http://dx.doi.org/10.14236/ewic/BCSHCI2023.11

Supporting dementia caregivers in Peru through chatbots: generative AI vs structured conversations.

Fernanda Espinoza
Imperial College London
London
fe319@imperial.ac.uk

Darren Cook
Imperial College London
London
darren.cook@imperial.ac.uk

Chris R. Butler
Imperial College London
London
christopher.butler@imperial.ac.uk

Rafael A. Calvo
Imperial College London
London
r.calvo@imperial.ac.uk

In Peru, dementia caregivers face burnout, depression, stress, and financial strain. Addressing their needs involves tackling the intricacies of caregiving and managing emotional burdens. Chatbots can serve as a viable support mechanism in regions with limited resources. This study delves into the perceptions of dementia caregivers in Peru regarding a chatbot tailored to offer care navigation and emotional support. We divided the study into three phases: the initial stage encompassed engaging stakeholders to define design requirements for the chatbot; the second stage focused on the creation of 'Ana', a chatbot for dementia caregivers; and the final stage assessed the chatbot through interviews and a caregiver satisfaction survey. 'Ana' was tested in two configurations - one employed pre-defined conversation patterns, while the other harnessed generative AI for more dynamic responses. The findings reveal that caregivers seek immediate access to information on handling behavioural symptoms and a platform for emotional release. Moreover, participants preferred the generative AI alternative of Ana, as it was perceived to be more empathic and human-like. The participants valued the generative approach despite knowing the potential risk of receiving inaccurate information.

Dementia. Chatbot. Conversational design. Human-centred design. Generative AI. Caregiver support. Caregiver intervention.

1. INTRODUCTION

Dementia is a syndrome where there is deterioration of cognitive functions beyond what might be expected from the usual consequences of biological ageing (World Health Organization, 2023). Peru, a low-middle-income country (LMIC), has a 6.85% prevalence of dementia in adults aged over 65 years old (Custodio, et al., 2008). This requires a significant number of caregivers to support them in managing the cognitive and behavioural symptoms they experience (Custodio & Montesinos, 2022). Being a caregiver is a mentally and physically straining role that demands a lot of time from caregivers and may result in caregiver burnout, depression, stress, and economic burden (Custodio, et al., 2014) (Parodi, et al., 2011). In Peru, 80% of primary caregivers are women, and most are relatives with no formal training, which makes caring for their loved one even more challenging (Custodio, 2016).

As of the current literature review, (Guerra, et al., 2010) remain the sole study addressing interventions for caregivers in Peru. Here, the authors evaluated an adaptation of the intervention: "Helping carers to help", which focused on the education and training of caregivers. The intervention was found to decrease strain after six months, but that did not impact caregivers' psychological distress or the quality of life of either caregivers or patients.

However, research indicates that psychosocial interventions can be effective in improving the quality of life of caregivers and People with Dementia (PWD) in high-income countries (HIC) (Aravena C, et al., 2016). For example, the Care Ecosystem (CE) is a US-based program for collaborative dementia care delivered over the telephone and Internet for caregivers and PWD (Possin, et al., 2019). An important aspect of this intervention being the care team navigators (CTNs), who are unlicensed yet trained individuals acting as a point of contact between caregivers and a clinical team offering educational resources and care protocols (Bernstein, et al., 2019). A key part of their role is to provide care navigation and emotional support, which is recognised as an effective way to support caregivers (Bernstein, et al., 2019). While this insight could benefit caregivers and PWD in Peru, its application would require a cultural adaptation strategy due to population differences.

In the present study, we hypothesize that conversational agents, such as a chatbot, can provide care navigation and emotional support to caregivers within a resource-restricted context like Peru. A chatbot is defined as a dialogue system that conducts natural language processing (NLP) and provides human-like responses to user input (Ruggiano, et al., 2021). Chatbots provide considerable opportunities for dementia care since they are a cost-effective method to increase healthcare access (Ruggiano, et al., 2018). Moreover, modern chatbots using generative artificial intelligence (GAI), such as ChatGPT, have demonstrated impressive ability in understanding and replicating human-like texts (Ayers, et al., 2023). For instance, a recent study compared ChatGPT's responses with physician responses to real-world patient questions and found that the chatbot responses were rated significantly higher for quality and empathy by a panel of licensed physicians (Ayers, et al., 2023). But GAI also can be unpredictable and unsafe, particularly in health-related applications. It is, therefore, important to understand the risks of its availability in the context of dementia in an LMIC.

This study engaged stakeholders to learn about their interests and concerns and explore the feasibility of chatbots to support dementia caregivers in Peru. Our research consists of three phases: stakeholder engagement, the design and development of two versions of the chatbot, and the evaluation to compare stakeholder perceptions on both versions.

2. BACKGROUND

2.1 Project-at-Large

This study is part of the "IMPACT" project aimed at strengthening the health service provision in Peru for PWD and their carers through sustainable, centrally-supported, community-delivered, technology-enabled innovation. IMPACT is a 4-year project conducted by researchers in Peru and the UK, involving participants from four Peruvian cities: Lima, Tumbes, Huancayo, and Iquitos. This study uses design research methods as part of Patient and Public Involvement activities to interest and risks of such interventions. Our research objectives involve adapting and applying aspects of the CE, hence the focus of this study on the CE and its materials. For the project-at-large the intention of this study is to investigate the utility of a chatbot for a randomized controlled trial (RCT) as part of IMPACT. The evaluation of the chatbot and its implications are thus highly significant.

