Peer Review File

Article information: https://dx.doi.org/10.21037/mhealth-23-55

Reviewer A

Comment 1: It would be useful to have more information about the data. It states that the paper analysed 101 comments posted on the app's website. Was this the total number of comments posted in that time frame? Is this data publicly available or was a gatekeeper needed?

Comment 2: There is no discussion of ethics. For example, when posting had people consented to their comments being used for research? Eysenbach and Till (2001) also recommend we adjust the content of any quotes used in publications to reduce the chances of the original being found if the extract is placed in search engines. Some discussion of this would be useful.
Eysenbach, Gunther, and James E. Till. "Ethical issues in qualitative research on internet communities." Bmj 323, no. 7321 (2001): 1103-1105.

Reply 1 & 2: Thank you for bringing these issues to our attention. We have answered all these questions in the Section titled "2.2 Study Design"

Updated text: The study was approved by the Institutional Review Board of our home institution. We obtained 300 user reviews from Wysa's Google Play page, where users had voluntarily posted their feedback on the application. The reviews were gathered using the Google-Play-Scraper [17], a Python-based set of APIs that allows for crawling the Play Store without relying on any external dependencies. After reading all the reviews, we excluded those that lacked useful information, were written before January 2020 or were irrelevant to our research, resulting in a final set of 159 reviews for analysis. The review selection happened at two instances, in October 2023 and then again in March 2024. The Google Play Store user reviews are publicly accessible, requiring no gatekeeper for access. According to Google Play's policy [18, 19], publicly posted reviews are visible to all Play Store users and may be utilized by developers to gain insights into user experiences. To safeguard the privacy of the reviewers, in this article, we adjusted the wording of original quotes slightly while preserving their original meaning. This approach aligns with the recommendations of Eysenbach and Till [20], who suggest modifying the content of publicly posted user reviews to reduce the likelihood of the original quotes being located through search engine queries. Additionally, we did not collect any identifiable information such as names or email addresses that may have been posted by the reviewers on the Google Pay page to ensure reviewers' anonymity.

Comment 3: We also need more on why user reviews are useful to analyse. At line 501 it says that user reviews often reflect extreme views, but no citation or evidence is given. The subjective data is criticized as it may not offer a comprehensive evaluation of the app's effectiveness (496). Qualitative research has many strengths but the paper needs to be clear why is the data being used, what can it do, what are its strengths. And then stay true to these claims.

Reply 3: We have updated Section 4.3 Strengths and Limitations to answer these questions.

Updated text: Despite its limitations, it is crucial to recognize the strengths of this research. User reviews offer valuable insights into lived experiences, filling gaps left by clinical research and presenting a grounded perspective on intervention effectiveness. Thematic analysis of user reviews excels in uncovering perceptions, and satisfaction nuances that quantitative metrics frequently overlook. Such detailed understanding is indispensable for developing AI systems for sensitive areas like mental health. Additionally, incorporating up-to-date user reviews enhances the study's relevance in the fast-evolving field of technology.

Comment 4: There is also little information given about how people can access the app.

Reply 4: This is a mobile app that can be used on a phone or tablet. The following text was added to Section 2.3 Application.

Addition to text: The app can be downloaded from both Apple App and Google Play stores on tablets or mobile phones.

Comment 5: The link between data and claim needs to be made clearer in places. It is not always clear how the data offered matches the claim being made. For example, 'tinkering with the AI's black box' seems to largely be about 'missing the context'. It seems that more is needed if the paper is going to return to the claim that people are trying to understand its inner workings (see discussion line 463). Similarly, the quote would suggest that it is "evolution and improvement over ELIZA' rather than 'traditional AI' as a whole.

Reply 5: We agree with reviewer's observation, hence, we took gathered additional data to update the existing themes and subthemes. We have inserted new comments that are more representative of the identified themes and subthemes. We have also removed Theme C 'tinkering with the AI's black box' and updated the subtheme "understanding and contextualization issue" to accommodate the observations made by the reviewer. The term "traditional AI" was changed into "ELIZA" as shown below.

