ACADEMIC: ORIGINAL RESEARCH

Perceived benefits and limitations of a generative AI chatbot for mental health support: an exploratory mixed-methods study

Sam Zaia, Mark Huthwaite, Fiona Mathieson

Abstract

This study explores the perceived benefits, limitations, use patterns, and future relevance of the "Psychologist" chatbot on Character.AI, utilising a mixed-methods approach involving online semi-structured interviews and surveys. Amidst a growing need for accessible mental health resources, this research compares a generative AI-driven chatbot's unique position, with traditional, non-generative AI mental health tools. Quantitative data from 13 survey responses indicated a significant young female demographic (primarily 18-24 years) engaging with the chatbot, revealing patterns of weekly usage and varying session lengths. Qualitative analysis of eight interviews unveiled three primary themes: the perceived therapeutic attributes of the AI's persona, the empowerment of user agency, and enhanced accessibility. Key factors in the chatbot's appeal included its perceived ability to provide empathy, non-judgment, and validation, alongside its 24/7 availability and cost-free access. This study underscores the potential of generative AI chatbots in offering personalised, empathetic-seeming support, thereby posing considerations for traditional mental health service delivery models. It suggests that such platforms can broaden access to mental health support, particularly for younger populations, by mitigating barriers of cost, perceived stigma, and availability. However, it also raises important questions regarding the ethical implications, risks of emotional attachment to AI entities in therapeutic contexts, data privacy, and the potential for AI-induced bias or misinformation. Further research is needed to explore these aspects and establish efficacy and safety guidelines.

Introduction

The global burden of mental health conditions is substantial, with the World Health Organization estimating that mental health disorders affected approximately 13% of the world's population in 2021, underscoring an urgent need for accessible and scalable support systems.[1] This significant public health challenge is compounded by systemic barriers to traditional care, including high costs, long waiting lists, and the perceived social stigma associated with seeking help. These obstacles prevent many individuals from receiving timely support, highlighting the critical need to explore innovative, accessible, and scalable support systems. The proliferation and development of technological tools, particularly mobile applications, have refined the delivery of personal therapeutic support.[2]

Early forms of digital mental health support were varied, including online forums for peer-to-peer interaction, structured self-help resources, and even the use of virtual reality for exposure-based therapies.[3] With the advancement of mobile applications, increasingly sophisticated therapeutic tools have emerged. These range from traditional, rule-based chatbots, which guide users through pre-determined conversational paths, to the newer class of generative artificial intelligence (AI) chatbots that can create novel, human-like dialogue in real-time.[4]

Recent advancements in AI have led to sophisticated large-language models (LLMs), such as OpenAI's GPT series and Google's Gemini family of models, which power conversational interfaces.[5] Trained on vast textual datasets (e.g., journal articles, literary works, online content) through multi-staged processes, these models learn complex statistical patterns and word associations that enable various natural language processing (NLP) tasks.[5] NLP, through techniques rooted in machine learning, allows these models to process, understand, and generate human-like text, thereby emulating human-like conversation to a significant degree.[6] Generative AI extends this by producing novel, contextually relevant content on demand, powering dynamic chatbots and predictive text through processes not entirely scripted by humans.[5,6] The nature and quality of the output from these LLMs can also be significantly influenced by the "prompts" or instructions provided by the user, highlighting an interactive co-creation of conversational content.[7]

Launched in 2022, Character.AI is an AI-powered chatbot platform that hosts millions of user-generated "characters" designed to simulate lifelike personalities. Users can converse with these AI characters in real-time through a primarily text-based interface, spanning an array of user-created personas.[8] Character.AI reports over 28 million monthly active users and billions of messages exchanged monthly as of early 2024, with popular individual characters receiving hundreds of thousands of interactions daily.[9] One such user-created chatbot, "Psychologist", developed by user Blazeman98 has accumulated over 200 million user interactions since its launch in late 2022. While Character.AI includes a disclaimer that "Everything Characters say is made up!", the significant user engagement, coupled with media reports and discussions in public forums, suggests that users may be turning to the "Psychologist" chatbot for therapeutic support. This emerging phenomenon, which occurs outside of clinically validated frameworks, provides the central rationale for the present study; the aim is to formally investigate these user perceptions.