2.2 Literature Review

2.2.1. Technologies for caregivers and PWD

There have been increasing efforts to develop technology-based interventions for dementia caregivers (Ruggiano, et al., 2018) which include but are not limited to, ICT devices, assistive technology, mobile health applications and conversational agents (Topo, 2008) (Kruse, et al., 2020). Despite such development, few applications are supported by evidence-based results, and even fewer are developed in resource-restricted contexts (Ruggiano, et al., 2018).

A review of 46 studies (1992-2007) that focused on technologies supporting PWD and their caregivers found 15 studies aimed to provide technology-mediated support services for family carers. They concluded that the computer and telephone-mediated services were widely accepted by the family carers (Topo, 2008). Most of the studies reported a reduction in caregiver burden and stress which showcases the opportunity of a technology-based intervention to support caregivers.

Additionally, a systematic review of assistive technologies' (AT) facilitators, barriers, and medical outcomes found that the main facilitator was caregivers' desire for AT, as it supports independence. Conversely, the main barriers were the PWD rejection of AT, partly due to cost (Kruse, et al., 2020). This suggests that caregivers are willing to adopt technology-based interventions.

2.2.2. Conversational agents in healthcare

Conversational agents are being increasingly utilised in healthcare applications (Ruggiano, et al., 2021), a consequence of modern technologies' ability to simulate complex human-like responses (Britz, et al., 2017).

A systematic review of fourteen conversational agents in healthcare concluded that only one study evaluated the efficacy or safety of the conversational agent (Laranjo, et al., 2018). However, this study reported a significant effect in reducing users' depression symptoms.

2.2.3. Conversational agents for caregivers

Using conversational agents to support PWD and their caregivers is an emerging field of research. Nevertheless, recent papers detail the use of chatbots to support PWD, including in the detection of cognitive impairment (de Arriba-Pérez, et al., 2022). Other studies provide design recommendations for conversational agents for PWD (Seah, et al., 2022). Such research is summarised in systematic reviews (Ruggiano, et al., 2021) and thematic analyses of research on chatbots for PWD (Rampioni, et al., 2021).

Moreover, a recent study discussed the unique challenges and opportunities of conversational agents to support dementia caregivers (Jiménez, et al., 2022). The authors conducted a qualitative study with 7 participants who identified forgetting things and repetitive questioning as the most challenging behaviours of the PWD, factors that have not yet been tackled with a conversational agent. These behaviours were found to cause feelings of frustration and stress for caregivers. The final recommendations included allowing personalisation, using it to provide activities for the PWD, and adapting to the situation and needs of the caregiver.

3. STAKEHOLDER ENGAGEMENT

The first phase of the study involved semi-structured interviews with dementia caregivers and experts in dementia. The purpose was to inform the design of our chatbot (Jiménez, et al., 2022). Our findings were transformed into chatbot design, system design, and conversational design requirements.

3.1 Method

3.1.1. Semi-Structured Interviews
We first created a discussion guide based on literature insights and the World Health Organization handbook for implementing mDementia, using mobile technology to support carers of PWD (World Health Organization, 2021). The caregiver interviews explored participants' challenges, needs, and wants, and their perception, requirements, and concerns regarding a chatbot. In the expert interviews, the topics explored were dementia care interventions, communication methods for caregivers, and participants' recommendations and concerns about a chatbot for caregivers.

Before the interview, we gave participants an information sheet and asked for their informed consent to participate. Interviews were scheduled once eligibility and consent requirements had been satisfied. All interviews took place one-on-one, with the first author as the interviewer. Interviews with caregivers were face-to-face (14%) or via WhatsApp video calls (86%). These were recorded and automatically transcribed. Conversely, every expert interview took place via Microsoft Teams video call and was recorded and transcribed within Teams.

3.1.2. Participants
To be eligible for the study, participants had to be the primary caregiver for a PWD and reside in Peru. We recruited *N*=7 caregivers to participate in the interviews. Most participants (86%) were female, and all were the primary caregiver for a family member. We initially sought to interview a larger number of participants. However, due to the time-consuming nature of care work, many potential respondents could not participate. Nevertheless, the gender distribution of our sample reflects the gender disparity of caregivers in Peru and is, therefore, a representative sample of our intended demographic (Custodio, 2016).

We also recruited four experts in dementia to participate. These experts included a Peruvian neurologist, a Peruvian public health administrator, an English academic neurologist, and an American psychologist. Both Peruvian experts work in the Peruvian Institute of Neuroscience (IPN) and have developed an education intervention for caregivers in Lima. The English neurologist is involved in the IMPACT project, and the American psychologist was one of the lead researchers of the CE intervention.

Both sets of participants were recruited based on a combination of personal and professional contacts.

3.1.4 Thematic Analysis
We used thematic analysis on the interview transcripts to identify insights relevant to chatbot design. Thematic analysis is a qualitative framework for describing unstructured data in rich detail. The framework we used followed the steps outlined in (Braun & Clarke, 2006). This process involved familiarization with the transcripts by gathering initial observations. We then coded interesting sections of the transcript and gathered them into themes and sub-themes. To improve methodological rigour, the analysis output was reviewed by Peruvian dementia experts experienced in qualitative research.

3.2 Results and Discussion

3.2.1. Caregivers Findings
The findings from the caregivers' interviews are split into the current situation, the problem, and a chatbot as a solution.