Changes in the text: lines 264-267: A user compared Wysa with an older chatbot., i.e. ELIZA, emphasizing the advancement, "I've tried ELIZA before, but Wysa is different. It's more human-like, with emotions and a sense of humor". (January 7, 2022).

Comment 6: The findings are also quite mixed; showing that some things are perceived positively, some negatively. It would be useful to get a feel for which were dominant themes. As the paper stands we get one quote per idea and no real indication if some themes are stronger in the data than others.

Reply 6 and changes made: We have added the frequency of occurrence of each theme and subtheme to the text.

Reviewer B

Comment 1: The manuscript reports the results of a thematic analysis of Wysa user reviews and provides insights into the users experience as well as offer insights into what users want from mental health chatbots more generally. The introduction and method sections are clear and well presented. Please see my comments and recommendations below.

Reply 1: Thank you for your comment. We have tried to address all the raised issues and concerns to their full extent. Please review our responses below.

2. Methods

Comment 2: The authors say the comments were posted on the app's website by Users. I think some more information is required here to clarify how these reviews were gathered. For example, were the comments posted as voluntary app reviews through the Google Play or Apple App store, or were they requested by the Wysa team either through the app or otherwise.
I'd like the authors to clarify whether the 101 comments downloaded were the total number of comments made by users between the dates provided or are they a sample of the total comments. If they were a sample, please explain how they were selected.

Reply 2: Thank you for bringing these issues to our attention. We have answered all these questions in the Section titled "2.2 Study Design"

Updated text: The study was approved by the Institutional Review Board of our home institution. We obtained 300 user reviews from Wysa's Google Play page, where users had voluntarily posted their feedback on the application. The reviews were gathered using the Google-Play-Scraper [17], a Python-based set of APIs that allows for crawling the Play Store without relying on any external dependencies. After reading all the reviews, we excluded those that lacked useful information, were written before January 2020 or were irrelevant to our research, resulting in a final set of 159 reviews for analysis. The review selection happened at two instances, in October 2023 and then again in March 2024. The Google Play Store user reviews are publicly accessible, requiring no gatekeeper for access. According to Google Play's policy [18, 19], publicly posted reviews are visible to all Play Store users and may be utilized by developers to gain insights into user experiences. To safeguard the privacy of the reviewers, in this article, we adjusted the wording of original quotes slightly while preserving their original meaning. This approach aligns with the recommendations of Eysenbach and Till [20], who suggest modifying the content of publicly posted user reviews to reduce the likelihood of the original quotes being located through search engine queries. Additionally, we did not collect any identifiable information such as names or email addresses that may have been posted by the reviewers on the Google Pay page to ensure reviewers' anonymity.

Comment 3: There appears to be an error in the Methods section line 127, I think the reference should be [5] Braun & Clarke (2006).

Reply 3 and changes made: Thank you for catching this error. We have now made sure that the correct reference is mentioned in the text.

3. Results

Comment 4: There are a couple of typos/errors in the Results section line 141 "judgement-Free", F should not be capitalized, and line 144 "important factors related to effective AI-based chatbot" consider "an effective AI-based chatbot" or "effective AI-based chatbots" depending on what you mean here.

Reply 4 and changes made: We have carefully reviewed the entire paper to address the problems mentioned in the comment.

Comment 5: Some results are presented quite strongly, for example, line 148 "Wysa has been able to…". The paper reviewed 101 comments, a reasonable amount but not exhaustive of Wysa's roughly 11 million users. Within this sample at least 7 of which were completely negative, and 56 were mixed. Therefore, I think some statements should be tempered. In this case (line 148), some user's felt Wysa offered a safe and non-judgmental environment but I imagine it was not all. There are other examples throughout the results usually in the theme overview and summary more so than the sub-themes, such as line 188, line 220, that are very strong and sound like a marketing pitch. These statements are not what I would expect in a qualitative study.

