"Not quite there yet": On Users Perception of Popular Healthcare Chatbot Apps for Personal Health Management

Ifunanya Barbara Onyekwelu
Dept. Mechanical, Electrical & Chemical Engineering, Oslo Metropolitan University, 0130 Oslo, Norway
s371147@oslomet.not

Raju Shrestha
Dept. Computer Science, Oslo Metropolitan University, 0130 Oslo, Norway
Raju.Shrestha@oslomet.no

Frode Eika Sandnes
Dept. Computer Science, Oslo Metropolitan University, 0130 Oslo, Norway
frodes@oslomet.no

ABSTRACT

Many individuals rely on digital resources for advice related to their health management such as passive information on web or more active resources such as chatbots. Chatbot technology has made rapid technical advances in recent years and holds potential for making health information accessible to a wider range of individuals including sparsely populated and rural areas. One challenge with new technology is the gap that can occur with the functionality offered by the technology and the actual needs of users. Not only is it important that health related information and advice is accurate and correct as it directly can affect individuals' health-related decisions, but the user experience must be positive, and users must trust the technology. This study explores how users perceive four popular health related chatbot apps by analyzing 708 reviews. The results confirm that there is a gap between users' needs and the user experience provided by the chatbots. Suggestions for chatbot developers are provided which could help reduce the gap between the available functionalities and users' needs.

CCS CONCEPTS

• **Information systems**; • **Information retrieval**; • **Users and interactive retrieval**; • **Computing methodologies**; • **Artificial intelligence**; • **Natural language processing**;

KEYWORDS

Healthcare technology, Chatbot, Artificial intelligence, Sentiment analysis, Qualitative analysis

ACM Reference Format:
Ifunanya Barbara Onyekwelu, Raju Shrestha, and Frode Eika Sandnes. 2024. "Not quite there yet": On Users Perception of Popular Healthcare Chatbot Apps for Personal Health Management. In *The PErvasive Technologies Related to Assistive Environments (PETRA) conference (PETRA '24), June 26–28, 2024, Crete, Greece*. ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3652037.3652042

1 INTRODUCTION

The exponential growth of artificial intelligence (AI) has allowed for its integration into multiple sectors, including healthcare. AI-based healthcare chatbots is one prominent application of this technology. Chatbots have emerged as a valuable resource for improving patient treatment and assisting healthcare practitioners through various AI-based technologies [34]. Chatbots could help increase access to health-related information in less populated areas, rural areas and low-income areas with lower density of health personnel. According to Salazar et al. [30] many medical professionals perceive chatbots as particularly useful for tasks such as scheduling doctor appointments, locating health clinics, or offering medication information. Chatbots are typically designed to imitate human conversation using text or voice interactions [2]. This paper focuses on text-based chatbots. Chatbots use a wide range of conversational data to learn and respond more effectively [14].

Recent advances in artificial intelligence, machine learning, and natural language processing have helped advance chatbot technology, and current AI chatbots can accomplish many advanced tasks, such as healthcare advice, symptom analysis, mental health support, and even medication management [27].

User Interface (UI) is the visible façade of a mobile app that directly influences users' impressions. User centered development processes are considered the golden standard for interactive systems, yet many advanced systems based on emerging technologies are developed using more traditional techno-centric methods. This is understandable as emerging technologies such as generative AI have non-deterministic behavior that is hard to fully explore with traditional formative user testing. Holmes et al. [20] argued that the traditional best practices normally applied to User Experience (UX) design cannot easily be applied to chatbots, nor can conventional usability testing techniques guarantee usability. An alternative and resource for learning about users' experiences with such technology is to solicit summative app reviews [12].

This paper details a document study, based on 708 user reviews, that analyzed UI-related issues in mobile healthcare chatbot apps. The goal was to understand what factors contribute to positive or negative perceptions in each app, and how these factors differ across apps. By examining user reviews, common themes related to barriers (such as technical issues, lack of personalization, or inadequate medical content) and facilitators (such as user-friendly design, accuracy of information, or helpfulness in managing health concerns) can be identified. Relevant information grounded in empirical evidence is a prerequisite for designing health chatbots that meet users' needs.

PETRA '24, June 26–28, 2024, Crete, Greece

ACM ISBN 979-8-4007-1760-4/24/06
https://doi.org/10.1145/3652037.3652042

2 RELATED WORK

Chatbots have a long history with a renewed interest due to recent advances in artificial intelligence (see for instance the general review [26]). Technological advances in text-to-speech and speech-to-text have also triggered related activity in voice assistants [16]. Chatbots typically allow free-text prompts that reflect natural language. This contrasts with traditional query-based systems that can be cognitively demanding [6]. Chatbots are used in many problem domains including healthcare [5, 21]. Several healthcare related chatbots have been implemented and tested in clinical settings [3, 22]. Health related chatbots have also been explored as a resource in medical and healthcare education [18].

