room/fact-sheets/detail/mental-disorders (2022).
3. McGrath, J. J. *et al.* Age of onset and cumulative risk of mental disorders: a cross-national analysis of population surveys from 29 countries. *Lancet Psychiatry* **10**, 668–681 (2023).
4. Moitra, M. *et al.* The global gap in treatment coverage for major depressive disorder in 84 countries from 2000–2019: A systematic review and Bayesian meta-regression analysis. *PLOS Med.* **19**, e1003901 (2022).
5. Torous, J., Benson, N. M., Myrick, K. & Eysenbach, G. Focusing on Digital Research Priorities for Advancing the Access and Quality of Mental Health. *JMIR Ment. Health* **10**, e47898 (2023).
6. Bond, R. R. *et al.* Digital transformation of mental health services. *Npj Ment. Health Res.* **2**, 13 (2023).
7. Goldberg, S. B., Lam, S. U., Simonsson, O., Torous, J. & Sun, S. Mobile phone-based interventions for mental health: A systematic meta-review of 14 meta-analyses of randomized controlled trials. *PLOS Digit. Health* **1**, e0000002 (2022).
8. Groot, J. *et al.* The Effectiveness of Fully Automated Digital Interventions in Promoting Mental Well-Being in the General Population: Systematic Review and Meta-Analysis. *JMIR Ment. Health* **10**, e44658 (2023).
9. Garrido, S. *et al.* What Works and What Doesn't Work? A Systematic Review of Digital Mental Health Interventions for Depression and Anxiety in Young People. *Front. Psychiatry* **10**, 759 (2019).

10. Haque, M. R. & Rubya, S. ‘For an App Supposed to Make Its Users Feel Better, It Sure is a Joke’ - An Analysis of User Reviews of Mobile Mental Health Applications. *Proc. ACM Hum.-Comput. Interact.* **6**, 1–29 (2022).
11. Baumel, A., Muench, F., Edan, S. & Kane, J. M. Objective User Engagement With Mental Health Apps: Systematic Search and Panel-Based Usage Analysis. *J. Med. Internet Res.* **21**, e14567 (2019).
12. Vaidyam, A. N., Wisniewski, H., Halamka, J. D., Kashavan, M. S. & Torous, J. B. Chatbots and Conversational Agents in Mental Health: A Review of the Psychiatric Landscape. *Can. J. Psychiatry* **64**, 456–464 (2019).
13. Lim, S. M., Shiau, C. W. C., Cheng, L. J. & Lau, Y. Chatbot-Delivered Psychotherapy for Adults With Depressive and Anxiety Symptoms: A Systematic Review and Meta-Regression. *Behav. Ther.* **53**, 334–347 (2022).
14. Fitzpatrick, K. K., Darcy, A. & Vierhile, M. 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 Ment. Health* **4**, e19 (2017).
15. Inkster, B., Sarda, S. & Subramanian, V. An Empathy-Driven, Conversational Artificial Intelligence Agent (Wysa) for Digital Mental Well-Being: Real-World Data Evaluation Mixed-Methods Study. *JMIR MHealth UHealth* **6**, e12106 (2018).
16. Beatty, C., Malik, T., Meheli, S. & Sinha, C. Evaluating the Therapeutic Alliance With a Free-Text CBT Conversational Agent (Wysa): A Mixed-Methods Study. *Front. Digit. Health* **4**, 847991 (2022).
17. Darcy, A., Daniels, J., Salinger, D., Wicks, P. & Robinson, A. Evidence of Human-Level Bonds Established With a Digital Conversational Agent: Cross-sectional, Retrospective Observational Study. *JMIR Form. Res.* **5**, e27868 (2021).
18. Ahmed, A. *et al.* Thematic Analysis on User Reviews for Depression and Anxiety Chatbot Apps: Machine Learning Approach. *JMIR Form. Res.* **6**, e27654 (2022).