Reply 5 and changes made: We have reviewed and revised all the themes and subthemes to ensure that the statements do not appear as marketing pitches, or strong statements tilting the opinion in one direction.

Comment 6: It would also be useful to provide the number of reviews that support each theme and sub-theme so readers can see how strong support for each is.

Reply 6 and changes made: We have added the frequency of occurrence of each theme and subtheme to the text.

Comment 7: Sometimes names are provided for the reviews, and sometimes not. It would be good to have this consistent throughout the results section.

Reply 7 and changes made: We have removed all names from the paper to present user's anonymity.

Comment 8: Theme C. Tinkering to Unlock the Ai's Blackbox is interesting, while I agree with some of the authors interpretations, I think they have attempted to present these reviews in a very positive light. Framing comments as "users' active engagement" rather than acknowledging that in the quotes presented some are quite critical of the AI system. Especially the sub-theme Experimenting with AI's Capabilities and Limitations, I do not think this user's comments aligns with the summary that this understanding enhances their experiences as that review is quite negative. Consider reframing to acknowledge the frustration of users when AI does not work how they expect.

Reply 8 and changes made: We agree with reviewer's observation and have expanded the theme, i.e. "AI Limitations Detract from User Experience" to highlight user's challenges. We have deleted Theme C altogether.

4.Discussion
4.2 Strengths and limitations
Comment 9: The authors state in the introduction the objective is to fill a knowledge gap, namely understanding of how AI-driven tools are perceived and utilized by users by investigating user engagement focused on user perceptions. The limitations then focus on the subjective nature of user reviews and how they do not offer a comprehensive evaluation of the apps effectiveness, but this was not the purpose of this paper. I would argue the best way to understand how AI is perceived by users is to ask them qualitatively and the authors go on to discuss the strengths of this approach. This appears as a way to flag the "extreme opinions" of reviews and diminish the criticism of the users. I think some of the limitations mentioned in the paragraph are important to state but should rather be reframed as not representing every users experience rather than speaking to the effectiveness of the app.

Comment 10: I think a limitation of the study not mentioned is the use of reviews (although as the authors say there are some merits to this approach) rather than engaging with users in a more meaningful way, such as focus groups or consumer panels. Other mental health apps are undertaking codesign phases and research to achieve this (see Bevan et al., 2020 https://europepmc.org/article/MED/32572961; Wrightson-Hester et al., 2023 https://pubmed.ncbi.nlm.nih.gov/37477969/; Thabrew et al., 2018 https://europepmc.org/article/MED/30405450) and an approach like this would address some of the limitations of the qualitative approach the authors mention, such as not knowing how long people use Wysa or their characteristics. I think this would be worth mentioning as a limitation and possible future direction.

Reply 9 & 10: We agree with reviewer's comments and have updated Section 4.3 Strengths and Limitations accordingly.

Updated Text: The overall research presents valuable insights but comes with its own set of limitations that warrant consideration. The user reviews are inherently subjective and captured in a snapshot in time, lacking longitudinal data that could offer insights into the evolving user experiences. It is also unclear how long people used the app for and what were their characteristics. Given that user reviews often reflect extreme opinions [27], user demographics can aid in their interpretation. Alternative methods, such as focus groups or consumer panels can help engage with users in a more meaningful way. Moreover, other mental health apps [28, 29, 30] are undertaking codesign phases and research to obtain a broader and more detailed understanding of user experiences, characteristics, and engagement nuances. Such methods could indeed complement our current approach by addressing some of its limitations, such as not knowing how long people use Wysa or their characteristics.