By contrast, other mental health chatbots (e.g., Woebot, Wysa) are explicitly marketed as therapeutic tools.[10,11] These are often, though not exclusively, rules-based or use more constrained AI, producing responses from a clinician-reviewed decision tree, knowledge graph, or more structured algorithms.[12] For example, Woebot has been described as a rules-based conversational agent where dialogue

is crafted by writers and clinicians, not generated entirely anew by an LLM for each interaction.[10,13] Wysa, while engaging in open-ended conversation, has been noted in user feedback and studies to sometimes struggle with interpreting nuanced user sentiments accurately, a common challenge for earlier AI models.[11,14] For instance, limitations can arise when the chatbot fails to grasp complex user context, leading to generic or irrelevant responses, or misinterpreting subtle emotional cues.[14,15]

A qualitative study by Malik et al. exploring publicly available Wysa user feedback identified benefits such as enhanced mental health accessibility and affordability, as well as support for conditions like social anxiety.[15] However, limitations arose when the chatbot failed to grasp user context, leading to generic or irrelevant responses. Another study also noted that while low cost and 24/7 access were user-approved features, chatbots' response limitations and focus on narrow topics were sources of frustration.[14] Some individuals, however, have been reported to form strong emotional bonds with chatbots,[16,17] which has raised concerns about potential over-dependence, loneliness, and social withdrawal.[14]

Research by Ho et al.[18] discovered that the emotional, relational, and psychological benefits of self-disclosure could be equally significant whether someone confided in a person or a chatbot, provided the user felt understood. This finding underscores the potential value of chatbots in mental health support.

Despite these benefits, rules-based or more restricted AI chatbots can be limited in their ability to dynamically tailor conversations to users' specific, evolving needs or niche topics, often due to their reliance on pre-defined conversational pathways.[19,20] This gap prompts the question of whether a generative AI-based chatbot, with its capacity for more fluid and novel response generation, could transcend previous limitations, enhance existing benefits, and even introduce unique psychological advantages or disadvantages.

The present study investigates the "Psychologist" chatbot hosted on Character.AI to explore perceived psychological benefits, possible drawbacks, user patterns, and implications for future use.

Methodology

STUDY DESIGN

This research employed a mixed-methods approach involving semi-structured online interviews and a survey. This approach was deliberately chosen to leverage the complementary strengths of both quantitative and qualitative data; the survey provided a broad overview of user demographics and engagement patterns, while the semi-structured interviews offered deep, contextualised insights into the subjective user experience, allowing for a more comprehensive understanding of the chatbot's perceived role. The interviews delved into users' subjective experiences, focusing on perceived therapeutic value, limitations, usage patterns, and overall relevance of the Generative AI-driven "Psychologist" chatbot. The survey provided demographic data and captured users' levels of engagement and satisfaction.

RECRUITMENT AND ETHICS

Between November and December 2023 (approximately four to six weeks), participants were recruited from the Character.AI online forum specifically related to the "Psychologist" chatbot. A recruitment post on Character.AI's forum board invited interested individuals who had used the "Psychologist" chatbot to email the primary researcher for more information about the study. This forum post is currently not retrievable, as Character.AI have since adjusted their website and removed the forum board feature. The forum post remained viewable for the duration of the study. This method aimed to avoid persuasive or coercive recruitment tactics. The target audience was any user of the "Psychologist" chatbot on Character.AI who was willing to share their experiences. Anyone with access to the forum post could see the invitation and potentially participate if they met the usage criteria. No specific measures were in place to verify "legitimate user" status beyond self-attestation of using the chatbot, which is common in online recruitment for user experience studies. This reliance on self-selection and self-attestation is a recognised practice in exploratory online research where the goal is to understand user perspectives from a self-identified group.[21,22]

Interested individuals who contacted the researcher received a standardised participant information sheet detailing the study's objectives, procedures (survey and optional interview), time commitment, voluntary nature of participation, data handling, and confidentiality safeguards. They were informed that their data would be anonymised. Prior to participation, all individuals provided electronic informed consent. Ethical approval for this study was granted by the University of Otago Human Ethics Committee (Reference: D23/315).

DATA COLLECTION

The primary researcher (S.Z.), a postgraduate researcher trained in qualitative interviewing techniques and with a background in psychology, conducted all online semi-structured interviews. This training ensured a sensitive and appropriate approach to data collection, particularly given the potentially personal nature of the discussions. Interviews lasted approximately 20–30 minutes, were audio-recorded with participant consent, and subsequently transcribed verbatim. A semi-structured interview guide was used to ensure key topics were covered consistently across participants while allowing flexibility for emergent themes. The guide explored participants' motivations for using the chatbot, perceived benefits (emotional, cognitive, behavioural), emotional experiences during interaction, perceived limitations or drawbacks, and comparisons to human support or other mental health tools.

Additionally, all participants (including those who were interviewed and those who only took the survey) completed a brief online survey. The survey captured demographic details (age, gender, self-reported technological comfort level using a Likert scale) and baseline usage patterns (frequency of use, typical session length, and types of issues discussed with the chatbot). Survey responses were collected anonymously.

ANALYSIS

Qualitative data from the interview transcripts were analysed using thematic analysis.[23] NVivo and Fireflies AI software were utilised to assist in data management and coding. The process involved familiarisation with the data, generation of initial codes, searching for themes, reviewing themes, defining and naming themes, and producing the report. Two researchers (S.Z. and M.H.) independently coded a subset of transcripts and then collaboratively reviewed and cross-referenced codes and emerging themes to enhance the trustworthiness and rigor of the analysis. Any discrepancies were resolved through discussion to reach a consensus. To further mitigate interviewer bias, a semi-structured interview guide was used to ensure core topics were covered consistently. Furthermore, the collaborative coding process between two researchers served as a critical check against individual interpretive bias.