Bhirud et al. [7] argued that most healthcare chatbots only provided answers to general health related problems based on FAQs. They suggested that chatbots should provide more natural communication through the use of natural language processing to mimic the responses of a medical doctor. A relatively recent review [29] found that most chatbots provided fixed output. However, recent advances in generative AI may change this, for example by providing summaries [24]. Shan and colleagues [33] provide an overview of how language is used in health related chatbots and its impact on the users.

Through an online survey, Sweeney et al. [35] found similar results, namely that about half of mental health professionals responding found chatbots to provide benefits for clients to manage their mental health, but that they are not yet sufficiently understanding or expressing human emotion. Cameron et al. [10] argued that such chatbots could be a useful remedy to counteract long waiting lists to see professionals and provide access to sparsely distributed individuals in rural areas. They also pointed out the risks and ethical considerations associated with implementing mental health chatbots as well as the importance of usability. One ethical issue is related to how users may be nudged by the technology with either positive or negative effects [37]. A review of health chatbot papers [25] revealed that a majority did not address security and privacy issues; they called for intensified research into chatbot privacy.

Using an online survey to probe 100 physicians' opinions about general health chatbots, Palanica and colleagues [28] found that most physicians acknowledged potential benefits to include supporting, motivating, and training patients with the potential of being "surrogate" caregivers. Scheduling of doctor appointments, locating health clinics, and providing medication information were explicitly listed. However, lack of human emotion and specialist knowledge to provide reliable diagnosis were the main perceived weaknesses of chatbots.

Studies of health related chatbots have shown that utility [36] and information quality, accuracy, and competence [8] are key success criteria. Using eye-tracking, Chen et al. [11] found that anthropomorphic appearances and human-like conversation styles in chatbots affected users' perception of social presence, trust, and satisfaction. Baek and Kim [4] studied the "creepiness" of ChatGPT and found it related to task efficiency and social interaction.

3 METHODS

The literature on chatbots typically evaluates performance according to usability, classifier performance, speed, comprehensibility, realism, repetitiveness, word error rate, concept error rate, and aesthetics [1]. This work relies on quantitative measures as well as qualitative opinions.

3.1 Chatbot health apps

Four Android mobile health apps were randomly selected from Google Play based on their popularity, namely Woebot, WebMD, Ada and Healthily. Woebot is an AI-powered mental health chatbot that uses principles of cognitive-behavioral therapy to assist users in managing mental health issues such as depression, anxiety, and stress. It is designed to provide users with emotional support and therapeutic conversations. WebMD, Ada and Healthily (also known as Your.MD) are symptom checker apps with chatbots. They are designed to help users manage their health independently. The smartphone platform was chosen due to its convenience and pervasiveness among users, despite a risk of being more cognitively demanding [31].

3.2 Data collection

User reviews for the four mobile apps were first scraped from Google Play using the 'google scraper' python tool. The reviews collected are all public and there are therefore no personal data or privacy issues related to the data collection. All reviews were also classified as verified reviews.

A total of 2000 random user reviews were collected, that is, 500 for each app. Next, only recent reviews were included (2021-2023) resulting in 708 reviews (Ada: 302, Woebot: 183, WebMD: 153, and Healthily: 68). For each user review, the author, date, score (star rating), review comment, developer response and time of response were collected. Both quantitative and qualitative analysis were performed on the collected data.

3.2.1 Quantitative Analysis. Quantitative analyses were performed on the collected data using rating analysis, sentiment analysis, temporal trends, and developer response analysis.

Rating analysis was performed by computing the mean rating and rating distribution for each app. The mean rating was to give an overall sense of user satisfaction while the rating distribution was used to identify if most users were generally satisfied (higher ratings) or not (lower ratings).

Sentiment analysis is the task of extracting and analyzing people's opinions, sentiments, attitudes, perceptions, etc., toward different entities such as topics, products, and services. It can be used for monitoring brand reputation, analyzing customer feedback, or gauging public opinion on social media by extracting meaningful insights from textual data. The sentiment analysis conducted herein was inspired by [9, 19] and was conducted using the TextBlob sentiment analysis tool in Python to classify the sentiment of each review as positive, negative, or neutral.

Temporal trends were analyzed to see how ratings changed over time by grouping the reviews by year and computing the mean rating per year.