Regarding their current situation, all caregivers felt they did not have "enough support" and they lacked knowledge and resources: "I don't have any knowledge in caring for older adults, what I put in practice only comes from the love for my dad". Thus, they all actively look for information to become more knowledgeable about dementia showing there is an opportunity to provide this.

Caregivers identified several key problems: the emotional and physical burden of caregiving, the difficulty of managing behavioural symptoms and the uncertain nature of dementia. Most of their needs and wants involved having the appropriate tools and information to give the PWD and themselves a "happy" life. All the caregivers want immediate advice on managing behavioural symptoms because these are the most distressing. The caregivers mentioned they want support groups and health services orientation. This indicates they value efficiency, accessibility, and the well-being of the PWD and their own. However, only one caregiver engaged in self-care activities, while the others admitted neglecting their own well-being.

The caregivers had mixed perceptions about a chatbot. These negative perceptions stemmed from distrust in the Peruvian health system and unpleasant experiences with other chatbots. Their positive perceptions stemmed from the idea that a chatbot would always be available and helpful in moments of crisis: "It would be very useful - there are some situations I don't know how to handle - I do what I can but I know that my abilities are limited - so this could help in those moments".

When prompted about chatbot requirements, all caregivers said it was essential for the chatbot to be built on "trustworthy sources" and to "respect their privacy". They highlighted the chatbot must have "empathy and patience" and should demonstrate this in the conversation. Moreover, caregivers emphasized the chatbot must give "immediate" and "useful" information to "calm their doubts". They stressed this information must be suitable for the Peruvian context. Their requirements show they value trust and privacy and reinforce they value efficiency and accessibility.

Furthermore, caregivers recommended that the chatbot uses "simple terminology" and a friendly tone for accessibility and a pleasant experience. They also want the chatbot to offer the ability to vent, contact a human operator, and provide referrals to specialists or health services, suggesting the chatbot should be part of a wider system. Similarly, some caregivers mentioned that a clinical team could provide additional support if the user's needs exceeded the chatbot's capabilities.

The caregivers also discussed their concerns about using a chatbot. They feared that it would not be useful, trustworthy, or accessible. They were concerned about being misunderstood by the chatbot and about it being "unreliable" in moments of crisis. Additionally, they said a chatbot would be unable to help with serious mental health issues: "the emotional burden can be very heavy, and I feel that to deal with that you do need a human".

3.2.2. Dementia Experts Findings

The findings from the interviews with dementia experts concern their previous work and their recommendations and concerns about a chatbot.

Regarding their experience, Peruvian experts highlighted that even though there are high-quality materials for caregivers, these have been mostly developed for HIC and as such "cannot be applied to our (Peruvian) reality". Thus, they have produced a manual for caregivers to understand dementia and its symptoms in Peru (Custodio & Montesinos, 2022). Additionally, they have a WhatsApp group for caregivers to "spread awareness of information sessions with doctors and materials for relaxation to prevent caregiver burnout". The group has over 250 members, evidencing caregivers' desire to access relevant information. It was pointed out that only the group administrators can send messages because if not, they receive over 100 messages per hour from caregivers asking questions. This finding demonstrates the potential utility of a question-answering feature in a caregiver-focused chatbot.

Features of a chatbot recommended by dementia experts included a friendly tone with "emojis", simple vocabulary and demonstrable empathy towards the users. These coincide with caregiver recommendations and are taken forward in the next phase of the study. The experts recommended the chatbot provides information on behavioural symptoms of dementia, as this is the most sought-after issue for carers. Like caregivers, they stressed that this information must be suitable for Peruvians.

The dementia experts discussed that a chatbot is limited in its capacity to treat mental health issues, and the technical limitations would make it unable to deal with uncommon issues arising from rare symptoms. An expert with previous experience of chatbots for caregivers in the pharmaceutical industry suggested that to avoid open-ended user inputs, which could be challenging for the chatbot to understand and frustrating for the user, having menus of options for user input is the best alternative. Furthermore, some experts recommended that due to these limitations, the chatbot should have a clinical team behind it, or a human operator, to support caregivers when the chatbot is unable to. They stressed that it is critical to determine when the chatbot has reached its limit and when it is time for a human, like a clinical specialist, to get involved.

The design requirements are detailed below.

3.2.3. Chatbot design requirements

(i) Provide information about users' privacy rights and request consent when the conversation is recorded, or other information is collected.
(ii) Enable users to feel empathy, patience, and support from the chatbot.
(iii) Provide users with relevant information on care management suited for their context through a question-answering (QA) feature.
(iv) Develop the chatbot's responses based on trustworthy sources verified by healthcare professionals and display these sources to the users.
(v) Provide users with clear instructions and information about the chatbot.
(vi) Provide orientation in health services for caregivers and PWD.

3.2.4. System design requirements

(vii) Provide alternative support from a human operator to manage topics and issues beyond the chatbot's scope.
(viii) Implement referrals to specialists (e.g., psychologists) and health service providers whenever sensitive topics or issues are mentioned beyond the chatbot's scope.
(ix) Develop the chatbot in native languages (Quechua, Aymara)

3.2.5. Conversational design requirements

(x) Provide a menu of options for users to select features of the chatbot.
(xi) Personalise responses by integrating personal information.
(xii) Incorporate emojis in the chatbot's responses to create a friendly tone.
(xiii) Employ simple vocabulary avoiding medical jargon.