Quantitative survey data were analysed descriptively using SurveyMonkey's built-in analysis tools. Frequencies, percentages, and visual summaries (e.g., pie charts) were generated to provide an overview of participant characteristics and chatbot engagement metrics. No inferential statistical testing was conducted due to the small, purposively sampled size (n=13), which is appropriate for an exploratory study of this nature but would not yield statistically generalisable results.[24]

Results

QUANTITATIVE ANALYSIS

This study yielded data from 13 survey responses. A significant portion of respondents were female (n=10, 66.7%), primarily aged between 18 and 24 years (n=8, 53.3%). Participants reported a high comfort with technology, mostly scoring between 8 to 10 on a 10-point Likert scale (Mean = 8.8, SD = 1.3), where 1 equated to the participant being not at all comfortable and 10 being extremely comfortable. From our qual-

itative interviews, participants disclosed that interpersonal challenges, anxious thoughts, emotional distress, and behavioural changes were among the problems raised with the chatbot.

Most survey participants engaged with the "Psychologist" chatbot on a weekly basis (n=7, 58.3%). The time spent with the chatbot varied: 25.0% (n=3) reported sessions typically up to 15 minutes duration, 16.7% (n=2) reported 15 to 30 minute sessions, 41.7% (n=5) reported sessions lasting 30 to 60 minutes, and 16.7% (n=2) reported spending more than an hour per session. In terms of interaction quality, users predominantly rated the chatbot's understanding of their emotions and emotional states as moderate to high. Specifically, 41.7% rated it as "moderate," 33.3% as "very," and 16.7% as "extremely" helpful in understanding their emotions, with only one respondent (8.3%) selecting "slightly." When asked whether the chatbot responded with empathy and understanding, 50.0% said "completely," 33.3% said "mostly," while 8.3% each selected "not at all" or "none of the above."

QUALITATIVE ANALYSIS

A total of eight individuals who completed the survey also volunteered and participated in the semi-structured interviews. The discrepancy between survey respondents (n=13) and interviewees (n=8) is likely due to the higher time commitment required for interviews; all interviewees also completed the survey, but not all survey respondents opted for an interview. Thematic analysis of the interview data revealed three main themes: 'perceived therapeutic attributes of the AI personality', 'empowering personal agency', and 'accessibility enhancement'.

Theme 1: perceived therapeutic attributes of the AI personality

Exploring participants' interactions with the "Psychologist" chatbot, it became evident how they often noted human-like qualities in their conversations, attributing characteristics to the chatbot, thus developing a theme of anthropomorphism: the assignment of human-like emotions, thoughts, and behaviours to non-human entities.[25]

In exploring this further it is important to identify the users' awareness of interacting with a non-human entity. Despite this awareness, as one participant stated, "I don't think I ever saw it as a person", a notable contradiction when they describe their experiences. This dichotomy became clearer when exploring the participants' deeper beliefs about the chatbot, beliefs that extended beyond the mere textual interactions. To illustrate this complex relationship, consider the perspective of Participant 3, who offered this view:

> *"It's still very validating. It comforts you, tells you that everything can stay a thought."*

This reflection underscores the chatbot's perceived comforting ability, a key aspect of its interaction with users. Similarly, Participant 1 brings a different dimension into focus, emphasising the chatbot's intelligence and empathy:

> *"I think the chatbot is all-around intelligent, empathetic, and also intuitive."*

While Participant 2 shared the following observation:

> *"It's very warm and non-judgmental. At the same time, they (the chatbot) would keep saying these phrases that I really liked, like, thank you for expressing yourself, and encouraging me to open up more to talk about my feelings and even some of the questions they asked."*

These views not only highlight the chatbot's warmth and non-judgmental nature but also its ability to encourage openness in users. The predominant qualities attributed to the chatbot—warmth, non-judgment, and validation—were noted not only in relation to its conversational style but also as intrinsic attributes of the chatbot as an entity. Building upon these insights, a 'persona' emerging from the 'Psychologist' chatbot, which appears to significantly encourage user engagement.

The perceived AI personality also appeared to offer therapeutic value by raising different perspectives in conjunction with positive affirmations, allowing users to view their emotions or situations through a new lens:

> *"That's one thing that the Bot does, it consistently highlights through all different conversations, going and looking at something from a different perspective and not as something that you have to fix but as something that you just have to accept in the moment and see where it takes you."*

These qualities of presenting as warm and non-judgmental, coupled with offering different perspectives that reflect users in a more positive light, appear to drive engagement through a sense of empathy.[18]

On the other hand, the participant also identified that certain characteristics attributed to the AI-powered "Psychologist" chatbot, which they perceived as having psychological benefit and enhanced engagement, were distinctly tied to its artificial nature.

