This is the author manuscript. The final edited version will be published on ACM DL after the CHI conference 2024. Please cite the published version:

Wang, X., Abubaker, S. M., Babalola, G. T. & Jesso, S. T. (2024, May). Co-Designing an AI Chatbot to Improve Patient Experience in the Hospital: A human-centered design case study of a collaboration between a hospital, a university, and ChatGPT. In *Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems* (pp. 1-10). https://doi.org/10.1145/3613905.3637149

Co-Designing an AI Chatbot to Improve Patient Experience in the Hospital: A human-centered design case study of a collaboration between a hospital, a university, and ChatGPT

Xin Wang

Department of Systems Science and Industrial Engineering, Binghamton University, Vestal, New York, USA, xwang314@binghamton.edu

Samer M. Abubaker

Department of Systems Science and Industrial Engineering, Binghamton University, Vestal, New York, USA, sabubak2@binghamton.edu

Grace T. Babalola

Department of Systems Science and Industrial Engineering, Human-Centered Mindful Technologies Lab, Binghamton University, Vestal, New York, USA, gbabalo1@binghamton.edu

Stephanie Tulk Jesso

Department of Systems Science and Industrial Engineering, Human-Centered Mindful Technologies Lab, Binghamton University, Vestal, New York, USA, stulkjesso@binghamton.edu

Abstract

Patient experience (PX) is an important reflection of healthcare quality and is highly related to patient health outcomes and hospital reputation of within the communities they serve. PX data reported by patients is also crucial for hospitals to improve the services they provide, however, current approaches to survey and analyze PX data have many limitations. Our team collaborated with United Health Services (UHS), a New York healthcare system, to co-design a prototype chatbot

application for patients to use while in the hospital, yielding more accurate PX data, but also an opportunity for staff to respond in real-time. We discuss our human-centered design process, which entailed interviews, data mining, qualitative analysis, and the application of ChatGPT and other algorithms to recognize relevant PX complaints from natural language data. Through ongoing collaboration, we are developing a chatbot application to elicit PX feedback and allow PX experts to improve patient experience in real-time.

Keywords

healthcare, patient experience, human-centered design, natural language processing, ChatGPT, large language model, data mining, qualitative analysis

1 INTRODUCTION

Patient experience (sometimes abbreviated as "PX") has emerged as a vital component of healthcare quality, linked to clinical outcomes and hospital reputation. Studies have found that higher patient satisfaction is associated with improved clinical measures like reduced readmission rates [1], while positive experiences lead to higher levels of trust and loyalty to healthcare providers [2]. Furthermore, patient experience scores directly impact hospital rankings and reimbursement rates as part of the HCAHPS (Hospital Consumer Assessment of Healthcare Providers and Systems) survey and value-based purchasing programs [3, 4]. However, traditional methods of gathering patient feedback via surveys administered after discharge have limitations including low response rates, biased samples, recall errors, and ineffective timing [5]. Without access to timely and informative patient experience and patient satisfaction indicators, healthcare organizations can struggle to identify and rectify causes and contributors to poor patient experience. Intelligent conversational agents like chatbots present a new opportunity to capture timely and accurate data on patient experience, and even intervene while patients are in the hospital to improve patient experience directly. Chatbots could be used to engage patients in natural dialogues at various touchpoints during their visit or stay, promote higher response rates, and provide real-time assistance [6]. Despite this potential, there is limited research on integrating chatbots into patient experience initiatives and evaluating their impact.

We conducted human-centered design research in collaboration with UHS, a New York based healthcare organization, to develop an early-stage chatbot application which could assist them with collecting more valuable and informative data on patient experience. Over the past six months, we have conducted qualitative and data mining research to develop a medium-fidelity chatbot prototype and design specifications which can satisfy security and usage requirements at UHS. Our findings contribute empirical insights into the desired functionality and potential value of chatbots for improving patient experience in hospitals.

2 RELATED WORK

In the realm of healthcare, chatbots have gained attention for their potential to complement traditional on-site interactions between patients and health professionals. Chatbots integrated into mobile applications have shown promise in providing accessible cognitive behavioral therapy for mental health conditions like panic disorder [8]. Specialized chatbots have been developed to provide mental health therapy [9, 10], alcohol use disorder [8] and childhood obesity [7], showcasing the versatility of chatbot technology in healthcare contexts.