Despite its limitations, it is crucial to recognize the strengths of this research. User reviews offer valuable insights into lived experiences, filling gaps left by clinical research and presenting a grounded perspective on intervention effectiveness. Thematic analysis of user reviews excels in uncovering perceptions, and satisfaction nuances that quantitative metrics frequently overlook. Such detailed understanding is indispensable for developing AI systems for sensitive areas like mental health. Additionally, incorporating up-to-date user reviews enhances the study's relevance in the fast-evolving field of technology.

Comment 11: It is not clear what the authors mean on line 515: "The study also implicitly captures the long-term….". I do not understand how user reviews demonstrate this?

Reply 11: In some reviews, users had explicitly mentioned the amount of time they had been using the app. For example, some users indicated that they have been using it for the past three months, a couple of users had been using it for the past four years, etc. However, we think that it is better to remove this from the text as it may require further explanation and may detract from the original arguments made in the paper.

4.2 Comparison with similar research and 4.3 Strengths and limitations
Comment 12: Error line 524 "those works differ from the one reported one."
Error line 533 "they their focus"
Error line 559 "is remain crucial"

Reply 12: We have made the corrections to the errors pointed out by the reviewer.

4.2 Comparison with similar research
Comment 13: This does not really discuss or compare the findings of the studies mentioned. The exception being the few sentences on Malik et al. I think it would be better to compare the findings of the current study to the findings of these papers and demonstrate how they are different or similar and why, rather than make these general statements.

Reply 13: Based on this comment, we have now added findings of the other studies mentioned in this section.

Updated text: Kettle and Lee [25] explored user experiences of two different types of conversational agents to identify important features for user engagement. However, they focused on understanding experiences of students transitioning into university life and facing challenges like the pandemic. Their themes are very similar to our findings. They also found that accessibility and availability, communication style, conversational flow and anthropomorphism are important features of a mental health chatbot. In addition, they found that user response modality, perceived conversational agent's role and question specificity are essential for user engagement.

Maharjan et al. [26] focused on studying the user experiences of individuals with diagnosed mental health conditions using speech-enabled conversational agents. Their findings highlight participants' strategies to mitigate conversational agents' flaws, their personalized engagement with these agents, and the value placed on conversational agents within their communities. The data suggests conversational agents' potential in aiding mental health self-reporting but emphasizes the urgent need to improve their technical and design aspects for better interaction and at-home use.

5. Conclusion
Comment 14: Like my comments in the Results section I think some of this is worded strongly and not what I would expect from a qualitative study. I think the authors should focus on what the study tells them about users' experience of Wysa rather than the app's effectiveness as the results cannot speak to that.

Reply 13: The wording of the conclusions section was adjusted to address the mentioned concern.

Updated text: Section 5 Conclusion.
Our findings suggest that, with AI-driven conversational agents, users can experience non-judgmental space to express their thoughts and emotions. The ubiquitous availability and real-time support can be particularly valuable to users during moments of acute stress or anxiety. Furthermore, the anonymity and confidentiality features encourage users to engage more deeply with the app's therapeutic offerings.

Reviewer C

Comment 1: In the introduction, the author briefly introduces fundamental concepts related to machine learning and natural language understanding. However, there is a lack of emphasis on research in the field of mental health support and the development of AI-driven mental health support. It is important to provide a concise overview of the evolution of this field to provide context for the study's focus.

Comment 2: The introduction of this study lacks a detailed explanation of its contributions and innovations. It is advisable to provide a more comprehensive description of the unique contributions and innovations of this research. Additionally, the author discusses the differences between this study and similar research in lines 522-536. It would be beneficial to incorporate this information into the introduction to highlight the contributions and innovations of this study right from the beginning. This will help readers understand the significance of the research and its distinctive features in the context of existing work in the field.

Reply 1 and 2: Based on these comments, we have rewritten the introduction, so it covers the the evolution, existing state and gaps in the field in a more balanced manner.

Updated text: Please review the updated Section 1.1 Introduction.