4. DESIGN AND DEVELOPMENT

The second phase of the study involved conversational design and implementation of the chatbot based on the stakeholder's requirements. Two chatbots were created: a non-GAI chatbot (V1) and a GAI chatbot (V2) which uses generative machine learning (ML) models for its responses. It is anticipated that a generative AI approach (V2) could be more engaging but less accurate.

Conversational design is defined as "the practice of making AI assistants more helpful and natural when they talk to humans – combining technology, psychology, and language to create human-centric experiences" (Conversation Design Institute, 2020). The development of both chatbots focused on the conversational design experience rather than constructing a larger system. Hence, the chatbots were developed in Voiceflow, a tool to design and test conversational assistants (Voiceflow, 2023). Voiceflow's flexibility facilitated feedback-based iterations involving a caregiver and a dementia expert who contributed to design decisions, like the chatbot's wording. Additionally, evidence-based communication methods, such as BATHE, were applied in the conversational design to demonstrate empathy, patience, and support (Lieberman & Stuart, 1999) (Jr., 2018).

The chatbot was named Ana, a common women's name in Peru. This gender affordance was based on research showing it can help caregivers, who are mostly women, feel more comfortable when speaking about dementia and well-being (Brahnam & De Angeli, 2012).

Table 1: *Comparison between both versions of Ana*

	Ana V1	**Ana V2**
Features	*Ask anything (QA), Navigate health services, Alert an emergency*	*Vent, Ask anything (QA), Navigate health services, Alert an emergency*
Uses GAI	*No*	*Yes*
QA Feature NLP Tasks	*Semantic similarity, passage ranking*	*Semantic similarity, passage ranking, text generation*
ML Models	*Sentence Transformers: Paraphrase-xlm-r-Multilingual-V1 - OpenAI: Ada*	*Sentence Transformers: Paraphrase-xlm-r-Multilingual-V1 - OpenAI: Ada and GPT-3.5*
Verified knowledge base	*Yes*	*No*
Verified responses	*Yes*	*No*

The conversation with Ana always begins with a greeting. Then, if the user is new, it follows with an introduction from Ana as a non-human assistant, a disclaimer about the user's privacy rights, and a question about whether the user wants to know more about the chatbot, followed with a 'Yes' or 'No' option. This gives users a chance to clarify any doubts they have about the chatbot. If the answer is 'Yes', then the user is prompted to ask questions about Ana, like where the content comes from. If the answer is 'No', or once the user does not have any more questions about Ana, then Ana asks for their name. Ana reaffirms their name by saying "Nice to meet you {name}" and then presents the main menu of features as buttons for users to pick from, each with a corresponding emoji. The welcome process described is identical for both versions of Ana.

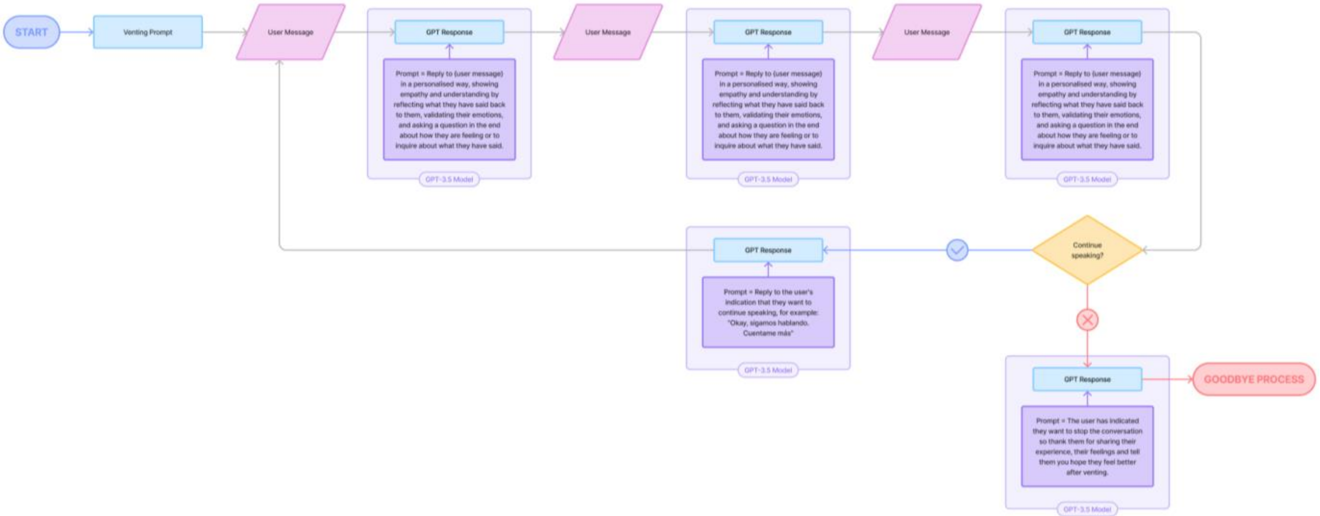
Ana-V1 has a menu with three features: "Ask anything", "Navigate health services" and "Alert an emergency", while Ana-V2 has the additional feature of "Vent". Given time constraints, we only developed the "ask anything" and "vent" features. The other two features require comprehensive research into the health service entities' operations, and they were not essential to evaluate user experience.

4.1.1. The "Ask anything" feature (QA)

This is the main feature of both chatbots and aims to tackle caregivers' need for immediate information about dementia and care management. With this feature, the user can ask a question either about dementia or how to manage a symptom and Ana answers using the software architecture.