Table 1: Thematic coding of review comments

Category	*Theme*	*Description*
Appearance	Layout	Clear or unclear visual structure
	Font size	Adequacy of the font size
Interaction	Navigation	Difficulty
	Notification	Too few or too many
	Responsiveness	Ability to respond to user text input
	Ease of Use	How easy to use the app
Functionality	Usefulness	Usefulness of the app
	Update bugs	Issues introduced with updates
	Login	Negative login experiences
	Crash issues	Apps crashing
	Accuracy	Content accuracy
Experience	Advertisement	Appropriateness of ads
	Feedback	Responsiveness of app developers
	Cost	Payment or free

Some users may read reviews to determine whether they want to download an app or not, and a developer's comment on a user's review can help influence their decision. Secondly, some users tend to change their reviews and give an app a higher rating after receiving a response from the app developers. The reason for most rating decrease is that app owners do not fully meet users' needs or respond too late [12]. Developer response analysis was conducted by calculating the percentage of reviews that received a developer response in a bid to analyze if the presence of a developer response correlates with higher subsequent ratings (indicating effective issue resolution). Thus, a correlation analysis was carried out between the developer response scores and the mean ratings of each app. Correlating the two allowed us to investigate the degree of consistency (or discrepancy) between users and developers ratings of the apps.

3.2.2 Qualitative Analysis. A thematic analysis was conducted by coding the reviews along the dimensions shown in Table 1. To ensure coding reliability, the reviews were extracted based on score ratings for each app; each user comment was manually reviewed to identify UI related keywords.

A deductive coding approach inspired by Chen et.al [12] was used to identify 14 UI-related themes. These themes were then organized into four categories, namely *Appearance*, *Interaction*, *Functionality* and *Experience*. For example, a user review, "*The UI is not user friendly at all for a health application. One would think they would take into account poor vision, poor coordination, shaking hands and other conditions where having small icons and buttons are a bad decision. For example, what is up with the redesigned medication reminder? The check marks require a magnifying glass to see. Could they be any smaller? What a ludicrously stupid design choice. The check marks should remain prominent and obvious at a glance*" was classified as a font-size related issue under the appearance category. Similarly, the review "*Im not sure what its actually supposed to do. The medical advice is dreadful, type your symptoms in and 99% of time it will just say go to the doctor to cover itself. The information on the app is just recycled from Google searches. Of zero use to me*" was classified as related to usefulness under the functionality category.

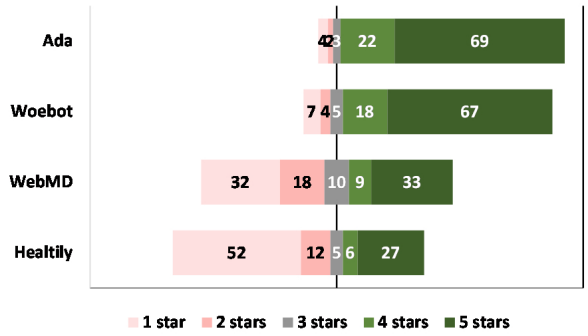

Figure 1: Diverging stacked bar-graph showing user rating (percentage) distributions of the four apps. Ratings were from 1 (lowest) to 5 (highest).

4 RESULTS

4.1 Quantitative findings

Figure 1 shows the distributions of the scores for the four apps. Healthily ($M = 2.44$, $SD = 1.74$) is the app with the most negative reviews with 51.5% of 1-star ratings, indicating a significant portion of users are not satisfied with the app. At the other end of the scale Ada ($M = 4.50$, $SD = 0.97$) was associated with the most positive reviews with 69.1% 5-star rating suggesting a high degree of user satisfaction. WebMD ($M = 2.92$, $SD = 1.68$) exhibited a balanced percentage (32%) of 1-star and 5-star ratings which suggests diverging responses where one portion of users found it useful and satisfactory, another portion had negative experiences with the app, while the remaining users were more neutral. The large portion (67.8%) of the 5-star rating for Woebot ($M = 4.37$, $SD = 1.12$) suggests a high level of user satisfaction though not as strong that of Ada.
Figure 2 shows the sentiment distributions for the four apps. It seems that the sentiments correlate with the ratings. Ada and Woebot are associated with the highest reviews of 77.9% and 73.8%, respectively, indicating strong user satisfaction. Although Healthily

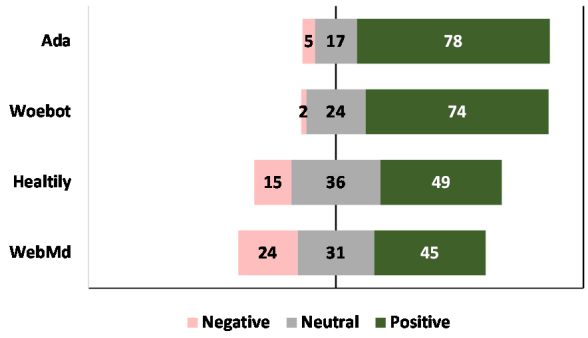

Figure 2: Diverging stacked bar graph showing the sentiment distributions for the four apps (percentages).