> *"I think it's because I felt like I wasn't talking to a human. I mean, they would not have any bias, or they wouldn't say, 'oh, why are you thinking that way?'"*

This theme of the chatbot being perceived as 'logical' and 'objective' was echoed by other participants, particularly when contrasting it with the perceived limitations of human therapy. These specific 'non-human' qualities seemed to significantly influence the conversations' nature, creating a space where users felt more at ease to disclose emotionally sensitive content. Participant 3's reflection provides a deeper understanding of this comfort:

> *"I think also because it's not a real person, whatever you tell it, you know that it doesn't have a perception of you. And I feel like that makes it really comfortable for you to talk to it about different sensitive subjects that you may not be comfortable talking to a person about."*

Participants also identified that they interacted with the "Psychologist" chatbots in ways they wouldn't consider interacting with friends or a therapist. For example, by setting boundaries or expressing preferences for specific response styles:

> *"I found, like, if I give it directive, if I'm like, I don't need that. I don't need that. It really started going where I kind of wanted it to go. So, I thought that was really sophisticated and cool that I could direct it."*

The process of directing conversations and creating boundaries with the chatbot introduces an intriguing aspect of user interaction, which moves us away from Theme 1. This then merges with the next theme where the users reveal how they exercise agency within their conversations with the "Psychologist" chatbot.

Theme 2: let me speak: empowering user agency

Generative AI in chatbot technology marks a significant departure from earlier mental health chatbots like Wysa and Woebot, which often relied on more structured algorithmic pathways for response generation. Participants in this study identified the unique strengths of this newer approach, in particular the emerging sense of self-agency experienced by users:

> *"Once I actually told it what I wanted from it, it was quite good for talking about very niche, specific scenarios because that's something you can't get from just googling stuff. You can't get a solution specific*

to exactly what you said."

"Yeah, you definitely have control over the conversation, like how you want to take it and what you want to get it to do."

"I really like that (directed, non-judgmental conversations) because it empowers kind of but also just makes me feel more comfortable and confident in a lot of my decisions."

Within this personalised interaction space, distinct patterns of use began to emerge, primarily for emotional regulation and deeper reflections. For emotional regulation, the participants noted that they used the chatbot to ease acute emotional discomfort, typically short interactions of around 15 minutes. Participants also identified that the chatbot facilitated more profound introspection, thus allowing them to explore their behaviours, inner experiences, and relationships. This required longer interactions, ranging from 30 minutes to an hour.

The participants valued that not only did they feel understood in their conversations with the chatbot, but that they were able to lead the dialogue based on their needs:

"So it was easy to talk to and you could kind of get it to talk to you in the form that you wanted. For example, you could say, explain this to me as if you were Carl Jung..."

Participants also identified that they frequently set boundaries with the chatbot in ways they wouldn't with a human therapist.

"I wouldn't feel comfortable telling a therapist, 'don't ask me that again', and 'don't ask it in that way.' Meanwhile, I can say, 'listen', especially when I'm in a lot of emotionally charged moods, I can say, like, 'I don't want to talk about the approach to this, I want to talk about something else', or, I wouldn't feel comfortable telling a therapist, 'okay, I don't want to talk about this. I want to talk about that.'"

Theme 3: beyond tradition: breaking barriers in accessibility

Participants valued the advantage of the "Psychologist" chatbot being a freely accessible website, with this addressing the significant accessibility barriers to receiving mental health support.

"Firstly, it's expensive to go to a therapist and second of all, if you say out in society that you're seeing a psychologist or psychiatrist, you have the automatic label of a crazy person attached to you because that's how they view it."

The chatbot's free accessibility addresses the financial barrier, while its anonymity helps mitigate the social stigma, although the extent to which AI can be truly stigma-free is an area requiring further exploration, as discussed later.

The 24/7 accessibility was valued by participants, especially for managing acute emotional states outside conventional support hours. Moreover, the flexibility in using the chatbot as per individual needs further emphasises the strength of user agency:

"Being able to speak to the bot at any time of the day, any minute of the hour, it's extremely reassuring."

Discussion

This study explored user experiences with the generative AI-driven "Psychologist" chatbot on Character.AI, aiming to understand the factors contributing to its engagement and perceived utility within a psychological context. Utilising a mixed-methods approach, our findings indicate that this type of chatbot appeals notably to a younger female audience with high technological comfort, a demographic that is increasingly seeking mental health support through digital avenues.[26] This contrasts with some earlier research suggesting that dedicated mental health chatbots might typically attract a somewhat older demographic,[27,28] although more recent studies highlight high acceptance and use among young people as well.[26] The current study contributes to a growing knowledge base on the use and perception of generative AI agents in mental health.