Ensuring an overall positive experience between patients and healthcare systems is vital for maintaining a patient's engagement in their own treatment and/or recovery, and for optimizing healthcare services and outcomes. Patient feedback data reflects the patient's needs and attitudes toward the healthcare service they've received, and is beneficial for the healthcare provider to improve their service [11]. Here, chatbots present opportunities in collecting data from patients to recognize and improve patient experience. When used as a tool for data collection, researchers have noted the importance of investigating the usability and acceptance of chatbot interactions as an alternative to traditional surveys [12]. One study compared patients' user experiences with a virtual conversational chatbot and a traditional online form and found that 70% of users prefer giving patient experience feedback through chatbots [13]. In the domain of patient experience data collection and analysis, some studies emphasize not only the collection of data but also how it is imperative to utilize this data to

improve care [11]. This includes understanding differences in patient experiences with respect to individual medical conditions [14], and individual characteristics. To ensure inclusivity and effectiveness, researchers have explored the design of chatbots catering to specific demographics who may experience additional barriers to care, such as Black Americans with chronic conditions [15, 16]. Language barriers are also impediments to capturing accurate patient experience data, and researchers have created multilingual chatbots to engage patients with limited proficiency in English [17, 18]. Overall, while some are recognizing the potential benefits of chatbots in the healthcare sector, more research is necessary to recognize exactly how healthcare organizations can make use of chatbots to improve patient experience.

3 METHODS

We conducted early-stage iterative design using the Human-Centered Design Cycle [19]. Our design team began by conducting interviews with the PX Team at UHS to assess their needs for a patient experience chatbot. The PX Team routinely uses PX data to investigate and address organizational challenges to providing high-quality patient experience. Insights gained from these interviews led us to develop an initial set of design specifications, which we presented to our collaborators at UHS for their feedback. This feedback in turn informed our second research method, which included mining patient reviews of the organization provided through Google Reviews, and qualitatively analyzing these reviews to recognize common types of patient experience and care quality complaints that were pertinent to our intended application. We then applied three types of machine learning (ML) approaches to automatically code this data, including ChatGPT, Convolutional Neural Network (CNN), and Long Short-Term Memory (LSTM), and compared their output to human performance. In following sections, we will discuss how these methods informed our current prototype and design specifications. All human-subjects research was overseen by UHS's Institutional Review Board.

3.1 Interviews

3.1.1 Participants and procedures.

We conducted semi-structured interviews with individuals on or connected to the PX Team at UHS (n = 4, mean years of experience = 19.8, SD = 14.2) to assess their needs for a chatbot application, and to gain a deeper understanding of the context and different aspects of patient experiences that they routinely examine. Interview questions focused on recognizing how the PX Team currently assesses PX, how they get their data, the major features they think about and know to affect PX, and what they would desire from an ideal PX surveying tool. Questions are presented in Table A in the Appendices. Interviews were conducted over Zoom software, and a transcript was automatically generated and stored as data. Interviews lasted 1 hour, and afterwards, participants were thanked for their participation.

3.1.2 Analysis and results.

Interview transcripts were qualitatively analyzed using content analysis. The team iteratively reviewed data and developed codes, and periodically assessed and improved agreement through discussion and code refinement until a final set of codes were established. These codes were applied to all statements made by participants by two raters (XW and SA) working independently. Rater agreement and interrater reliability were evaluated and reported below.

Table 1 shows a summary of the results from the content analysis. Seven codes were established, including Data Collection used by the PX team, User Experience related to how people currently report or use PX data, PX Influencing Factors which were brought up by PX team members as common and/or significant, Patient Characteristics including demographics that were top-of-mind to the PX team, Data Analysis, Types of Data which are used, and Sharing Findings from the data with others at UHS.

Table 1: Content analysis results on interview scripts.

Codes	Definition	Representative quote	Total count	Percent agreement	Cohen's Kappa

Data Collection	Collection of data, including surveys, 3rd party services or other employees relaying information to PX team	"For in-patients, they typically send home a letter, like a paper copy survey. And then an outpatient world I believe it's typically an email link where they're sent the survey."	86	94%	0.94
User Experience	Experiences with how patients and staff interact with existing types of PX collection or analysis tools	"Most people are much more likely to respond to an electronic invitation then to open up their mail and fill out a piece of paper with a pen and a pencil and return it by mail."	61	97%	0.97
PX Influencing Factors	Factors which influence patient experience	"The food: obviously we're looking at temperatures and courtesy of the person serving food."	50	97%	0.97
Patient Characteristics	Characteristics of patients, such as demographics like sex, age, ethnicity or race	"But we could do everything from, you know, bringing it down by type of patient. So I could see, you know, is our patients happy based on their age?"	26	99%	0.99
Data Analysis	Concepts related to data analysis used by the PX team	"I utilize all of that information to kinda see, you know, if grievances are high, satisfaction is low, you have kind of validated our potential opportunity [for improvement]...so I use [software] for that purpose to kinda like validate one or over the other."	17	99%	0.99
Types of Data	Any information pertaining to patient experience which is used as data	"[Surveys include measures of] likelihood to recommend, nursing communication and provider communication."	14	98%	0.98
Sharing Findings	Sharing findings of patient experience with others, including presentations to organization leaders or feedback to individual members of staff	"And we would talk to the nurse managers about, you know, what their unit scores were, looking at scores, but then also looking at comments the patients would write."	14	99%	0.99