Comment 3: Using tables to compare and contrast with relevant research is indeed a valuable way to present information, and incorporating 1-2 illustrative figures to showcase the work done in the study can enhance the presentation. Tables can help summarize key findings, and figures can provide visual representations of data or concepts, making the content more accessible and engaging for readers. Consider including tables for comparative analysis and relevant figures in the manuscript to improve the clarity and visual appeal of the research.

Reply 3: Thank you for the comment. We have included one table and one figure in response to this comment.

Comment 4: In the methods section, providing a more detailed explanation of the qualitative analysis techniques and offering specific examples would be beneficial. This additional detail can help readers better understand how the data were analyzed and interpreted in the study. Consider elaborating on the qualitative analysis process, including the steps involved, any coding or thematic analysis methods used, and examples of how data were categorized or interpreted to support the study's findings. This added clarity can enhance the rigor and transparency of the research methodology.

Reply 4: We have expanded the Section 2.4 Data Analysis section to address the issues raised in this comment.

Updated Text:
We used Braun and Clark's thematic analysis technique [21] to understand how people are using the Wysa app, how they feel about it, and what they think are its benefits and drawbacks. We began with becoming thoroughly familiar with the corpus through repeated readings, allowing for an understanding of content and the identification of key ideas. By the end of this phase, we had categorized the reviews into four types based on sentiment, as described below:

(a) Positive reviews (n=69), which solely highlighted app benefits, for example, "Very helpful practices, prompts, activities and 'chats'!! Even on the free version it is a BIG help for managing anxiety, getting perspective, and prioritizing self-care."
(b) Negative reviews (n=24), which focused on drawbacks, for example, "The bot doesn't understand me. It seems to work by recognizing words like stressed or sad. It commonly misunderstands me because of this and it's usually in a set path and unable to talk about what you really want. Unhelpful."
(c) Mixed reviews (n=57), which noted both strengths and weaknesses, for example, "Sometimes the responses seem like they would be the same no matter what I say, but just being able to bounce my thoughts off Wysa and think about different perspectives is helpful."
(d) Neutral reviews (n=9), which made feature suggestions for improving the application, for example,

> "I would like to have the option to customize the screen with themes. Perhaps choose a character to interact with."

We then used an open and inductive approach to code the reviews such that significant ideas were labeled (coded) in each user review. To ensure accuracy and reliability, one author first coded all the reviews, and then the second re-reviewed and refined them. Once the codes were finalized, based on similarities and differences, the first author grouped the codes such that each group could become a potential theme or sub-theme of the recurrent ideas in the reviews. The two authors worked together through the iterative process of refining and revising the groups such that a set of themes emerged that accurately represented our dataset. Each theme was then clearly defined and named, and a detailed explanation was formulated to articulate its meaning and fit within the overall narrative.

Comment 5. It is essential to provide relevant information about the dataset used in the study to give readers context and transparency about the data source. Specifically, consider showcasing certain aspects of the dataset, such as representative examples of positive, negative, and neutral reviews or opinions. Additionally, in your study, where thematic analysis played a significant role in determining positive, negative, and neutral sentiments, it's crucial to address the potential subjectivity of this technique and whether different individuals may have different evaluation criteria for the same content. Acknowledging and discussing the subjectivity and potential variations in evaluation standards among different individuals in the text will help clarify the methodology and ensure transparency in the research process.

Reply 5: We have expanded the Section 2.2 Study Design and Section 2.4 Data Analysis sections to address the issues raised in this comment. Section 2.4 has been pasted in the previous response. Section 2.2 is pasted below.