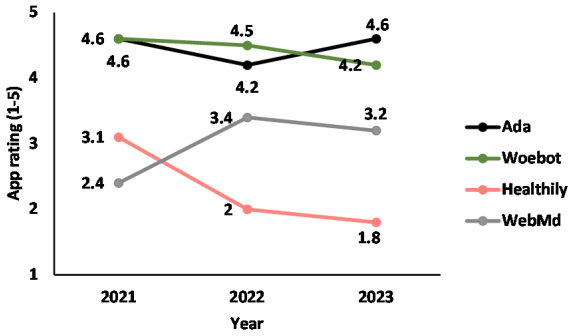

Figure 3: App rating over time.

has a lower mean rating compared to WebMD, it has a higher percentage of positive reviews.

Figure 3 shows that the ratings for Woebot (4.6 to 4.2) and Healthily (3.1 to 1.8) exhibit downward trends, while no clear upward or downward trends could be observed for Ada (4.6 → 4.2 → 4.6) or WebMD (2.4 → 3.4 → 3.2). The rating for Healthily has the largest decrease with nearly a halving of ratings during the last three years.

Figure 4 shows the relationship between the mean app ratings and the developer response rate. The data gives weak support to the proposal that there is a pattern for apps with higher developer response rates to have higher average ratings. Ada and Woebot both have high ratings and high developer response rates, while WebMD has comparatively low user ratings and developer responses. Healthily deviates somewhat from this pattern with the lowest app rating despite also having the highest developer response rate.

The written reviews give some insight as one user of the Ada app wrote: "*Was relatively ok, until today when the app refused to start. It freezes at the entry logo and that is it. Later edit: problem has been solved by the developers of the app. It's a pretty good app. True diagnostics*".

4.2 Qualitative findings

Table 2 summarizes the common strengths and weaknesses for each app.

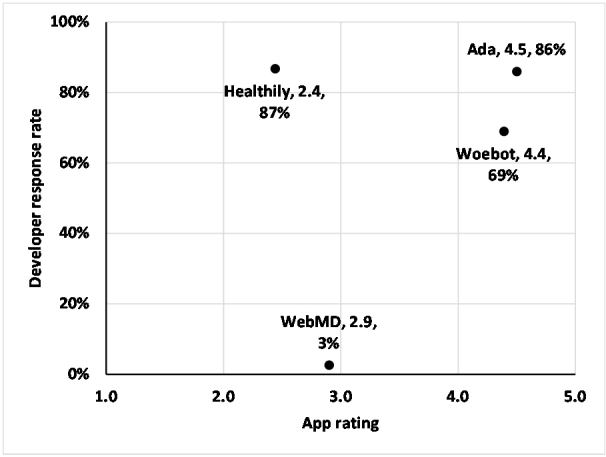

Figure 4: Developer response rate vs user's app rating.

The results indicate that many users found all four apps helpful and easy to use. It was also observed that each app had distinct strengths catering to different aspects of health management. The Ada app was recognized for its accurate health information and diagnostic capabilities. Users commended its precision and responsiveness, particularly in urgent situations. For instance, one user's experience underscores this: "*Honestly helped before going to the doctor and it was helpful. Ada symptom checkers is pretty accurate in my experience. It was right, I had bronchitis.*" Another significant aspect of Ada was its emergency care detection, as highlighted in a review: "*Very great app to assess your symptoms and find out what could be wrong with you. You can tell Ada all your symptoms and it usually is correct on the possible diagnosis. Ada will tell you when to seek emergency care or when to talk with a doctor. . .*"

A unique feature of WebMD was its allergy tracker, which has been a highly valued feature. A user stated that "*For me this is a wonderful app. When I start getting sniffles, itchy eyes, and congestion, all I do is look at the app.*" However, many users expressed dissatisfaction with the recent discontinuation of this tracker feature. As pointed out by one disappointed user "*They are taking away the main reason I got this app, the allergy tracker!... But now I will probably just uninstall...*"

Woebot was commended for its crisis detection feature and user-friendly interface. It was particularly highlighted for mental health support, as one user noted, "*I've been using Woebot for nearly 2 weeks and have found it very helpful. I especially appreciate the crisis detection - it's comforting to know that there's a safety net if things get too heavy.*" The app is also lauded for being ad-free, enhancing its usability. A user review encapsulates this sentiment: "*Great app. Was really helpful when I was just starting to learn about CBT. Continues being helpful as I find it useful to revisit the lessons and use the tools it offers. The interactive nature makes it easy and more fun to learn and do exercises. The chat interface makes it more engaging. It's completely free from what it seems, no ads, no subscriptions or anything which is GREAT. . .*"

The Healthily app was acknowledged for its health tracking features and its health article suggestions, noting that the articles add value to the user experience. This is captured in a user's review:

Table 2: Summary of positives and negatives across all the apps. Note that the Allergy tracker* for WebMD was dropped in 2022 and most reviews documented how not useful the app is without the allergy tracker.