Based on the content analysis, we found that PX data collection and analysis can be improved in a number of ways. First, we learned that the hospital must use other forms of data collection in addition to HCAHPS surveys due to various limitations. Patients find this survey too long and cumbersome to fill and mail back, leading to a low response rate. A PX member imagined a better tool to be, "s*omething that allows for a lot of different ways of responding or reaching out to people. And then just something that's going to really choose what are the most important questions"*; and this was key because *"not everyone has, you know, half an hour to take a survey" (Participant 1, PX Coordinator*). The same participant noted that "*a lot of [older patients] don't love stuff on the internet or don't even bother with it*", indicating the need to consider preferences in different populations carefully within design.

We learned that UHS already uses a number of third-party platforms to increase access patient feedback data, including Reputation.com, which compiles hospital reviews across social media like Google Reviews, Facebook, and Yelp. They also use Press Ganey, which is a third party survey and analysis service, and CipherRounds. Participant 1 expressed that CipherRounds, which is an app installed on PX Team smartphones, is helpful because it allows them to survey patients on-the-spot and communicate with other hospital departments in real-time:

> [CipherRounds] does a lot to also notify other departments, like if someone says their food's been cold, it'll send an automatic message to the food team and they can send someone up to talk to

the patient. So that's kind of a more on-the-spot type survey that we do. Aiming to try and help, you know, that patient's experience. (Participant 1)

However, they noted that a major limitation was the inability of any staff to round with enough frequency to collect much data or use the real-time response feature. PX Team members pointed to various causes and contributors of low patient experience, including dissatisfaction with care, professionalism of staff, cleanliness and comfort in room, and dissatisfaction with food, but communication was indicated as something that was very influential: "*When I worked in patient satisfaction, I used to say, 'it's like real estate. Real estate is location, location, location. With satisfaction, it's communication, communication, communication" (Participant 4, a hospital Senior Director)*. Overall, insights gained from interviews led to our recognition that a chatbot would be best positioned to capture patient experience data if patients could interact with it while hospitalized or in the emergency department (ED), and that there were many possibilities on how that data could be made valuable to UHS.

3.1.3 Initial design concept and feedback from UHS.

After conducting and analyzing interviews, we developed our initial design concept, which involved a web-or app-based chatbot that could respond in natural language to encourage users to share their hospital experiences as a form of data collection. The intended context of use would be in the hospital or ED so that patients could describe their experiences more fully as they were occurring, rather than as a recollection after leaving the hospital. We thought that patients could use their phones or hospital units could provide tablets to individuals who would not normally attempt to connect to the chatbot through their own phones (e.g., older patients). We then discussed our interview findings and design concept with our collaborators at UHS to get their feedback. Due to privacy and security standards at UHS, it was determined that the most viable platform would be a web-based application that could be hosted within the organizational firewall, which would afford additional functionality if the chatbot knew which patient they were interacting with. If the chatbot could use the patient's ID, it could also synchronize with data in the electronic health record, enabling the application to be used for communication between the patients and their care team. It could also summarize patient complaints and provide insights to PX Team members in real-time, allowing them to locate patients who were having bad experiences and intervene. While this new concept was desirable and exciting to our UHS collaborators and our design team, this constituted a major expansion of scope that we then adopted in our next design iteration.

3.2 Data Mining

Based on what we learned from the interviews and discussions with our UHS collaborators, and the expansion of our scope to include a chatbot that collects data and facilitates patients getting help in real-time, we decided to collect data on patient complaints by scraping Google Review data in order to learn what types of issues could be identified from natural language responses.

3.2.1 Data collection and analysis.

We collected patients' reviews from Google Review pages of 4 UHS hospitals using a Pythonic web scraper. These 4 UHS hospitals are UHS Wilson Medical Center, UHS Binghamton General Hospital, UHS Chenango Memorial Hospital, and UHS Delaware Valley Hospital. For each patient's review, we also collected star rating score from Google Review, which is given by the patient to express the level of satisfaction on their hospital visiting experience. As shown in Table 2 on the statistics of the collected reviews, UHS Wilson Medical Center has 52.91% non-five star reviews, while all the reviews of UHS Delaware Valley Hospital are from full star reviews. UHS Binghamton General Hospital and UHS Chenango Memorial Hospital have 29.79% and 26.79% non-five star reviews, respectively.

Table 2: The statistics on the collected reviews of the UHS hospitals.