Updated Text:
The study was approved by the Institutional Review Board of our home institution. We obtained 300 user reviews from Wysa's Google Play page, where users had voluntarily posted their feedback on the application. The reviews were gathered using the Google-Play-Scraper [17], a Python-based set of APIs that allows for crawling the Play Store without relying on any external dependencies. After reading all the reviews, we excluded those that lacked useful information, were written before January 2020 or were irrelevant to our research, resulting in a final set of 159 reviews for analysis. The review selection happened at two instances, in October 2023 and then again in March 2024. The Google Play Store user reviews are publicly accessible, requiring no gatekeeper for access. According to Google Play's policy [18, 19], publicly posted reviews are visible to all Play Store users and may be utilized by developers to gain insights into user experiences. To safeguard the privacy of the reviewers, in this article, we adjusted the wording of original quotes slightly while preserving their original meaning. This approach aligns with the recommendations of Eysenbach and Till [20], who suggest modifying the content of publicly posted user reviews to reduce the likelihood of the original quotes being located through search engine queries. Additionally, we did not collect any identifiable information such as names or email addresses that may have been posted by the reviewers on the Google Pay page to ensure reviewers' anonymity.

Comment 6: The practical application of research findings to the design and implementation of AI-driven mental health support systems is an important aspect that should be explored in the article. It is advisable to discuss how the research outcomes can be translated into real-world applications to enhance the effectiveness and usability of AI-driven mental health support systems. This discussion can include recommendations for system developers, mental health professionals, or policymakers on how to integrate the insights gained from the study into the development and improvement of AI-driven mental health solutions. Addressing this aspect will provide valuable guidance and make the research more actionable and impactful in the field of mental health support.

Reply 6: Thank you for the comment, we have revised the Implications section according.

Updated Text: Please review Section 4.4 Implications.

Comment 7: Integrating artificial intelligence into mental healthcare is of significant importance. However, in this field, ethical concerns and data protection measures are paramount to safeguarding user privacy and rights. Have there been specific solutions in previous research to address this issue? Are they effective? How should these issues be addressed in future research? It is recommended to elaborate on these relevant aspects.

Reply 7: We have address issues raised in this comment to some extent. Particularly, we have discussed different data protection techniques employed by AI algorithms and some made some suggestions for addressing this issue in future research.

Changes in Text: Please review lines 615-629

Reviewer D

Comment 1: You have developed and conducted really useful results here that I'm sure will be of benefit to the mhealth readership. To make the most research impact, I believe that some major changes are needed, particularly in the structuring of the paper and level of detail in some areas. Please have a look at my comments below.

Reply 1: Thank you for your comment. We have tried our best to address to the fullest, all aspects of each comment.

Broad comments:
Comment 2: Intro should be restructured. Objectives should go at the end of the intro, just before the methods.

Reply 2: The introduction was restructured and rewritten to address the mentioned deficiencies.

Comment 3: The context and paper overview should be listed in the methods.

Reply 3: The context and paper overview have been moved to the methods section as suggested.

Comment 4: Would be good to include a figure with screenshots of the app that exemplify some of your description of the app

Reply 4: Figure has been included.

Comment 5: Methods need more work and detail. e.g. How did you download comments? What date did you extract this information, where from and what kind of information was provided with the comment? Which software did you use for analysis? Were both authors involved equally in each stage? How was a consensus reached? Did authors code separately and then discuss these, or did one author code and the other revise these? How did you identify "good", "negative" or "mixed" - was there some criteria or system you used?

Reply 5: Thank for the feedback. We have updated Section 2.4 Data Analysis to answer all the questions rasied in the comment.

Updated Text:
We used Braun and Clark's thematic analysis technique [21] to understand how people are using the Wysa app, how they feel about it, and what they think are its benefits and drawbacks. We began with becoming thoroughly familiar with the corpus through repeated readings, allowing for an understanding of content and the identification of key ideas. By the end of this phase, we had categorized the reviews into four types based on sentiment, as described below:

(e) Positive reviews (n=69), which solely highlighted app benefits, for example, "Very helpful practices, prompts, activities and 'chats'!! Even on the free version it is a BIG help for managing anxiety, getting perspective, and prioritizing self-care."