Subsequently, Ana asks if the response has answered the user's question. If the user indicates it hasn't then Ana replies, "Sorry for the confusion" and prompts the user to repeat the question, but if the user indicates it has, then Ana asks whether they want to ask another question. If the user wants to ask another question, this process is repeated until they do not. If the user has no more questions, then Ana follows the goodbye process.

Even though both versions follow the same process for this feature, the different conversational design and system architectures governing the chatbot's responses means that each chatbot's reply differs and therefore create different experiences.



Prompt = Reply to {user message} in a personalised way, showing empathy and understanding by reflecting what they have said back to them, validating their emotions, and asking a question in the end about how they are feeling or to inquire about what they have said.

Prompt = Reply to {user message} in a personalised way, showing empathy and understanding by reflecting what they have said back to them, validating their emotions, and asking a question in the end about how they are feeling or to inquire about what they have said.

Prompt = Reply to {user message} in a personalised way, showing empathy and understanding by reflecting what they have said back to them, validating their emotions, and asking a question in the end about how they are feeling or to inquire about what they have said.

Prompt = Reply to the user's indication that they want to continue speaking, for example: "Okay, sigamos hablando. Cuentame más"

Prompt = The user has indicated they want to stop the conversation so thank them for sharing their experience, their feelings and tell them you hope they feel better after venting.

Figure 1: *Ana V2 – "Vent" feature back-end*

For Ana-V1, the back-end process conducts two sequential NLP tasks: semantic similarity and passage ranking. The semantic similarity task finds the most similar verified question in the Ana database to the question asked by the user, using the model Paraphrase-xlm-r-Multilingual-V1 from Sentence Transformers. This results in a similarity score between the user's question and the most similar question from the database. If the similarity score is equal to or greater than 0.9, then the corresponding verified answer from the database is given to the user. On the other hand, if the similarity score is below 0.9, then a passage ranking task is executed to find the most relevant document from the database for the user's question. This task is completed by the embedding model Ada from OpenAI and results in a relevance score of the most relevant document for the user's question. If the relevance score of the document is equal to or greater than 0.8, then Ana's response is "I can't answer your question, but I think this document could help you: {link}", where {link} is the URL of the most relevant document from the database. However, if the relevance score is below 0.8, Ana's response is "Sorry, I can't answer your question". The back-end architecture described is the safest and most trustworthy method to accomplish the QA feature, since only verified answers from the database can be given to the user and if this isn't possible, they are directed to a verified document.

For Ana-V2, the back-end process conducts three sequential NLP tasks: semantic similarity, passage ranking, and text generation. Ana-V2 uses the same models and parameters for the first two tasks as Ana-V1 and follows the same process until the second task, passage ranking. Once the passage ranking is executed and the most relevant document for the user's question has been found, if the relevance score calculated is equal to or greater than 0.8, the text generation task is conducted. This third task sends the following prompt to ChatGPT-3.5: "Use the information below about dementia and caregiving to answer the subsequent question in Spanish: {most relevant document}, {user's question}". The prompt includes the most relevant document and the user's question. Alternatively, if the relevance score is below 0.8, the answer given is "Sorry, I can't answer your question". In this architecture, if there is not a question similar enough to the user's question in the Ana database, then the response is generated instantly by ChatGPT. Therefore, the answers given to the users by Ana-V2 are not verified. This architecture is potentially unsafe and unreliable for a healthcare context since there is no reliable way to avoid ChatGPT giving the user an incorrect or undesirable answer.

4.1.2. The "vent" feature (Ana-V2)

The idea of the "vent" feature is to create a safe space for caregivers to vent about anything that is an emotional burden to them. This feature is only available in Ana-V2 because its functionality depends on generative AI (illustrated in Figure 1). The responses are generated with the GPT-3.5 model because these are the models being tested in similar literature (Ayers, et al., 2023).

When this feature is selected, Ana says she is "ready to listen" and asks the user to reply in one message. Then, the response is generated with the prompt: "Reply to user's message in a personalised way, showing empathy and understanding by reflecting what they have said back to them, validating their emotions, and asking a question in the end about how they are feeling or to inquire about what they have said." The response reflects the user's message to show empathy, as suggested by the BATHE communication method (Lieberman & Stuart, 1999), and includes a question about their feelings, as suggested by dementia experts. This repeats three times and then, Ana asks whether they want to continue or end the conversation, to avoid a never-ending conversation. If the user wants to continue, then this process is repeated. If the user wants to stop the conversation, then Ana replies with a message generated with the prompt: "The user has indicated they want to stop the conversation so thank them for sharing their experience, feelings, and tell them you hope they feel better after venting". This is followed by the goodbye process.

5. EVALUATION

5.1 Method

5.1.1. Semi-Structured Interviews

To evaluate the chatbots, semi-structured interviews were conducted with caregivers and experts. The same eligibility criteria as for the Stakeholder Engagement were applied for caregivers because the participants agreed to be re-contacted to test the chatbot. For experts to be eligible, they had to be professionals working on dementia interventions or chatbots for healthcare. Participants were given an information sheet and a consent form. Once their eligibility and consent were confirmed, the interview was scheduled. Interviews with caregivers consisted of three parts.