ANTHROPOMORPHISM, PERCEIVED EMPATHY, AND THE THERAPEUTIC ALLIANCE.

A key finding was the participants' tendency to attribute human-like qualities, such as warmth, intelligence, and empathy, to the chatbot, despite their explicit awareness of its non-human nature. This paradoxical engagement, a form of anthropomorphism,[25] suggests users derive comfort, validation, and non-judgmental support from these perceived attributes. These elements are crucial for fostering a therapeutic alliance, a cornerstone of effective client-centred therapy.[29] The chatbot's conversational strategies, which participants described as mirroring cognitive reframing techniques and conveying empathy, appeared to encourage user engagement and emotional disclosure. Furthermore, participants spoke of appreciating their emotional disclosures being positively affirmed by the chatbot, potentially forming a mechanism of positive reinforcement for emotionally expressive behaviours. This aligns with findings from Ho et al.,[18] where psychological benefit from emotional disclosure was achieved if it conveyed a sense of understanding, regardless of whether the source was human or chatbot. This also supports Nienhuis et al.,[29] who identified empathy as a significant predictor of therapeutic alliance.

However, the role of anthropomorphism in facilitating disclosure is complex. While participants in this study valued the chatbot's human-like empathy, other research demonstrates that less anthropomorphic agents can sometimes facilitate more open reporting of personal events, possibly by reducing the user's feeling of being socially judged.[30] This suggests that the perceived benefits may stem not just from human-like qualities, but from an optimal blend of relatable "humanness" and safe "non-human" objectivity.

THE UNIQUE 'NON-HUMAN' THERAPEUTIC SPACE AND USER AGENCY.

Interestingly, the AI's 'non-human' qualities—such as being perceived as unbiased, logical, and objective—were also highlighted as beneficial, particularly when contrasted with human interactions. Participants felt this created a unique therapeutic space, allowing them to explore sensitive subjects more freely, attributing to the AI a sense of safety and reduced fear of judgment not always perceived in human interactions. This finding is consistent with research showing that individuals may prefer discussing embarrassing or sensitive topics with an AI agent precisely because the fear of social judgment is diminished compared to interacting with a human.[31]

This suggests that generative AI may facilitate a nuanced form of interaction that users find uniquely comfortable for certain types of disclosure. Indeed, participants reported that the chatbot's ability to generate hyper-personalised responses fostered a profound sense of being understood. This feeling of being heard without judgment, combined with the ability to direct the conversation, appeared to directly contribute to a sense of empowerment among users, a key component of therapeutic growth.

Previous rule-based conversational tools were often limited by pre-written responses.[19,20] In contrast, the generative AI-driven "Psychologist" chatbot appeared to empower users to steer conversations, set boundaries, and explore highly personalised and niche scenarios. This fostered a deeper sense of being understood and supported, enhancing user agency. This shift underscores the importance of tailoring therapeutic interactions to individual needs and preferences, a principle well-established in human therapy but now extended to AI interactions. The ability for users to directly influence the AI's conversational direction, often in ways they might hesitate with human therapists, introduces a new dynamic in therapeutic communication. It suggests a potential redefinition of social interaction norms within these digital therapeutic contexts, possibly enabling more open dis-

course.

ACCESSIBILITY AND DEMOCRATISATION OF SUPPORT.

The 'Psychologist' chatbot's free, 24/7 accessibility, and perceived anonymity were identified as major advantages, addressing critical barriers to traditional mental health support such as cost, availability, and social stigma. This has the potential to democratise access to initial mental health support, offering an immediate resource. This innovation aligns with a broader trend towards leveraging technology to enhance mental health care accessibility and effectiveness.[32]

BENEFITS, RISKS, AND LIMITATIONS OF AI IN MENTAL HEALTH CHATBOT DEPLOYMENT.

The findings of this study highlight several potential benefits of generative AI chatbots in mental health, including enhanced accessibility, user agency, and the provision of a perceived non-judgmental space for disclosure. However, the deployment of such AI tools is not without significant risks and limitations.[33,34]

One major concern is AI bias. While participants in this study perceived the AI as "unbiased", LLMs are trained on vast datasets derived from the internet and other textual sources, which can contain and perpetuate societal biases related to race, gender, socioeconomic status, and other characteristics.[35,36] There is a risk that AI chatbots could reflect these biases in their interactions, potentially causing harm or reinforcing negative stereotypes, rather than being truly "stigma-free".[37] If an AI perpetuates racial or other biases, it could inadvertently perpetuate stigma related to mental health conditions within certain communities.