(f) Negative reviews (n=24), which focused on drawbacks, for example, "The bot doesn't understand me. It seems to work by recognizing words like stressed or sad. It commonly misunderstands me because of this and it's usually in a set path and unable to talk about what you really want. Unhelpful."

(g) Mixed reviews (n=57), which noted both strengths and weaknesses, for example, "Sometimes the responses seem like they would be the same no matter what I say, but just being able to bounce my thoughts off Wysa and think about different perspectives is helpful."

(h) Neutral reviews (n=9), which made feature suggestions for improving the application, for example, "I would like to have the option to customize the screen with themes. Perhaps choose a character to interact with."

We then used an open and inductive approach to code the reviews such that significant ideas were labeled (coded) in each user review. To ensure accuracy and reliability, one author first coded all the reviews, and then the second re-reviewed and refined them. Once the codes were finalized, based on similarities and differences, the first author grouped the codes such that each group could become a potential theme or sub-theme of the recurrent ideas in the reviews. The two authors worked together through the iterative process of refining and revising the groups such that a set of themes emerged that accurately represented our dataset. Each theme was then clearly defined and named, and a detailed explanation was formulated to articulate its meaning and fit within the overall narrative.

Comment 6: Also, you later mention that there is a free and paid version. How did you filter out the paid version to ensure that the comments on the app store only reflected the free version's comments (it appears this is what was done)

Reply 6: It was not always possible to tell whether the reviewer was using the paid or free version of the application. Therefore, we did not use this criterion to select the reviews.

Comment 7: Results: Surely not every comment contributed to every theme. It would be good to know the number of comments that contributed to each of these themes. Also, you mentioned above there are "good" "negative" and "mixed" reviews, the breakdown of these for each theme would be good. If possible, it would also be good to mirror this approach with the sub-themes.

Reply 7: We have revised the description of the themes, wherever possible, to clarify how many positive or negative comments were present within each theme/subtheme.

Comment 8: Make sure the quotes used exemplify the statement you are making. e.g. "Pretty cool. Great AI interface. And it helps me to talk, especially since I'm not talking to a judgmental human. The AI seems unconditional and set on being helpful." This comment doesn't necessarily show self-reflection. Is there another comment you can use for this?

Reply 8: We have updated the carefully revised all the themes to ensure that the captured user reviews exemplify the theme/subtheme more closely.

Comment 9: In the results, anonymise all names (e.g. saying "the user"). There is no benefit to writing out the user names so this should be anonymised.

Reply 9: All the user reviews are now anonymized.

Comment 10: The strengths/limitations can be more concise, particularly the strengths section which largely outlines the strengths of this type of qualitative research (rather than this specific study)

Reply 10: We have revised the Section 4.3 Strengths and Limitations to make it more concise.

Updated Text: 4.3 Strengths and limitations

The overall research presents valuable insights but comes with its own set of limitations that warrant consideration. The user reviews are inherently subjective and captured in a snapshot in time, lacking longitudinal data that could offer insights into the evolving user experiences. It is also unclear how long people used the app for and what were their characteristics. Given that user reviews often reflect extreme opinions [27], user demographics can aid in their interpretation. Alternative methods, such as focus groups or consumer panels can help engage with users in a more meaningful way. Moreover, other mental health apps [28, 29, 30] are undertaking codesign phases and research to obtain a broader and more detailed understanding of user experiences, characteristics, and engagement nuances. Such methods could indeed complement our current approach by addressing some of its limitations, such as not knowing how long people use Wysa or their characteristics.

Despite its limitations, it is crucial to recognize the strengths of this research. User reviews offer valuable insights into lived experiences, filling gaps left by clinical research and presenting a grounded perspective on intervention effectiveness. Thematic analysis of user reviews excels in uncovering perceptions, and satisfaction nuances that quantitative metrics frequently overlook. Such detailed understanding is indispensable for developing AI systems for sensitive areas like mental health. Additionally, incorporating up-to-date user reviews enhances the study's relevance in the fast-evolving field of technology.