First, caregivers engaged in conversations with Ana. They were instructed to use the QA feature in Ana-V1 first to ask questions about behavioural symptoms. Then, they repeated the same questions to Ana-V2. Finally, they used the "vent" feature to speak about anything they wanted, and if requested, they were given examples. Second, caregivers completed a caregiver satisfaction questionnaire based on eight success metrics: empathy, understanding, knowledge, trustworthiness, usefulness, reliability, giving carers confidence, and supporting carers. For each question, caregivers rated both chatbots on a scale of 1 to 5 (1 = "no, never", 2 = "no, rarely", 3 = "maybe sometimes", 4 = "yes, quite frequently", 5 = "yes, nearly always"). Four of the questions were based on the CE's caregiver satisfaction questionnaire for CTNs because they reflected success metrics for Ana (Bernstein, et al., 2019). Likewise, one question comes from the Health Care Climate Questionnaire because it reflected one of Ana's success metrics (Center for Self-determination Theory, 2023). All other questions were based on the insights from Stakeholder Engagement. Third, caregivers answered three open-ended questions: "What worked?", "What did not work and could be better?", and "Which chatbot do you prefer and why?".

The experts' interviews followed a similar structure. First, they interacted with the chatbots following the same instructions as caregivers. Second, they answered open-ended questions. These questions were the same questions asked to caregivers with two additional questions about the possible intended and unintended consequences of Ana.

Participants interacted with Ana via a web browser with a URL generated by Voiceflow. All interviews were held on Zoom, recorded, transcribed, and translated. The quantitative results were aggregated to calculate average scores and the qualitative data was analysed thematically.

5.1.2 Participants

The seven caregivers from the Stakeholder Engagement were scheduled for testing, but only five caregivers participated due to medical emergencies. All the participants were women (100%) which is not an accurate representation of the caregiver population limiting this study. Two Peruvian public health administrators participated as experts. One works for the IPN and has experience with caregiver interventions. Both have experience with chatbots for caregivers in the pharmaceutical industry in Peru.

5.3 Results and Discussion

5.3.1. Questionnaire Findings

Table 2: *Questionnaire Results*

Question	**Ana V1 Av. Score**	**Ana V2 Av. Score**	**Human CTN Av. Score**
Did you feel understood by the chatbot?	*3.0*	*4.4*	*4.84*
Did you feel that the chatbot was empathetic?	*3.0*	*4.6*	*-*
Do you feel that you can trust the chatbot?	*4.2*	*4.8*	*4.91*
Do you feel that the chatbot gives you confidence in your ability to be a caregiver?	*2.8*	*4.6*	*-*
Do you feel the chatbot is knowledgeable?	*2.6*	*4.2*	*4.58*
Do you feel that you can depend on the chatbot?	*2.6*	*4.0*	*-*
Do you feel that the chatbot is useful for you?	*3.4*	*4.6*	*-*
Do you feel supported by the chatbot?	*3.2*	*4.0*	*4.71*

Table 2 shows that Ana-V2 has higher average satisfaction scores for all questions. All caregivers scored Ana-V2 equal or higher than Ana-V1, demonstrating that Ana-V2 is preferred in all aspects. The difference in scores between Ana-V1 and V2 indicates that, for this application, incorporating GAI responses positively impact the user's satisfaction and experience.

Additionally, the results evidence that the current state of Ana-V2 does not reach the same levels of caregiver satisfaction that human CTNs do, exposing the shortcomings of chatbots (Bernstein, et al., 2019). However, the CTN study was conducted with 262 caregivers in an RCT so to make an appropriate comparison, Ana should also be evaluated in an RCT, once issues of safety are addressed, with comparable number of participants.

Moreover, the interview setting facilitated further insights about the scores. The key finding is that caregivers felt Ana-V2 was more understanding and empathic because the generative responses were more human-like. For example, a caregiver expressed that speaking to Ana-V2 was "speaking to someone I knew who understood what I felt". This coincides with the study comparing ChatGPT's responses to doctor's responses, demonstrating the potential of GPT for emotional support and healthcare applications (Ayers, et al., 2023).

5.3.2. Interviews Findings

The interview findings highlight the positive and negative aspects of the chatbots and recommendations for further development.

The most significant finding is that all caregivers preferred Ana-V2 over Ana-V1, even when they knew the risks involved in Ana-V2's responses. This was mainly due to the "vent" feature because caregivers claimed it was "essential" as it focused on them, instead of the PWD: "V2 has a plus because it's also concerned about you". They also thought that Ana-V2 was friendlier and more useful than Ana-V1. Even though the caregivers acknowledged that Ana-V1 would be "more trustworthy in terms of information", they concluded that Ana-V2's benefits outweighed the risk of unreliable information.

The favourite feature of all participants was Ana-V2's "vent" feature. The interviews highlight that the feature is unique: "no other app has this option", that it fulfils a need: "sometimes there is no one I can speak to so I keep everything in… it's nice having this option instead" and that it exceeds exceptions on a chatbot's ability to offer emotional support: "I thought a chatbot would be cold, but it has proved the opposite". Through this feature, Ana-V2 demonstrated empathy and understanding which created a satisfying experience for caregivers, for example they said: "it is strange but also rewarding to engage in a dialogue with someone you can't see but know they understand what you think or feel". The dementia experts recommended that the space created in the "vent" feature could be leveraged to promote caregiver well-being by offering self-care resources during the conversation.

Concerning the welcome process, participants' favourite aspects were having the opportunity to make questions about Ana and being asked for their name. Knowing the chatbot's sources was a key factor contributing to the caregivers' trust in Ana, so some "wished" this was presented at the start of the conversation: "having this at the start makes you trust that the responses have grounds".