App	Positives	No. of Reviews	Negatives	No. of Reviews
Healthily	1. Ease of use		1. Payment	32
	2. Useful trackers	4	2. Update bugs	4
	3. Article suggestions	3	3. Lack of detailed info	3
			4. Inaccuracy	7
WebMD	1. Allergy tracker*	15*	1. Instability	11
	2.Healthy guidelines	3	2. Adverts	5
	3. Ease of use		3. Limited Symptoms list	9
	4. Useful	32	4. Login issues	5
			5. Inaccuracy	8
			6. Notification abuse	7
Ada	1. Detailed	48	1. Instability	8
	2. Ease of use		2. Verification/Login	6
	3.Emergency detection	29	3. Navigation issues	3
			4. No provision to type in symptoms	7
			5. Limited symptom list	2
			6. Font size	2
Woebot	1. Ads-free	11	1. Payment	5
	2. Ease of use		2. Too generic	6
	3. Easy Navigation	32	3. Not responsive	8
	4. Good UI	64		
	5. Crisis detection	37		

"*This app is pretty good. It's suitable for people who want to improve their health. The articles are a nice touch.*"

While the four apps offer unique features that users find valuable, there were also mentioning of notable challenges. Users have encountered a range of issues from technical glitches to limitations in app functionality. The Healthily app had the highest percentage of negative reviews mainly due to its subscription fee. As stated in one review: "*Not a free app, requires an expensive subscription to use at all.*" Concerns about inaccurate information were also noted: "*. . .but always took medical advice with a grain of salt. Googling medical symptoms has better results. . .*." Users of WebMd and Ada reported issues with app stability and performance. One Ada user wrote: "*So far this app is so good, its helped me determine what I may have and its pretty accurate. Only problem is it crashes sometimes.*" There were also mentions of login and limited symptoms issues on both apps. For example, a WebMd user wrote: "*Looks okay, however I can't sign in to the app, keeps saying wrong email or password. Tried resetting password, still the same issue.*" The Woebot app received feedback about its generic responses and lack of interactive understanding as highlighted in this user review: "*The information given is great, however when the bot asks for your input, it is not read. You can type anything you want in the text box and the bot will tell you that you did well and completed the assignment. It is definitely not interactive.*"

Key user suggestions based on the reviews can be summarized as:

- Introduce 'medicine cabinet' and 'medicine saving' features.
- Facilitate search queries as an alternative to selecting symptoms from a limited list.
- Expand the list of symptoms and medications.
- Paid apps should have free versions.
- Introduce camera features allowing users to take pictures of areas of concern on their bodies.
- Introduce weekly 'activity monitoring' and 'discussion history' features.
- Improve AI algorithm accuracy and personalization to prevent inaccurate and generic responses.
- Careful curation of features based on relevance and usefulness to avoid overloading users with less useful features.

5 DISCUSSION

One possible explanation for the discrepancy between the review score and sentiment for Healthily and WebMD could be that the content and tone of the reviews can vary significantly. For example, a review with a higher score might still contain constructive criticism or neutral remarks, leading to a neutral sentiment classification. Conversely, a review with a lower score could include some positive comments, resulting in a positive sentiment. For example, one user gave the WebMD app a 5-star rating but was very critical in her review comment: "*the allergy screen is partially not visible the ad screen seems to be in the way, also it looks very plain and needs the different colors for the intensity of the pollutants and the map*". This led the comment to be classified as neutral.

The correlation between the mean rating and developer response scores were only moderate, indicating that other factors also play a significant role in determining the average rating of an app. For instance, the quality of the app, user experience, and app functionality are likely to be major factors. For example, the healthily

app, despite having the highest response rate, has a relatively low average rating. This could suggest that while the response rate is important, it cannot offset other aspects of the app that might be affecting user satisfaction. Conversely, the WebMD app, with the lowest response rate, does not have the lowest average rating. This again reinforces the idea that many factors contribute to an app's average rating.

The qualitative analysis of user reviews for all four apps reveals the emphasis users place on availability of useful specialized features, accuracy, ease of use, and responsiveness in health apps. Ada's high level of information accuracy and diagnostic capabilities, as evidenced by user experiences, underscores the importance of reliable and precise health information. The positive feedback about Ada's emergency care detection feature further highlights the need for health apps to be responsive to urgent health situations. This suggests that developers should prioritize not only the informational content but also the decision-support capabilities of health apps.