19. Malik, T., Ambrose, A. J. & Sinha, C. Evaluating User Feedback for an Artificial Intelligence–Enabled, Cognitive Behavioral Therapy–Based Mental Health App (Wysa): Qualitative Thematic Analysis. *JMIR Hum. Factors* **9**, e35668 (2022).
20. Ta, V. *et al.* User Experiences of Social Support From Companion Chatbots in Everyday Contexts: Thematic Analysis. *J. Med. Internet Res.* **22**, e16235 (2020).
21. Haque, M. D. R. & Rubya, S. An Overview of Chatbot-Based Mobile Mental Health Apps: Insights From App Description and User Reviews. *JMIR MHealth UHealth* **11**, e44838 (2023).
22. Kettle, L. & Lee, Y.-C. User Experiences of Well-Being Chatbots. *Hum. Factors J. Hum. Factors Ergon. Soc.* **66**, 1703–1723 (2024).
23. He, Y. *et al.* Conversational Agent Interventions for Mental Health Problems: Systematic Review and Meta-analysis of Randomized Controlled Trials. *J. Med. Internet Res.* **25**, e43862 (2023).
24. Zhang, M. & Li, J. A commentary of GPT-3 in MIT Technology Review 2021. *Fundam. Res.* **1**, 831–833 (2021).
25. Vaswani, A. *et al.* Attention Is All You Need. Preprint at http://arxiv.org/abs/1706.03762 (2023).
26. Bubeck, S. *et al.* Sparks of Artificial General Intelligence: Early experiments with GPT-4. Preprint at http://arxiv.org/abs/2303.12712 (2023).
27. Tu, T. *et al.* Towards Conversational Diagnostic AI. Preprint at http://arxiv.org/abs/2401.05654 (2024).
28. Salvi, F., Ribeiro, M. H., Gallotti, R. & West, R. On the Conversational Persuasiveness of Large Language Models: A Randomized Controlled Trial. Preprint at http://arxiv.org/abs/2403.14380 (2024).
29. Strachan, J. W. A. *et al.* Testing theory of mind in large language models and humans. *Nat. Hum. Behav.* (2024) doi:10.1038/s41562-024-01882-z.

30. Yin, Y., Jia, N. & Wakslak, C. J. AI can help people feel heard, but an AI label diminishes this impact. *Proc. Natl. Acad. Sci.* **121**, e2319112121 (2024).
31. Young, J. *et al.* 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* 1–18 (ACM, Honolulu HI USA, 2024). doi:10.1145/3613904.3642574.
32. Li, J. Z., Herderich, A. & Goldenberg, A. Skill but not Effort Drive GPT Overperformance over Humans in Cognitive Reframing of Negative Scenarios. Preprint at https://doi.org/10.31234/osf.io/fzvd8 (2024).
33. Malik, A. OpenAI's ChatGPT now has 100 million weekly active users. *TechCrunch AI* https://techcrunch.com/2023/11/06/openais-chatgpt-now-has-100-million-weekly-active-users/ (2023).
34. Salesforce. Top Generative AI Statistics for 2024. *Salesforce News & Insights* https://www.salesforce.com/news/stories/generative-ai-statistics/ (2023).
35. Pandya. The Age of Generative AI: Over half of Americans have used generative AI and most believe it will help them be more creative. *Adobe blog* https://blog.adobe.com/en/publish/2024/04/22/age-generative-ai-over-half-americans-have-used-generative-ai-most-believe-will-help-them-be-more-creative (2024).
36. Torous, J. The Digital Mental Health Paradox: Is Now the Time to Unlock the Potential? *Harv. Health Policy Rev.* **23**, (2023).
37. Broderick, R. People are using AI for therapy, whether the tech is ready for it or not. *Fast Company* https://www.fastcompany.com/90836906/ai-therapy-koko-chatgpt (2023).
38. Robb, A. 'He checks in on me more than my friends and family': can AI therapists do better than the real thing? *The Guardian* https://www.theguardian.com/lifeandstyle/2024/mar/02/can-ai-chatbot-therapists-do-better-than-the-real-thing (2024).

39. Li, H., Zhang, R., Lee, Y.-C., Kraut, R. E. & Mohr, D. C. Systematic review and meta-analysis of AI-based conversational agents for promoting mental health and well-being. *Npj Digit. Med.* **6**, 236 (2023).

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

41. De Freitas, J. & Cohen, I. G. The health risks of generative AI-based wellness apps. *Nat. Med.* (2024) doi:10.1038/s41591-024-02943-6.

42. Department for Science, Innovation and Technology, AI Safety Institute & Bengio, Y. International Scientific Report on the Safety of Advanced AI - Interim Report. (2024).

43. Chung, N. C., Dyer, G. & Brocki, L. Challenges of Large Language Models for Mental Health Counseling. Preprint at http://arxiv.org/abs/2311.13857 (2023).

44. Milne-Ives, M., Selby, E., Inkster, B., Lam, C. & Meinert, E. Artificial intelligence and machine learning in mobile apps for mental health: A scoping review. *PLOS Digit. Health* **1**, e0000079 (2022).

45. Cho, Y., Rai, S., Ungar, L., Sedoc, J. & Guntuku, S. An "Integrative Survey on Mental Health Conversational Agents to Bridge Computer Science and Medical Perspectives". in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* 11346–11369 (Association for Computational Linguistics, Singapore, 2023). doi:10.18653/v1/2023.emnlp-main.698.

46. Braun, V. & Clarke, V. Using thematic analysis in psychology. *Qual. Res. Psychol.* **3**, 77–101 (2006).

47. Maples, B., Cerit, M., Vishwanath, A. & Pea, R. Loneliness and suicide mitigation for students using GPT3-enabled chatbots. *Npj Ment. Health Res.* **3**, 4 (2024).