The positive comments from participants about the QA feature apply to both chatbots. All participants valued that the answers included the sources of information because it increased their trust in those answers. Caregivers thought the answers were useful because it calmed their doubts while being clear and concise: "It doesn't overwhelm me with information, it gives me precise answers".

The most notable negative aspect of both chatbots was that their scope of knowledge and capabilities were unclear, confusing, and even frustrating for some caregivers. This was partly due to the lack of information provided to users about its functionality. For example, the QA feature focused on symptoms management, but Ana did not communicate this to the users. All caregivers expressed that "knowing upfront what type of questions I can do" would avoid wasting their time. A caregiver suggested that to improve this a "leaflet on how to formulate questions" could be provided. The lack of clarity was also due to ineffective wording in Ana's phrases. For example, almost no caregiver understood Ana's first question ("do you want to know more about me?") and thought that they were already inside the QA feature. The dementia expert pointed out that this could be improved by using different phrasing such as: "Do you want my help, or do you want to know more about how I work?". This new phrasing provides two clear distinct options to the user.

The experts also raised concerns about the phrasing of the main menu. They pointed out that "alert an emergency" is unclear in terms of its functionality, as they wondered if this would automatically call an ambulance service. Based on their previous chatbot experience, they stressed that an "emergency" option can have legal repercussions. For instance, if a user only relies on that button to receive urgent medical attention, they might not get the attention they need in time which makes the chatbot legally responsible for any of the negative consequences. They recommend changing the phrasing to "get emergency contacts" to better reflect the functionality and avoid unintended consequences.

6. CONCLUSION

The study reveals that of the two chatbots designed from stakeholder requirements, and insights from the literature, only Ana-V2, the chatbot employing generative AI, succeeded in providing caregivers with care navigation and emotional support at a high satisfactory level. The chatbots were evaluated by stakeholders through interviews and a satisfaction questionnaire based on eight success metrics. Ana-V2 outperformed Ana-V1, the chatbot that does not employ GAI, in every success metric. Ana-V2 scored above 4 for all metrics on a 1-5 scale and approximated the scores of human CTNs. In contrast, Ana-V1 scored over 3.5 for only one metric, making it less successful in supporting caregivers. A remarkable deduction is that despite acknowledging the potential limitations of GAI, caregivers expressed a strong preference for Ana-V2. The human-like interaction and opportunity for emotional venting were deemed more valuable than the assurance of reliable information.

Furthermore, the study results should be considered alongside its limitations. The most significant limitation is that all study participants are from Lima, the capital city. As said by one of them: “Lima is not Peru”, so the study cannot appropriately represent people of other Peruvian regions, and the findings are not translatable to other Peruvian communities. Similarly, the lack of participants from indigenous communities limits the study because their needs and values can be different from the stakeholders living in Lima. To improve this study, more effort should be placed into recruiting stakeholders from different regions and cultural communities to account for the diversity and disparities in Peru (World Bank, 2023).

Another key limitation is that participants tested the chatbots with the researcher being present. Since the participants knew the researcher had developed both chatbots, it could have influenced their judgment. They were also able to ask questions to the researcher and receive further clarifications during their interaction which may have impacted their experience. To address this limitation, participants should use the chatbot at their own disposal without anyone else present, and after that, the interviews and questionnaires should be carried out for evaluation.

6.1 Project-at-Large Implications

The results indicate that a chatbot could become an effective strategy for adapting principles of the CE intervention for caregivers in Lima. Therefore, there is value in investigating the chatbot further within the IMPACT project. Rural community stakeholders should also be involved to address cultural biases. However, the study suggests that as chatbots with GAI, like ChatGTP, become more common, they will be the preferred option. Currently, this would not be a safe intervention given concerns about inaccuracies and biases of large language models. Hence, if the IMPACT project continues the development of a chatbot to adapt the CE, it should investigate how to add safeguards.

6.2 Next Steps

Following the positive response of stakeholders and their encouraging comments about the chatbot, the next steps for this study would be to continue the development of Ana so it can be tested in a full feasibility study and eventually an RCT in the four regions of Peru with IMPACT participants.

The further development of Ana should involve implementing the recommendations from the evaluation interviews and continuing with the iterative process of user engagement and testing. However, to improve the testing process, it should be conducted over a longer period where the chatbot is available to caregivers to accurately test its performance in real-time and real-world scenarios such as moments of crisis. This will require real time monitoring by trained carers.

Since participants preferred Ana-V2, further research should be conducted about the safety measures required to prevent undesirable unintended consequences from using generative AI responses. This must employ Responsible Design methods, like ‘Design for Wellbeing’ workshops, and involve stakeholders such as caregivers, neurologists, psychologists, public health experts, and legal advisors.

ACKNOWLEDGMENTS

The project was partially funded by the natioanl Institute of Health Research, UK: NIHR150287 IMPACT: Innovations using Mhealth for People with dementiA and Co-morbidiTies.

8. REFERENCES

Aravena C, J. et al. (2016) Calidad de vida en cuidadores informales de personas con demencia: una revisión sistemática de intervenciones psicosociales. Revista chilena de neuro-psiquiatría, 54(4), pp. 328-341.

Ayers, J. W. et al. (2023) Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. JAMA Internal Medicine.

Bernstein, A. et al. (2019) The Role of Care Navigators Working with People with Dementia and Their Caregivers. Journal of Alzheimer's Disease, 71(1), pp. 45-55.

Brahnam, S. & De Angeli, A. (2012) Gender affordances of conversational agents. Interacting with Computers, 24(3), pp. 139-153.