Overall, the study offers a balanced and insightful contribution to the field.

Comment 11: Move the comparison with similar research to before the strengths/limitations

Comment 12: There were various instances of incorrect spelling and grammar, correct these. e.g. sentence structure errors in the first couple of sentences of the "comparison with similar research" paragraph.

Reply 11 & 12: The requested changes have been made to the text.

Comment 13: In the discussion, you stated you used an inductive approach. This should be stated in the methods. Also, explain how this inductive approach adds to the literature, is there any particular reason why an inductive approach may have different outcomes/value to the literature than the deductive approach from the other study?

Reply 13: Thank for the comment. The inductive approach allows for new and fresh data to emerge from the text corpus.

Updated Text: Please review the updated Sections 2.4 and 4.3 where the issues discussed in the comment have been addressed.

Comment 14: "The study also implicitly captures the long-term use and engagement levels with the app, which is a vital metric for any digital health tool". It did not seem to do this from what I have read. How did it capture these metrics (if applicable, you can explain this in the methods)?

Reply 14: Some reviews explicitly indicated the length of time for which the reviewer was using the application. However, we think that this comment was unnecessary in the light of the revisions made, so it has been removed.

Comment 15: Many of the research implications don't seem to be based on this study alone but on expectations for the whole conversational chatbot field. It may be better labelled as "future directions", or reformulated to explicitly show how your study's results led to each of these listed implications

Comment 16: The discussion section, particularly the strengths/limitations section onwards, is quite verbose and should be made more concise. There are multiple instances of repetition that can be removed

Reply 15 & 16: Thank for the comment. We have revised the Implications section so that it is more in line with the study findings. We have also made it more concise and ensured that there are no repetitions

Updated text: Please review the updated Section 4.4 Implications.

Specific/minor comments:

Comment 1: LLMs: State what this is and explain as it is a core tenant of the study
Reply 1: A brief description of LLM has been added to the text.

Changes in the text: Please review lines 593-596.

Comment 2: Background can start broader by first identifying the overall utility of AI in healthcare. Then you can link this up with discussing conversational chatbots

Comment 3: Too long spent on explaining the first chatbot. More effort is needed to explain the current utility for conversational chatbots. Overall, more detail is required in the background section.

Comment 4: For the aim: [13,14,15]. Not clear how these citations are relevant to support your aim. I would remove and instead discuss these studies earlier in your intro before you state your aim.

Reply 2, 3 & 4: Thank you for the comment. The Background section has been rewritten according to the guidelines provided.

Changes in the text: Please review Section 1.1 Background.

Comment 5: I would give different headings for context and paper overview. e.g. Context could be titled "The Wysa app", and Paper overview could be titled "Design".

Comment 6: I recommend using standard headings under the methods, such as "participants", "design", "procedure" and "data analysis"

Reply 5 & 6: The recommended changes were made to the text.

Changes in the text: Review Sections 2.1, 2.2 and 2.4.

Comment 7: First paragraph of results: these themes don't match the themes listed in your abstract. Make sure the wording is consistent
Comment 8: Results: titles should exactly match the themes you have listed

Reply 7 & 8: We have made sure that the theme and subtheme labels are consistent throughout the paper.

Comment 9: The term black box was used. If you're using these terms they should be defined in the intro
Comment 10: You can correct the spelling in comments made, such as "sonetimes" into "sometimes"

Reply 9 & 10: We are no longer using the term black box and the spelling correction has been made.

Comment 11: When discussing involving users in the design process, provide one or two examples here of new (2023) studies that have used co-design for AI-informed apps and the kinds of outcomes they have reported.

Reply 11: We have added some studies to that have used co-design for AI-informed apps. We did not discuss the outcomes of these studies as it would not fit the context in which these studies have been mentioned.

Changes in text: Please review lines 569-571.