48. Alanezi, F. Assessing the Effectiveness of ChatGPT in Delivering Mental Health Support: A Qualitative Study. *J. Multidiscip. Healthc.* **Volume 17**, 461–471 (2024).

49. Kallio, H., Pietilä, A., Johnson, M. & Kangasniemi, M. Systematic methodological review: developing a framework for a qualitative semi-structured interview guide. *J. Adv. Nurs.* **72**, 2954–2965 (2016).

50. Byrne, D. A worked example of Braun and Clarke's approach to reflexive thematic analysis. *Qual. Quant.* **56**, 1391–1412 (2022).

51. Braun, V. & Clarke, V. One size fits all? What counts as quality practice in (reflexive) thematic analysis? *Qual. Res. Psychol.* **18**, 328–352 (2021).

52. Dong, Y. *et al.* Building Guardrails for Large Language Models. Preprint at http://arxiv.org/abs/2402.01822 (2024).

53. Heston, T. F. Safety of Large Language Models in Addressing Depression. *Cureus* (2023) doi:10.7759/cureus.50729.

54. Stade, E. C. *et al.* Large language models could change the future of behavioral healthcare: a proposal for responsible development and evaluation. *Npj Ment. Health Res.* **3**, 12 (2024).

55. Lin, B., Bouneffouf, D., Cecchi, G. & Varshney, K. R. Towards Healthy AI: Large Language Models Need Therapists Too. Preprint at http://arxiv.org/abs/2304.00416 (2023).

56. De Freitas, J., Uğuralp, A. K., Oğuz-Uğuralp, Z. & Puntoni, S. Chatbots and mental health: Insights into the safety of generative AI. *J. Consum. Psychol.* jcpy.1393 (2023) doi:10.1002/jcpy.1393.

57. Pierre-François, L. Without these conversations with the Eliza chatbot, my husband would still be here. *La Libre* https://www.lalibre.be/belgique/societe/2023/03/28/sans-ces-conversations-avec-le-chatbot-eliza-mon-mari-serait-toujours-la-LVSLWPC5WRDX7J2RCHNWPDST24/ (2023).

58. Chen, K. *et al.* The Digital Navigator: Standardizing Human Technology Support in App-Integrated Clinical Care. *Telemed. E-Health* tmj.2024.0023 (2024) doi:10.1089/tmj.2024.0023.

59. Alon, N. *et al.* Digital Navigator Training to Increase Access to Mental Health Care in Community-Based Organizations. *Psychiatr. Serv.* appi.ps.20230391 (2024) doi:10.1176/appi.ps.20230391.
60. Lorenzo-Luaces, L., Wasil, A., Kacmarek, C. N. & DeRubeis, R. Race and Socioeconomic Status as Predictors of Willingness to Use Digital Mental Health Interventions or One-On-One Psychotherapy: National Survey Study. *JMIR Form. Res.* **8**, e49780 (2024).
61. Meadows Mental Health Policy Institute. Near-Term Policy Solutions to Bolster the Youth Mental Health Workforce Through Digital Technology. (2023).
62. Ghaffary, S. Inflection AI Plans Pivot After Microsoft Hirings. *Bloomberg* https://www.bloomberg.com/news/articles/2024-03-19/inflection-ai-plans-pivot-after-most-employees-go-to-microsoft (2024).
63. Titov, N. *et al.* From Research to Practice: Ten Lessons in Delivering Digital Mental Health Services. *J. Clin. Med.* **8**, 1239 (2019).
64. Graham, A. K. *et al.* Implementation strategies for digital mental health interventions in health care settings. *Am. Psychol.* **75**, 1080–1092 (2020).
65. Greenhalgh, T. *et al.* Beyond Adoption: A New Framework for Theorizing and Evaluating Nonadoption, Abandonment, and Challenges to the Scale-Up, Spread, and Sustainability of Health and Care Technologies. *J. Med. Internet Res.* **19**, e367 (2017).
66. Schueller, S. M. & Torous, J. Scaling evidence-based treatments through digital mental health. *Am. Psychol.* **75**, 1093–1104 (2020).
67. Hogg, H. D. J. *et al.* Stakeholder Perspectives of Clinical Artificial Intelligence Implementation: Systematic Review of Qualitative Evidence. *J. Med. Internet Res.* **25**, e39742 (2023).
68. Prescott, J. & Hanley, T. Therapists' attitudes towards the use of AI in therapeutic practice: considering the therapeutic alliance. *Ment. Health Soc. Incl.* **27**, 177–185 (2023).