Braun, V. & Clarke, V. (2006) Using thematic analysis in psychology. Qualitative Research in Psychology, 3(2), pp. 77-101.

Britz, D., Goldie, A., Luong, M.-T. & Le, Q. (2017) Massive Exploration of Neural Machine Translation Architectures. Cornell University.

Center for Self-determination Theory. (2023) Health-Care Self-Determination Theory Questionnaire. https://selfdeterminationtheory.org/health-care-self-determination-theory-questionnaire/

Conversation Design Institute. (2020) A guide to conversation design: why, what and how. https://www.conversationdesigninstitute.com/communications/what-is-conversation-design

Custodio, N. (2016) Vivir con demencia en Perú: ¿El sistema de salud está enfrentando la sobrecarga?. Revista de Neuro-Psiquiatría, 79(1), pp. 1-2.

Custodio, N. et al. (2008) Prevalencia de demencia en una población urbana de Lima-Perú: estudio puerta a puerta. Anales de la Facultad de Medicina, 69(4).

Custodio, N. et al. (2014) Informal caregiver burden in middle-income countries Results from Memory Centers in Lima - Peru. Dementia & Neuropsychologia, 8(4), pp. 376-383.

Custodio, N. & Montesinos, R. (2022) Reconociendo los sintomas de pacientes con demencia.

de Arriba-Pérez, F., García-Méndez, S., González-Castaño, F. J. & Costa-Montenegro, E. (2022) Automatic detection of cognitive impairment in elderly people using an entertainment chatbot with Natural Language Processing capabilities. Journal of Ambient Intelligence and Humanized Computing.

FastAPI. (2023) https://fastapi.tiangolo.com/lo/.

Guerra, M., Ferri, C. P. & Fonseca, M. (2010) Helping carers to care: the 10/66 dementia research group's randomized control trial of a caregiver intervention in Peru. Revista Brasileira de Psiquiatria, 33(1), pp. 47-54.

Hugging Face.(2023) https://huggingface.co/models

Jiménez, S. et al. (2022) Towards Conversational Agents to support Informal Caregivers of People with Dementia: Challenges and Opportunities. Programming and Computer Software, 48(8), pp. 606-613.

Jr., W. E. C. (2018) Four Evidence-Based Communication Strategies to Enhance Patient Care. Family Practice Management, 25(5), p. 13–17.

Kruse, C. S. et al. (2020) Evaluating the Facilitators, Barriers, and Medical Outcomes Commensurate with the Use of Assistive Technology to Support People with Dementia: A Systematic Review Literature. Healthcare, 8(3), p. 278.

Laranjo, L. et al. (2018) Conversational agents in healthcare: a systematic review. Journal of the American Medical Informatics Association, 25(9), pp. 1248-1258.

Lieberman, J. A. & Stuart, M. R. (1999) The BATHE Method. The Primary Care Companion to The Journal of Clinical Psychiatry, 1(2), pp. 35-38.

Ngrok. (2023) Ngrok. https://ngrok.com/

OpenAI. (2023) Products. https://openai.com/product (2023).

Parker, D., Mills, S. & Abbey, J. (2008) Effectiveness of interventions that assist caregivers to support people with dementia living in the community: a systematic review. International Journal of Evidence-Based Healthcare, 6(2), pp. 137-172.

Parodi, J., Montoya, J. C., Rojas, D. & Morante, R. (2011) Factores de Riesgo Asociados al Estrés Del Cuidador Del Paciente Adulto Mayor. Rev. Asoc. Colomb. Gerontol. Geriatr., 25(2), pp. 1504-1514.

Possin, K. L. et al. (2019) Effect of Collaborative Dementia Care via Telephone and Internet on Quality of Life, Caregiver Well-being, and Health Care Use. JAMA Internal Medicine, 179(12), p. 1658.

Rampioni, M. et al. (2021) Embodied Conversational Agents for Patients With Dementia: Thematic Literature Analysis. JMIR mHealth and uHealth, 9(7), p. e25381.

Ruggiano, N., Brown, E. L., Li, J. & Scaccianoce, M. (2018) Rural Dementia Caregivers and Technology: What Is the Evidence?. Research in Gerontological Nursing, 11(4), pp. 216-224.

Ruggiano, N. et al. (2021) Chatbots to Support People With Dementia and Their Caregivers: Systematic Review of Functions and Quality. Journal of Medical Internet Research, 23(6).

Seah, C. E. L. et al. (2022) Designing Mindfulness Conversational Agents for People With Early-Stage Dementia and Their Caregivers: Thematic Analysis of Expert and User Perspectives. JMIR Aging, 5(4), p. e40360.

Sevey, R. (2017) Data Robot. https://www.datarobot.com/blog/how-much-data-is-needed-to-train-a-good-model/ (2023).

Topo, P. (2008) Technology Studies to Meet the Needs of People With Dementia and Their Caregivers. Journal of Applied Gerontology, 28(1), pp. 5-37.

Voiceflow. (2023) Voiceflow | Design, prototype, & launch voice apps. https://www.voiceflow.com/

World Bank. (2023) Rising Strong: Peru Poverty and Equity Assessment, Washington, D.C.: The World Bank.

World Health Organization, I. T. (2021) Be healthy, be mobile. A handbook on how to implement mDementia.https://www.who.int/publications/i/item/9789240019966

World Health Organization. (2023) World Health Organization.https://www.who.int/news-room/fact-sheets/detail/dementia (2023).