Hospital	Number of reviews	Number of full star rating (5 star)	Number of non-five star rating (1-4 star)	Non-five star rate
UHS Wilson Medical Center	344	162	182	52.91%

UHS Binghamton General Hospital	376	264	112	29.79%
UHS Chenango Memorial Hospital	112	82	30	26.79%
UHS Delaware Valley Hospital	33	33	0	0.00%

Our study focused on non-five star reviews to identify patient complaints, so we selected 159 non-five star (1-3 star) reviews from UHS Wilson Medical Center for analysis. UHS Wilson Medical Center was selected, because it is the main hospital site at UHS and a regional referral center providing different high-level medical and surgical services, such as cardiac surgery and cancer care. We then used thematic analysis to qualitatively analyze each review and generated themes. Using our themes and human-coded data, we then trained and compared the performance of ChatGPT and two traditional ML models to human performance as a step to evaluate if these approaches to NLP (natural language processing) would provide adequate recognition of common PX complaints.

3.2.2 Thematic analysis and results.

Our team developed seven relevant themes based on what we found to be pertinent characteristics of patient experience through interviews and collaboration of UHS partners, including Missing Essentials, Quality or Safety Concern, Professionalism/Competence, Comfort in Facility, Waiting on Tests, Waiting on Treatment, and Waiting on Clinician (in Table 3). We then reviewed and coded the negative PX reviews according to these themes. The most frequently occurring code was Professionalism/Competence, often related to a complaint about rude staff. Waiting on Tests, Treatment, and Clinician were also frequent complaints. Some of the patients commented that they didn't know how long time they still needed to wait there, since there is no notification or communication from any hospital staff. We also found a number of complaints related to Patient Safety and Quality issues, such as nearly receiving the wrong medication, and placing an IV (intravenous line) with dirty gloves or without gloves or hand sanitization. Patients also pointed to Missing Essential items as a contributor to bad experience, for instance food, a bathroom, and a wheelchair. These examples represent the input that a PX chatbot could receive directly from patients and family, and could be directed to PX Team members for real-time interventions, and are considered as another component that will be incorporated into our chatbot prototype.

Table 3: Thematic analysis results on patient reviews of UHS hospital Google Review.

Themes	Definition	Representative quote	Total count	Percent agreement	Cohen's Kappa
Missing Essentials	Complaints related to inability to access basic essentials, like food and water, or even medical equipment	"The worse part was when they did not have supplies to access my chest port, my daughter had to go back home to get my supplies that I had."	15	95%	0.95
Quality or Safety Concern	Concerns over errors, unsafe conditions, or the quality of care.	"Almost given the wrong medication twice - stopped by family member."	38	89%	0.88
Professionalism/ Competence	Complaints related to staff being unprofessional or incompetent, including clinical and non-clinical staff	"Zero bedside manner staff is extremely egotistical."	93	92%	0.81
Comfort in Facility	Comfortability of the hospital environments	"Rooms so small and such a lack of privacy! Every time a visitor came for my roommate, they have to walk by my bed, bumped into it, were terribly nosey!"	33	91%	0.90
Waiting on Tests	Waiting for tests to be performed, or for results from a medical test.	"The RNs were very nice but it felt like I waited a very long time for	73	72%	0.66

		any doctors to talk to me and for any MRI or CT scans to be done."			
Waiting on Treatment	Waiting for a medical treatment, including waiting for medication	"They order the drip. An hour later had not arrive and the RN said they would have to reschedule me for another day."	49	74%	0.73
Waiting on Clinician	Waiting for a clinician to come	"We took my daughter to the walk in clinic and were told to take her to ER, after six hours she still hasn't seen a doctor yet."	32	86%	0.85

3.2.3 Applying ChatGPT and traditional ML models to automatically perform thematic analysis.

Next, our study explored ML methods of automatically applying our themes, including the use of ChatGPT, a Large Language Model (LLM) trained by OpenAI, which can generate human-like responses to progress a conversation [20]. Our goal was to see if our application could use an LLM within the chatbot to identify our seven major themes automatically, which could then be used in different workflows to get assistance to patients in the hospital in real-time. We developed a Pythonic script that passed our individual themes and definitions to ChatGPT (see Table 4), then asked it to code each individual review. We then had our script put all ChatGPT-coded output into a data frame to compare it to the coded data generated by human raters.

Table 4: The designed ChatGPT prompt.