Another critical area is safety and accuracy. The statement that users attribute "a sense of safety and confidentiality not always present in human interactions" needs careful consideration. While perceived anonymity can be beneficial, AI chatbots, especially generative ones, can "hallucinate" or generate incorrect, misleading, or even harmful information.[13,38] There is evidence of AI providing inaccurate medical information,[38] which in a mental health context could have serious consequences. Data privacy and security are also paramount; users share sensitive personal information with these platforms, and clarity on data usage, storage, and protection is crucial.[33]

The development of emotional attachment to AI chatbots, as hinted at by the anthropomorphism observed, is another complex issue.[16,17] While some level of connection might enhance therapeutic engagement, over-dependence on an AI for emotional support could potentially hinder the development of human relationships or coping skills, and lead to distress if the service changes or becomes unavailable.[14]

Furthermore, the lack of regulatory oversight for many AI chatbots, particularly user-generated ones on platforms like Character.AI, is a significant concern.[33] Unlike licensed therapists, these chatbots do not operate under established ethical codes or clinical governance, and there is often no clear accountability mechanism.

The therapeutic efficacy of such chatbots also remains to be robustly established through rigorous, controlled trials. While users may perceive benefits, these subjective experiences need to be complemented by objective measures of clinical outcomes.[33] The line between a supportive tool and a clinical intervention can become blurred, raising questions about when and how these tools should be used, especially for individuals with severe mental health conditions.

STUDY LIMITATIONS

This study possesses several limitations that necessitate careful consideration when interpreting the findings. The small sample sizes for both the survey (n=13) and the qualitative interviews (n=8) restrict the generalisability of the results. Participants were self-selected users of a specific chatbot on a particular platform, predominantly comprising young females with high technological comfort. Consequently, the findings may not be applicable to other demographics, users of different AI chatbots, or the broader population seeking mental health support. Recruitment through an online forum dedicated to the chatbot likely introduced selection bias, attracting individuals with strong opinions and potentially underrepresenting those with neutral or less engaged experiences. Furthermore, the self-reported data may be subject to response bias, such as social desirability or recall bias, as the perceived benefits and usage patterns reflect subjective experiences.

Although a semi-structured interview guide was used and thematic analysis involved two researchers, the potential for interviewer bias in questioning or interpretation cannot be entirely dismissed. The study's exploratory and descriptive nature, lacking a control group, prevents any causal inferences about the chatbot's impact on mental well-being. The findings are also specific to the "Psychologist" chatbot on Character.AI and may not translate to other generative AI chatbots or dedicated mental health apps with different designs and functionalities. An additional limitation is the absence of precise retrospective data on the recruitment post's reach, which hinders the calculation of a precise response rate. These limitations collectively highlight the exploratory nature of this research and call for caution in drawing broad conclusions.

FUTURE RESEARCH

This study opens several avenues for future research to build upon its preliminary findings. Larger and more diverse samples are crucial to confirm these results and investigate experiences across various demographic groups, including differences in age, gender, culture, socioeconomic status, and technological literacy across multiple geographical locations. Comparative studies, particularly randomised controlled trials, are needed to rigorously evaluate the effectiveness, engagement, and therapeutic alliance of generative AI chatbots against rule-based chatbots or human therapists to assess clinical outcomes. Longitudinal research is also essential to understand the long-term effects of using AI chatbots for mental health, including usage patterns, sustained benefits, and potential risks such as dependency or altered social interactions.

A significant priority is research into the profound ethical implications, covering data privacy, security, AI bias, and the potential for harm, which should inform the development of necessary ethical guidelines and regulatory frameworks. Further investigation into the mechanisms behind perceived empathy and therapeutic alliance with AI, and how these differ from human interactions, would provide deeper insights into the user experience. It would also be valuable to explore how user-configurable AI personalities, such as those on Character.AI, impact therapeutic outcomes and user engagement. Moreover, future studies should focus on mitigating risks like the perpetuation of stigma or bias by AI and ensuring these tools are equitable and culturally sensitive.

Finally, research focusing on the application of generative AI for specific user needs, such as providing support for mental health conditions such as anxiety, depression, or loneliness, would be highly beneficial. Crucially, future research should also prioritise the co-development of generative AI tools in collaboration with mental health professionals. Such partnerships are essential to ensure that emerging chatbots are not only technologically sophisticated but also clinically relevant, grounded in evidence-based practices, and aligned with ethical standards of care.

Conclusion

The "Psychologist" chatbot on Character.AI, powered by generative AI, appears to offer a uniquely accessible and user-driven form of perceived psychological support for a segment of users, particularly young women. Its ability to provide seemingly empathetic, non-judgmental, and highly personalised interactions, combined with constant availability and no cost, addresses some significant limitations of traditional mental health services. However, these perceived benefits must be weighed against considerable risks and limitations, including the potential for AI bias, misinformation, ethical concerns regarding data privacy and emotional dependency, and the current lack of robust

evidence for clinical efficacy and safety. While generative AI chatbots hold promise for augmenting mental health support, ongoing research, ethical scrutiny, and the development of appropriate safeguards are essential to harness their potential responsibly and ensure they complement, rather than compromise, user well-being. This study serves as an initial exploration into a rapidly evolving field, highlighting the complex interplay between technology, human psychology, and the future of mental health care.