The dissatisfaction expressed by WebMD users due to the removal of the allergy tracker reveals how specific features can become integral to user experience. It emphasizes that changes in app functionalities can significantly impact user reliance and satisfaction. This scenario presents a critical learning point for developers: understanding and maintaining the features most valued by their users is crucial for sustaining app relevance and user loyalty.

Woebot's praise for its crisis detection feature and ad-free, engaging interface highlights a growing user demand for mental health support through digital platforms. However, feedback about Woebot's generic responses points towards the need for more personalized and interactive experiences in mental health apps.

The negative response to Healthily's subscription fee brings to the forefront the issue of accessibility in health apps. It underscores the need for affordable and accessible health solutions and the need for developers to balance monetization with user accessibility. Additionally, concerns about content accuracy in Healthily highlight the vital responsibility of health app developers in providing reliable and trustworthy health information, given the potential consequences of misinformation.

5.1 Limitations

This study was limited to Google Play reviews. It is thus a possibility that the results are affected by representation bias [32]. Future work could include data from other platforms such as the Apple store. More importantly, steps should be taken to corroborate the findings through triangularization methodologies relying on different sources.

Another weakness of the current study is that the authenticities of the reviews were not validated [23]. It was assumed that the reviews were valid and not manipulated. Although challenging, future work should attempt to validate review authenticity to compensate for manipulation using one of the sampling methods proposed in the literature (for example [17]).

Next, issues related to hallucination were not addressed herein. One practical limitation of modern large language models is their tendency to hallucinate [15, 38] making it harder for users to trust the provided information. Clearly, more work is needed to prevent hallucinations [13] and understand how technological hallucinations are perceived by users and how such hallucinations affect their decisions.

6 CONCLUSION

There has been a rise in the development of healthcare chatbot smartphone apps as studies have shown that chatbots can be used for "therapeutic" healthcare interventions or for at least augmenting traditional healthcare interventions. However, as reported in user reviews, many of these apps have significant problems. App developers would benefit from more insight into major user concerns to improve the quality and adoption of their apps.

While most of the users of these apps found the app useful, the analysis carried out in this study demonstrates that there is an ongoing need for improvement in areas including information accuracy, app stability, navigation, and functionality across all apps. The evolving nature of user needs and expectations calls for a user centric approach in health app development, where user feedback is continually integrated to enhance app functionality and user experience. Insight into users' perceptions and experiences with popular chatbots provided in this study can help identify app improvements to increase the user experience and user engagement.

REFERENCES

[1] Alaa Abd-Alrazaq, Zeineb Safi, Mohannad Alajlani, Jim Warren, Mowafa Househ, and Kerstin Denecke. 2020. Technical metrics used to evaluate health care chatbots: scoping review. Journal of medical Internet research, 22(6), e18301.

[2] Eleni Adamopoulou and Lefteris Moussiades Adamopoulou. 2020. Chatbots: History, technology, and applications. Machine Learning with Applications, 2(100006). https://doi.org/https://doi.org/10.1016/j.mlwa.2020.100006.

[3] Flora Amato, Stefano Marrone, Vincenzo Moscato, Gabriele Piantadosi, Antonio Picariello, and Carlo Sansone. 2017. Chatbots Meet eHealth: Automatizing Healthcare. In WAIAH@ AI* IA (pp. 40-49).

[4] Tae Hyun Baek and Minseong Kim. 2023. Is ChatGPT scary good? How user motivations affect creepiness and trust in generative artificial intelligence. Telematics and Informatics, 83, 102030.

[5] Mary Bates. 2019. Health care chatbots are here to help. IEEE pulse, 10(3), 12-14.

[6] Gerd Berget and Frode Eika Sandnes. 2019. Why textual search interfaces fail: a study of cognitive skills needed to construct successful queries. Information Research: An International Electronic Journal, 24(1), n1.

[7] Nivedita Bhirud, Subhash Tataale, Sayali Randive, and Shubham Nahar. 2019. A literature review on chatbots in healthcare domain. International journal of scientific & technology research, 8(7), 225-231.

[8] Svetlana Bialkova. 2023. How to Optimise Interaction with Chatbots? Key Parameters Emerging from Actual Application. International Journal of Human–Computer Interaction, 1-10.

[9] Marouane Birjali, Mohammed Kasri, and Abderrahim Beni-Hssane. 2021. A comprehensive survey on sentiment analysis: Approaches, challenges and trends. Knowledge-Based Systems. 226(107134). https://doi.org/https://doi.org/10.1016/j.knosys.2021.107134.

[10] Gillian Cameron, David Cameron, Gavin Megaw, R. R. Bond, Maurice Mulvenna, Siobhan O'Neill, C Armour, and Michael McTear. 2018. Best practices for designing chatbots in mental healthcare–A case study on iHelpr. In: British HCI Conference 2018. BCS Learning & Development Ltd.