Prompt
1. Label the text as "Missing Essentials" if the reviewer indicates that they needed something and were upset or unable to get the object. For instance, if the patient is hungry or thirsty, not having access to food or drinks, or the food is very bad, or needing things to make a patient more comfortable or safe, like a blanket or a pillow or wheelchair, or even needing to charge their phone and not having a phone charger. It could even include needing a medical device that is out of stock or inaccessible at the hospital 2. Label the text as "Waiting on Tests" if the reviewer is complaining that they have not received the results from a medical test, or they are waiting for someone to come and collect a sample to run a medical test in order to understand their medical conditions or prescribe a treatment. This could include waiting for someone to perform an imaging procedure, like an MRI, or if the patient has not received the results from a lab testing their blood or urine samples. This can include other instances of waiting on tests. 3. Label the text as "Quality or Safety Concern" if the reviewer complains about something that they experience that is unsafe or dangerous. This could include being prescribed the wrong medication or given the wrong medication or almost given the wrong medication, waiting in the emergency room too long and having additional illness due to a very long wait, getting an infection from the hospital, and falling in the hospital. 4. Label the text as "Professionalism/Competence" if the reviewer complains that the staff within the hospital is rude, incompetent, bad at their job, makes mistakes, doesn't help or is not respectful. Staff can include receptionists, nurses, physicians, nursing aids, or anyone else working at the hospital. 5. Label the text as "Comfort in Facility" if the reviewer complains that the environment is uncomfortable for any reason. For instance, if they complain that it is dirty, crowded, old, loud, or lacks privacy. 6. Label the text as "Waiting on Treatment" if the reviewer complains that they are waiting for a medical treatment. For instance, if they are waiting for medication or waiting for a procedure, surgery, or waiting to be admitted into the hospital to get care. 7. Label the text as "Waiting on Clinician" if the reviewer complains that they are waiting for a clinician to give them medical care. For instance, they are waiting on clinical staff, a nurse, a provider or a doctor to help them. === Please label each sentence with any of the defined labels. Please reply only with the defined codes that you want to label for the sentence.

Additionally, we used two traditional ML models, applied with deep neural networks, to compare to the performance of ChatGPT: CNN and LSTM. With these two models, we trained a binary classifier for each defined theme (in Table 2). The CNN model was constructed with one embedding layer, one 1-D convolutional layer, one pooling layer, two dropout layers, and two dense layers. The LSTM model was set with one embedding layer, two LSTM layers, two dropout layers, and two dense layers. For both CNN and LSTM, adam optimizer, loss function of binary cross entropy, and accuracy metrics were performed for model compiling purposes. We randomly selected 115 comments as the training set and took the remaining 44 comments as the testing set. Table 5 shows the performance of all methods when compared to human raters as a percent agreement.

Table 5: Theme labeling performance by ChatGPT and ML models.

Themes	Missing Essentials	Quality or Safety Concern	Professionalism/ Competence	Comfort in Facility	Waiting on Tests	Waiting on Treatment	Waiting on Clinician
Humans to ChatGPT	92%	78%	52%	47%	53%	65%	72%
Humans to CNN	70%	64%	57%	55%	73%	52%	59%
Humans to LSTM	77%	66%	59%	73%	52%	64%	73%

Table B in the Appendices shows example output generated by ChatGPT. The model performance of the CNN and LSTM for each code's classifier is evaluated through three performance parameters, including accuracy, F1 score, and AUC score, in Table C and Table D in the Appendices. ChatGPT had the best overall performance. As seen in Table 5, all three types of ML outperform the other models in match to human raters for some themes, and all show poor performance on other themes. In other words, each model seems to have unique strengths and weaknesses, and therefore, a Mixture of Experts (MoE) approach could be used to improve performance and achieve a closer match to human judgment [21].

3.2.4 Discussion on data mining and application of ChatGPT and other ML models.

Our analysis of PX data on Google Reviews and application of ChatGPT and ML models led us to conclude that this is a promising method for automatically tagging relevant kinds of PX issues, which can be incorporated into an application to improve how care is provided to patients. Further model refinement can be accomplished in additional design iterations and after receiving more feedback from UHS regarding their needs, and how output could be integrated into various staff workflows to improve PX in real-time.

4 PROTOTYPE PATIENT EXPERIENCE CHATBOT

An AI (artificial intelligence) chatbot was designed based on the insights gained through the iterative design process. From interview, we learned that a chatbot might serve as a channel for communication, which is needed by both healthcare organizations and patients. For healthcare organizations, they need to hear the reason for why patients have bad experiences during their healthcare services, and this was a struggle given that patients typically only reported their experiences in retrospect, sometimes weeks after leaving the hospital, if they chose to share them at all. For patients, they need a quick way to request help when they need it, stay informed on their wait time and relevant details about their care, and a way to report their perceptions in the moment to provide a more accurate account of their experiences. As a first step towards addressing these needs, our current prototype uses OpenAssistant LLaMA-based model, a large language model pre-trained with 161,433 messages in 35 different languages [22], and searches for the relevant themes within text from patients who interact with the chatbot. The natural language responses are meant to engage patients in longer interactions to elicit more data about their experiences, which can be used by UHS in efforts to improve care and experience. We will later incorporate a MoE approach into this chatbot to improve its ability to detect relevant complaints. Figure 1 shows an early-stage version of the patient-facing interface of the proposed AI chatbot application. This AI chatbot is developed by Streamlit framework and can be visited through the application link (https://app-ai-chatbot-xwang.streamlit.app/). The intention is to embed this