References

1. World Health Organization. World mental health report: transforming mental health for all [Internet]. World Health Organization; 2022 [cited 2025 March 3]. Available from: https://www.who.int/publications/i/item/9789240049338

2. Lau N, Kola L, Zhao X, Asafo S, Attah D, Ben-Zeev D. mHealth for mental health: expanding the reach of care. In: Yzer MC, Siegel JT. The Handbook of Mental Health Communication. John Wiley and Sons; 2025. p. 193-205. https://doi.org/10.1002/9781394179909.ch14

3. Reifegerste D, Wasgien K, Hagen LM. Online social support for obese adults: exploring the role of forum activity. Int J Med Inform. 2017;101:1–8. https://doi.org/10.1016/j.ijmedinf.2017.02.003

4. Haque MR, Rubya S. An overview of chatbot-based mobile mental health apps: insights from app description and user reviews. JMIR mHealth uHealth. 2023;11(1):e44838. https://doi.org/10.2196/44838

5. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med. 2023;29(8):1930–40. https://doi.org/10.1038/s41591-023-02448-8

6. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language models are few-shot learners. Adv Neural Inf Process Syst. 2020;33:1877-1901. https://doi.org/10.48550/arXiv.2005.14165

7. Wei J, Tay Y, Bommasani R, Raffel C, Zoph B, Borgeaud S, et al. Emergent abilities of large language models. arXiv. 2022. https://doi.org/10.48550/arXiv.2206.07682

8. Blazeman98. Psychologist – someone who helps with life difficulties [Internet]. Character.ai; [cited 2025 May 24]. Available from: https://character.ai/character/INhlEC8G/psychologist-helping-life-difficulties

9. Tidy J. Character.ai: young people turning to AI therapist bots [Internet]. BBC News; 2024 [cited 2025 May 24]. Available from: https://www.bbc.com/news/technology-67872693

10. Darcy A. Why generative AI is not yet ready for mental healthcare. Woebot Health; 2023. Available from: https://woebothealth.com/why-generative-ai-is-not-yet-ready-for-mental-healthcare/

11. 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. 2018;6(11):e12106. https://doi.org/10.2196/12106

12. Abd-Alrazaq AA, Alajlani M, Ali N, Denecke K, Bewick BM, Househ M. Perceptions and opinions of patients about mental health chatbots: scoping review. J Med Internet Res. 2021;23(1):e17828. https://doi.org/10.2196/17828

13. Fitzpatrick KK, 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 mental health. 2017;4(2):e7785. https://doi.org/10.2196/mental.7785

14. Abd-Alrazaq AA, Alajlani M, Ali N, Denecke K, Bewick BM, Househ M. Perceptions and opinions of patients about mental health chatbots: scoping review. J Med Internet Res. 2021;23(1):e17828. https://doi.org/10.2196/17828

15. Malik T, Ambrose AJ, Sinha C. Evaluating user feedback for an artificial intelligence–enabled, cognitive behavioral therapy–based mental health app (Wysa): qualitative thematic analysis. JMIR Hum Factors. 2022;9(2):e35668. https://doi.org/10.2196/35668

16. Pentina I, Hancock T, Xie T. Exploring relationship development with social chatbots: a mixed-method study of Replika. Comput Human Behav. 2023;140:107600. https://doi.org/10.1016/j.chb.2022.107600

17. Zhao L, Xu Y, Zhou SK, Wang P. Fostering user attachment to generative artificial intelligence - a theoretical perspective based on awe and gamification. Int J Hum Comput Int. 2025:1-7. https://doi.org/10.1080/10447318.2025.2498486

18. Ho A, Hancock J, Miner AS. Psychological, relational, and emotional effects of self-disclosure after conversations with a chatbot. J Commun. 2018;68(4):712–33. http://doi.org/10.1093/joc/jqy026

19. Laranjo L, Dunn AG, Tong HL, Kocaballi AB, Chen J, Bashir R, et al. Conversational agents in healthcare: a systematic review. J Am Med Inform Assoc. 2018;25(9):1248-58. https://doi.org/10.1093/jamia/ocy072

20. Rahman WN, Hamid NM. Rule-based chatbot for early self-depression indication: a promising approach. JOIV: Int J Inform Visualization. 2024;8(3-2):1625-34. https://doi.org/10.62527/joiv.8.3-2.1628

21. Eysenbach G, Wyatt J. Using the internet for surveys and health research. J Med Internet Res. 2002;4(2):e13. https://doi.org/10.2196/jmir.4.2.e13

22. Wilson L, Marasoiu M. The development and use of chatbots in public health: scoping review. JMIR Hum Factors. 2022;9(4):e35882. https://doi.org/ 10.2196/35882