[11] Jiahao Chen, Fu Guo, Zenggen Ren, Mingming Li, andJaap Ham. 2023. Effects of anthropomorphic design cues of chatbots on users' perception and visual behaviors. International Journal of Human–Computer Interaction, 1-19.

[12] Qiuyuan Chen, Chunyang Chen, Safwat Hassan, Zhengchang Xing, Xin Xia, and Ahmed E. Hassan. 2021. How Should I Improve the UI of My App? A Study of User Reviews of Popular Apps in the Google Play. ACM Trans. Softw. Eng. Methodol. 30, 3, Article 37 (July 2021), 38 pages. https://doi.org/10.1145/3447808

[13] Yuyan Chen, Qiang Fu, Yichen Yuan, Zhihao Wen, Ge Fan, Dayiheng Liu, Dongmei Zhang, Zhixu Li, and Yanghua Xiao. 2023. Hallucination Detection: Robustly Discerning Reliable Answers in Large Language Models. In Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM '23). Association for Computing Machinery, New York, NY, USA, 245–255. https://doi.org/10.1145/3583780.3614905

[14] Jayati Dev and L. Jean Camp. 2020. User Engagement with Chatbots: A Discursive Psychology Approach. In Proceedings of the 2nd Conference on Conversational User Interfaces (CUI '20). Association for Computing Machinery, New York, NY, USA, Article 52, 1–4. https://doi.org/10.1145/3405755.3406165

[15] Hanyu Duan, Yi Yang, and Kar Yan Tam. 2024. Do LLMs Know about Hallucination? An Empirical Investigation of LLM's Hidden States. arXiv preprint arXiv:2402.09733. https://doi.org/10.48550/arXiv.2402.09733

[16] Faruk Lawal Ibrahim Dutsinma, Debajyoti Pal, Suree Funilkul, and Jonathan H. Chan. 2022. A systematic review of voice assistant usability: An iso 9241–11 approach. SN Computer Science, 3(4), 267.

[17] Amir Fayazi, Kyumin Lee, James Caverlee, and Anna Squicciarini. 2015. Uncovering Crowdsourced Manipulation of Online Reviews. In Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '15). Association for Computing Machinery, New York, NY, USA, 233–242. https://doi.org/10.1145/2766462.2767742

[18] Fotos Frangoudes, Marios Hadjiaros, Eirini C. Schiza, Maria Matsangidou, Olia Tsivitanidou, and Kleanthis Neokleous. 2021. An overview of the use of chatbots in medical and healthcare education. In International Conference on Human-Computer Interaction (pp. 170-184). Cham: Springer International Publishing.

[19] Emitza Guzman and Walid Maalej. 2014. How do users like this feature? a fine grained sentiment analysis of app reviews. 2014 IEEE 22nd international requirements engineering conference. IEEE, 153-162.

[20] Samuel Holmes, Anne Moorhead, Raymond Bond, Huiru Zheng, Vivien Coates, and Michael Mctear. 2019. Usability testing of a healthcare chatbot: Can we use conventional methods to assess conversational user interfaces? In Proceedings of the 31st European Conference on Cognitive Ergonomics (ECCE '19). Association for Computing Machinery, New York, NY, USA, 207–214. https://doi.org/10.1145/3335082.3335094

[21] Mlađan Jovanović, Marcos Baez, and Fabio Casati. 2020. Chatbots as conversational healthcare services. IEEE Internet Computing, 25(3), 44-51.

[22] Tobias Kowatsch, Marcia Nißen, Chen-Hsuan I.Shih, Dominik Rüegger, Dirk Volland, Andreas Filler, Florian Künzler, Filipe Barata, Sandy Hung, Dirk Büchter, Björn Brogle, Katrin Heldt, Pauline Gindrat, Nathalie Farpour-Lambert. 2017. Text-based healthcare chatbots supporting patient and health professional teams: preliminary results of a randomized controlled trial on childhood obesity. Persuasive Embodied Agents for Behavior Change (PEACH2017).

[23] Shanshan Li, James Caverlee, Wei Niu, and Parisa Kaghazgaran. 2017. Crowdsourced App Review Manipulation. In Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '17). Association for Computing Machinery, New York, NY, USA, 1137–1140. https://doi.org/10.1145/3077136.3080741

[24] Maryam Lotfigolian, Christos Papanikolaou, Samaneh Taghizadeh, and Frode Eika Sandnes. 2023. Human Experts' Perceptions of Auto-Generated Summarization Quality. In Proceedings of the 16th International Conference on PErvasive Technologies Related to Assistive Environments (PETRA '23). Association for Computing Machinery, New York, NY, USA, 95–98. https://doi.org/10.1145/3594806.3594828

[25] Richard May and Kerstin Denecke. 2022. Security, privacy, and healthcare-related conversational agents: a scoping review. Informatics for Health and Social Care, 47(2), 194-210.