chatbot as a web-based application that can be incorporated into UHS's website and within their firewall to ensure privacy, and patients could only use it when connected to the UHS Wi-Fi. Patients can use a tablet provided by staff or scan a QR (Quick Response) code placed in individual rooms which allows for tracking of exactly which patients are interacting with the chatbot. The functionality related to what occurs once UHS captures and tags this real-time data has yet to be developed, and will be part of the next design cycle conducted along with UHS. Additionally, we will further explore potential ways to build the connection between the proposed AI chatbot and the hospital's electronic health records to facilitate communication of relevant status updates between the organization and the patients. This may be especially useful for patients who are less likely to use existing apps (e.g., older patients), or patients who do not speak English as their primary language.

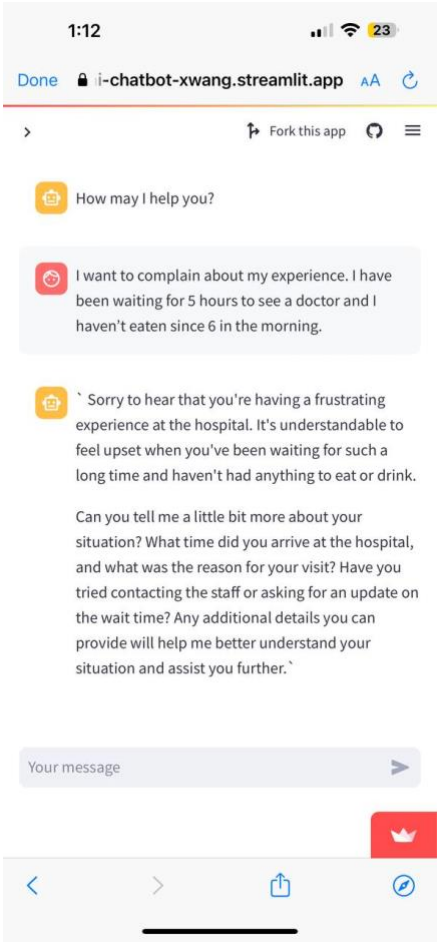

Figure 1: Early-stage AI chatbot prototype.

5 LESSONS LEARNED

During our study, we learned several meaningful lessons that could be helpful for other researchers as they attempt to use ChatGPT and other LLMs within their research and designs:

(1) In our study, we conducted interviews, mined Google Review data, applied qualitative analysis, and finally used deep learning approaches to test capabilities for automatic text coding to recognize PX concerns from real natural language

patient complaints. We first learned that this type of approach yields great insights into user needs, and it served as a strong foundation for co-designing the chatbot with UHS.
(2) When applying machine learning approaches, we found utility as well as limitations. With ChatGPT, it would sometimes randomly lose memory on prompts or be distracted by the previous information for unknown reasons, making it unreliable in some instances. We found that it is necessary to create programming to repeatedly prompt ChatGPT in a certain time span to ensure that it follows the prompt and replies to codes with high quality.
(3) In this study, we applied two traditional ML approaches, including CNN and LSTM. While ChatGPT had better overall performance, and did not require us to provide any training data, both ML models achieved better performance than ChatGPT for some thematic codes, therefore indicating potential utility in the MoE approach. Both ML methods were trained on a relatively small set of labeled data, and performance would likely be improved with the addition of more human-labeled codes, but this takes a considerable amount of time.

6 LIMITATIONS

To analyze patients' help requests during healthcare service, we collected patients' reviews only from UHS hospital pages on Google Review. However, as we found in interview process, the patients also give their reviews through other social media, including Twitter, Yelp, and Facebook. In the future, we can continue collecting more patients' reviews from different sources to gain further insights into patient experience needs in the hospital settings.

At present, since our prototype AI chatbot utilizes a pre-trained OpenAssistant LLaMA-based model from Hugging Face, users need to have a Hugging Face account for login to access this chatbot prototype. Therefore, we can use this prototype to perform user tests and elicit more feedback, but will eventually need to create our own web-based application with a built-in LLM.

Additionally, we have only begun to consider the potential security or systems issues introduced when patients use this application. Future work will focus on developing features and functions to protect patient users' privacy when using the chatbot. We also anticipate the need for topic filtering to avoid specific conversations that could lead users to leak personal information, engage in toxic interactions, or inappropriately elicit medical advice from the chatbot.

Moreover, our current case study is only focused on UHS, which is a US based non-profit healthcare system in New York State, therefore others who wish to develop similar applications may need to tailor this approach to suit their own unique context of healthcare delivery.