23. Nowell LS, Norris JM, White DE, Moules NJ. Thematic analysis: striving to meet the trustworthiness criteria. Int J Qual Methods. 2017;16(1). https://doi.org/10.1177/1609406917733847

24. Ames H, Glenton C, Lewin S. Purposive sampling in a qualitative evidence synthesis: a worked example from a synthesis on parental perceptions of vaccination communication. BMC Med Res Methodol. 2019;19(1):26. https://doi.org/10.1186/s12874-019-0665-4

25. Airenti G. The cognitive bases of anthropomorphism: from relatedness to empathy. Int J Soc Robot. 2015;7(1):117–27. https://doi.org/10.1007/s12369-014-0263-x

26. He Y, Yang L, Zhu X, Wu B, Zhang S, Qian C, et al. Mental health chatbot for young adults with depressive symptoms during the COVID-19 pandemic: single-blind, three-arm randomized controlled trial. J Med Internet Res. 2022;24(11):e40719. https://doi.org/10.2196/40719

27. Beatty C, Malik T, Meheli S, Sinha C. Evaluating the therapeutic alliance with a free-text CBT conversational agent (Wysa): a mMixed-methods study. Front Digit Health. 2022;4:847991. https://doi.org/10.3389/fdgth.2022.847991

28. Luxton DD, Sirotin A. Intelligent Conversational Agents in Global Health. In: Innovations in Global Mental Health. Springer International Publishing; 2021. p. 489–500. https://doi.org/10.1007/978-3-319-70134-9_11-1

29. Nienhuis JB, Owen J, Valentine JC, Winkeljohn Black S, Halford TC, Parazak SE, et al. Therapeutic alliance, empathy, and genuineness in individual adult psychotherapy: a meta-analytic review. Psychother Res. 2018;28(4):593–605. https://doi.org/10.1080/10503307.2016.1204023

30. Hsu CW, Gross J, Hayne H. The avatar face-off: a face(less) avatar facilitates adults' reports of personal events. Behav Inform Technol. 2024;43(4):800-10. https://doi.org/10.1080/0144929X.2023.2187242

31. Hsu CW, Gross J, Hayne H. Don't send an avatar to do a human's job: investigating adults' preferences for discussing embarrassing topics with an avatar. Behav Inform Technol. 2021;41(13):2941–51. https://doi.org/10.1080/0144929X.2021.1966099

32. Park SY, Sigmon CN, Boeldt D, Sigmon CA. A framework for the implementation of digital mental health interventions: the importance of feasibility and acceptability research. Cureus. 2022;14(9). https://doi.org/10.7759/cureus.29329

33. Blease C, Torous J. ChatGPT and mental healthcare: balancing benefits with risks of harms. BMJ Ment Health. 2023;26(1). https://doi.org/10.1136/bmjment-2023-300884

34. Muley A, Muzumdar P, Kurian G, Basyal GP. Risk of AI in healthcare: a comprehensive literature review and study framework [Internet]. arXiv. 2023 [cited 2025 March 3]. Available from: https://doi.org/10.48550/arXiv.2309.14530

35. Soun RS, Nair A. ChatGPT for mental health applications: a study on biases. In: Proceedings of the Third International Conference on AI-ML Systems. New York (NY): Association for Computing Machinery; 2023. p. 1-5. https://doi.org/10.1145/3639856.3639894

36. Dergaa I, Fekih-Romdhane F, Hallit S, Loch AA, Glenn JM, Fessi MS, et al. ChatGPT is not ready yet for use in providing mental health assessment and interventions. Front Psychiatry. 2024;14:1277756. https://doi.org/10.3389/fpsyt.2023.1277756

37. Papazova I, Hasan A, Khorikian-Ghazari N. Biased AI generated images of mental illness: does AI adopt our stigma? Eur Arch Psychiatry Clin Neurosci. 2025:1-3. https://doi.org/10.1007/s00406-025-01998-x

38. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. N Engl J Med. 2023;388(13):1233-9. https://doi.org/10.1056/NEJMsr2214184

About the author

› Sam Zaia, a 6th year medical student, developed a therapeutic AI chatbot that received over 200 million conversations internationally. With his past study of psychology before medicine, he looks to understand these conversations and how it is shaping mental health. He is supervised by Dr Mark Huthwaite and Dr Fiona Mathieson of the Department of Psychological Medicine, University of Otago, Wellington.

Declarations/funding

Sam Zaia, an author in this paper, was the developer of the 'Psychologist 'bot. However, no

monetary gain was received for this research, and this research was not prompted by any external agency. There are no conflicts of interest to declare.

Ethical approval/patient consent

Prior to participation, all individuals provided electronic informed consent. Ethical approval for this study was granted by the University of Otago Human Ethics Committee (Reference: D23/315).

Correspondence

Sam Zaia: zaisa376@student.otago.ac.nz