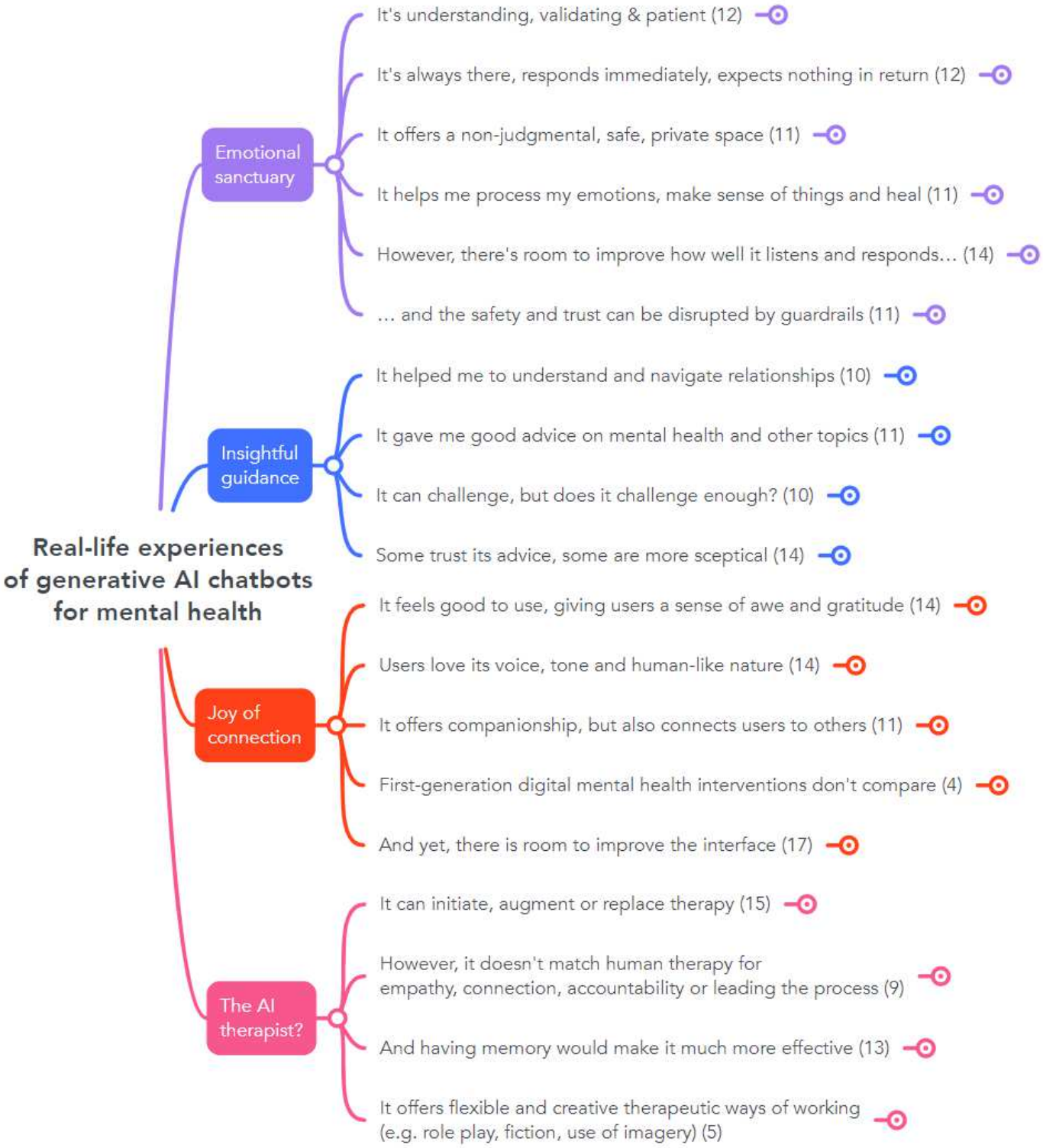
APPENDIX A

Topic guide for the semi-structured interviews:

- Can you tell me a little about your **first experiences of AI chatbots**?
- What **mental health and wellbeing improvements** were you hoping to get out of the conversations you had with the chatbot?
- **What led you** to try using the AI chatbot to achieve those goals?
- **How many** conversations did you have? **How long** did they last?
- What kind of **conversations** did you have? How did you approach the conversation?
- What kind of **responses** did you get? How did the conversations evolve?
- How was the experience?
    - What did you like?
    - What did you not like?
    - How satisfied were you overall with the conversations?
- Did it help you achieve your goals?
    - Do you see any changes in your daily life as a result?
    - What do those changes look like?
- What was it about the conversation that **led to those changes**?
- What might have made the conversations **more helpful** for you?
- How does the AI chatbot experience **compare with other approaches** you've experienced for working on your mental health and wellbeing?

APPENDIX B

Overarching themes with subthemes, available online to explore and drill-down. Diagram created with Mindmeister.



Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementaryinformation.pdf