7 CONCLUSION

Our team conducted human-centered research and design activities to develop a chatbot prototype that could recognize common patient experience challenges to aid in data collection and the development of interventions aimed at improving patient experience in real-time. The insight gained from interviews with PX professionals, as well as our ongoing conversations with UHS led our team to dramatically expand the original scope of our project, from an application that used AI to elicit better PX feedback from patients, into something more akin to an AI "concierge", which not only collected feedback, but also enabled the PX Team to sense and respond to poor experiences in real-time. We then scraped patient experience data from Google Reviews, qualitatively analyzed it, and applied ChatGPT along with other traditional ML approaches in order to explore the reliability of this type of automatic thematic coding by AI, which could be employed in an AI chatbot. Finally, we created design specifications for a PX chatbot application, and an AI chatbot prototype, powered by OpenAssistant LLaMA-based model, that can provide human-like responses to engage with patients and elicit better PX data that can be used by UHS to improve care. In future work, through co-design with our UHS collaborators, we will continue to iteratively refine our design so that it may one day support patients and hospital staff as they strive to improve patient experience in the hospital.

ACKNOWLEDGMENTS

We thank our participants for sharing their valuable insights with us. We thank our collaborators at UHS, who will continue to partner with us as we improve our prototype. We also thank our reviewers for their constructive feedback on this research.

REFERENCES

[1] Boulding, W., Glickman, S. W., Manary, M. P., Schulman, K. A., & Staelin, R. (2011). Relationship between patient satisfaction with inpatient care and hospital readmission within 30 days. The American journal of managed care, 17(1), 41-48.

[2] Lee, Y. Y., & Lin, J. L. (2010). Do patient autonomy preferences matter? Linking patient-centered care to patient–physician relationships and health outcomes. Social science & medicine, 71(10), 1811-1818.

[3] Sofaer, S., Crofton, C., Goldstein, E., Hoy, E., & Crabb, J. (2005). What do consumers want to know about the quality of care in hospitals?. Health services research, 40(6p2), 2018-2036.

[4] Chen, H. C., Cates, T., Taylor, M., & Cates, C. (2020). Improving the US hospital reimbursement: how patient satisfaction in HCAHPS reflects lower readmission. International Journal of Health Care Quality Assurance, 33(4/5), 333-344.

[5] Bland, C., Zuckerbraun, S., Lines, L. M., Kenyon, A., Shouse, M. H., Hendershott, A., ... & Djangali, A. L. (2022). Challenges facing CAHPS surveys and opportunities for modernization.

[6] Laranjo, L., Dunn, A. G., Tong, H. L., Kocaballi, A. B., Chen, J., Bashir, R., ... & Coiera, E. (2018). Conversational agents in healthcare: a systematic review. Journal of the American Medical Informatics Association, 25(9), 1248-1258.

[7] Kowatsch, T., Nißen, M., Shih, C. H. I., Rüegger, D., Volland, D., Filler, A., ... & Farpour-Lambert, N. (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).

[8] Haque, M. R., & Rubya, S. (2023). An Overview of Chatbot-Based Mobile Mental Health Apps: Insights From App Description and User Reviews. JMIR mHealth and uHealth, 11(1), e44838.

[9] Zhang, J., Oh, Y. J., Lange, P., Yu, Z., & Fukuoka, Y. (2020). Artificial intelligence chatbot behavior change model for designing artificial intelligence chatbots to promote physical activity and a healthy diet. Journal of medical Internet research, 22(9), e22845.

[10] Koulouri, T., Macredie, R. D., & Olakitan, D. (2022). Chatbots to support young adults' mental health: an exploratory study of acceptability. ACM Transactions on Interactive Intelligent Systems (TiiS), 12(2), 1-39.

[11] Coulter, A., Locock, L., Ziebland, S., & Calabrese, J. (2014). Collecting data on patient experience is not enough: they must be used to improve care. Bmj, 348.

[12] Pittman, Z. C., John, S. G., & McIntyre, C. W. (2017). Collection of daily patient reported outcomes is feasible and demonstrates differential patient experience in chronic kidney disease. Hemodialysis International, 21(2), 265-273.

[13] Soni, H., Ivanova, J., Wilczewski, H., Bailey, A., Ong, T., Narma, A., ... & Welch, B. M. (2022). Virtual conversational agents versus online forms: Patient experience and preferences for health data collection. Frontiers in Digital Health, 4, 954069.

[14] Pittman, Z. C., John, S. G., & McIntyre, C. W. (2017). Collection of daily patient reported outcomes is feasible and demonstrates differential patient experience in chronic kidney disease. Hemodialysis International, 21(2), 265-273.

[15] Kim, J., Muhic, J., Robert, L. P., & Park, S. Y. (2022, April). Designing chatbots with black americans with chronic conditions: Overcoming challenges against covid-19. In Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (pp. 1-17).

[16] Kwame, A., & Petrucka, P. M. (2021). A literature-based study of patient-centered care and communication in nurse-patient interactions: barriers, facilitators, and the way forward. BMC nursing, 20(1), 1-10.