[26] Quim Motger, Xavier Franch, and Jordi Marco. 2022. Software-Based Dialogue Systems: Survey, Taxonomy, and Challenges. ACM Comput. Surv. 55, 5, Article 91 (May 2023), 42 pages. https://doi.org/10.1145/3527450t

[27] Quynh N. Nguyen, Anna Sidorova, and Russell Torres. 2022. User interactions with chatbot interfaces vs. Menu-based interfaces: An empirical study. Computers in Human Behavior, 128(107093). https://doi.org/https://doi.org/10.1016/j.chb.2021.107093

[28] Adam Palanica, Peter Flaschner, Anirudh Thommandram, Michael Li, Yan Fossat. 2019. Physicians' perceptions of chatbots in health care: cross-sectional web-based survey. Journal of medical Internet research, 21(4), e12887.

[29] Zeineb Safi, Alaa Abd-Alrazaq, Mohamed Khalifa, Mowafa Househ. 2020. Technical aspects of developing chatbots for medical applications: scoping review. Journal of medical Internet research, 22(12), e19127.

[30] Gabriel Zúñiga Salazar, Diego Zúñiga, Carlos L. Vindel, Ana M. Yoong, Sofia Hincapie, Ana B. Zúñiga, Paula Zúñiga, Erin Salazar, Byron Zúñiga. 2023. Efficacy of AI Chats to Determine an Emergency: A Comparison Between OpenAI's ChatGPT, Google Bard, and Microsoft Bing AI Chat. Cureus, 15(9), e45473. https://doi.org/10.7759/cureus.45473

[31] Frode Eika Sandnes and Hua-Li Jian. 2004. Pair-wise variability index: Evaluating the cognitive difficulty of using mobile text entry systems. In International Conference on Mobile Human-Computer Interaction (pp. 347-350). Berlin, Heidelberg: Springer Berlin Heidelberg.

[32] Nima Shahbazi, Yin Lin, Abolfazl Asudeh, and H. V. Jagadish. 2023. Representation Bias in Data: A Survey on Identification and Resolution Techniques. ACM Comput. Surv. 55, 13s, Article 293 (December 2023), 39 pages. https://doi.org/10.1145/3588433

[33] Yi Shan, Meng Ji, Wenxiu Xie, Xiaobo Qian, Rongying Li, Xiaomin Zhang, Tianyong Hao. 2022. Language Use in Conversational Agent–Based Health Communication: Systematic Review. Journal of Medical Internet Research, 24(7), e37403.

[34] Supawadee Suppadungsuk, Charat Thongprayoon, JingMiao, Pajaree Krisanapan, Fawad Qureshi, Kianoush Kashani, and Wisit Cheungpasitporn. 2023. Exploring the Potential of Chatbots in Critical Care Nephrology. Medicines (Basel), 10(10). https://doi.org/10.3390/medicines10100058

[35] Colm Sweeney, Courtney Potts, Edel Ennis, Raymond Bond, Maurice D. Mulvenna, Siobhan O'neill, Martin Malcolm, Lauri Kuosmanen, Catrine Kostenius, Alex Vakaloudis, Gavin Mcconvey, Robin Turkington, David Hanna, Heidi Nieminen, Anna-Kaisa Vartiainen, Alison Robertson, and Michael F. Mctear. 2021. Can Chatbots Help Support a Person's Mental Health? Perceptions and Views from Mental Healthcare Professionals and Experts. ACM Trans. Comput. Healthcare 2, 3, Article 25 (July 2021), 15 pages. https://doi.org/10.1145/3453175

[36] Chenxing Xie, Yanding Wang, and Yang Cheng. 2024. Does artificial intelligence satisfy you? A meta-analysis of user gratification and user satisfaction with AI-powered chatbots. International Journal of Human–Computer Interaction, 40(3), 613-623.

[37] Thea Bratteberg Ytterland, Siri Fagernes, and Frode Eika Sandnes. 2022. Perceptions of Digital Nudging for Cervical Testing: A Comparison Four Nudge Types. In International Conference on Human-Computer Interaction (pp. 212-228). Cham: Springer International Publishing.

[38] Jia-Yu Yao, Kun-Peng Ning, Zhen-Hui Liu, Mu-Nan Ning, and Li Yuan. 2023. Llm lies: Hallucinations are not bugs, but features as adversarial examples. arXiv preprint arXiv:2310.01469. https://doi.org/10.48550/arXiv.2310.01469</bib>