[17] Al Shamsi, H., Almutairi, A. G., Al Mashrafi, S., & Al Kalbani, T. (2020). Implications of language barriers for healthcare: a systematic review. Oman medical journal, 35(2), e122.

[18] Rainey, J. P., Blackburn, B. E., McCutcheon, C. L., Kenyon, C. M., Campbell, K. J., Anderson, L. A., & Gililland, J. M. (2023). A Multilingual Chatbot Can Effectively Engage Arthroplasty Patients With Limited English Proficiency. The Journal of Arthroplasty.

[19] Harte, R., Glynn, L., Rodríguez-Molinero, A., Baker, P. M., Scharf, T., Quinlan, L. R., & ÓLaighin, G. (2017). A human-centered design methodology to enhance the usability, human factors, and user experience of connected health systems: a three-phase methodology. JMIR human factors, 4(1), e5443.

[20] Ray, P. P. (2023). ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. Internet of Things and Cyber-Physical Systems.

[21] Yuksel, S. E., Wilson, J. N., & Gader, P. D. (2012). Twenty years of mixture of experts. IEEE transactions on neural networks and learning systems, 23(8), 1177-1193.

[22] Köpf, A., Kilcher, Y., von Rütte, D., Anagnostidis, S., Tam, Z. R., Stevens, K., ... & Mattick, A. (2023). OpenAssistant Conversations--Democratizing Large Language Model Alignment. arXiv preprint arXiv:2304.07327.

APPENDICES

Table A: Interview questions.

No.	Interview question
1	Can you describe your job role?
2	How long have you worked at UHS?
3	How long have you worked in healthcare?
4	How do you gauge patient experience?
5	What kind of data do you use most frequently to understand patient experience and patient needs?
6	Can you describe your current process of surveying patient experience? What software, survey measure, and analysis methods do you use? Do you like these?

7 How well do your current methods of data capture and data analysis explain patient experience, and patient needs?

1. Is there any data or deeper understanding of patient experience that you wish you had?

2. Do you feel like the data shows a fair representation of strengths and weaknesses of care at UHS?

3. Are there any demographic groups (for instance, older people, or members of some specific ethnicity group) which you think are not represented in the data you have? If so, do you know why?

8 What do you think are the main factors that influence patients' feedback?

1. Are there general categories, like facilities, or clinical staff experience, that you think are most influential?

9 If you were able to design the best tool to capture and analyze patient experience data, what features would it include?

1. How could this tool encourage more participation and better data from patients?

2. How could this tool help you and your organization better understand patient needs and experience?

Table B: Representative code results by ChatGPT.

Patient review	ChatGPT response
I am currently sitting in a hallway with my ninety-eight year old father. We arrived by ambulance at 9 AM. We sat in the Emergency Room for three hours before being seen by a Doctor. Finally, seven hours later, a Doctor drained the fluid from his twisted knee. The Doctor wanted him to stay in the hospital for treatment for a few days. The staff moved him to a hallway like it was a loading zone for patients that want rooms. This is how the masses are treated. I wonder how Bethesda that's their patients?	• Waiting on Clinician • Comfort in Facility
53 hours without food waiting for tests...Would you want this for your 88 yr old father? Different doctor and different nurse EVERY shift. No notes read prior to interaction with patient, staff didn't know patient's needs or why they were there/what has happened. Almost given the wrong medication twice - stopped by family member. The assigned doctors appear unaccountable & unreachable. No communication/commitment on what is going on, when patient can have food, when tests will occur. Waited 17 hrs in chaotic ER for a hospital bed after entering. I pray for anyone entering without a family advocate. Wilson Admin - God's speed in improving for when any of your family members need help.	• Missing Essentials • Quality or Safety Concern • Professionalism/Competence • Waiting on Treatment

Table C: CNN model performance for each code.

Performance parameter	Missing Essentials	Quality or Safety Concern	Professionalism/ Competence	Comfort in Facility	Waiting on Tests	Waiting on Treatment	Waiting on Clinician
Accuracy	0.719	0.562	0.566	0.640	0.621	0.547	0.683
F1	0.754	0.618	0.556	0.634	0.636	0.527	0.675
AUC	0.698	0.565	0.552	0.654	0.426	0.558	0.573

Table D: LSTM model performance for each code.

Performance parameter	Missing Essential	Quality or Safety Concern	Professionalism/ Competence	Comfort in Facility	Waiting on Tests	Waiting on Treatment	Waiting on Clinician
Accuracy	0.701	0.533	0.543	0.621	0.610	0.501	0.661

F1	0.733	0.620	0.653	0.600	0.651	0.519	0.654
AUC	0.640	0.573	0.669	0.610	0.566	0.527	0.552