OXFORD

Interacting with Computers, 2025, **37**, 430–443
https://doi.org/10.1093/iwc/iwae057
Advance access publication date 1 April 2025
Article

Designing AI-Powered Chatbots for Dementia Care in Peru: Stakeholder Engagement and Field Observations

Fernanda Espinoza [1,*], Darren Cook[1], Marco Da Re[1], María Sofia Cuba Fuentes[2], Chris R. Butler[3], and Rafael A. Calvo[1]

[1]Dyson School of Design Engineering, Imperial College London, Imperial College Rd, South Kensington, London SW7 2DB, UK
[2]Centre for Research in Primary care, School of Medicine, Universidad Peruana Cayetano Heredia, Av. Honorio Delgado 430, San Martín de Porres 15102, Lima, Peru
[3]Department of Brain Sciences, Imperial College London, Imperial College Rd, South Kensington, London SW7 2DB, UK
Corresponding author: fe319@imperial.ac.uk

Abstract

In Peru, dementia caregivers face burnout, depression, stress and financial strain. Addressing their needs involves tackling the intricacies of resourcing, caregiving and managing emotional burdens. Chatbots could serve as a viable support mechanism for these issues in regions with limited resources. We study the perceptions of dementia caregivers in Peru regarding a chatbot tailored to offer care navigation and emotional support. We divided the study into four phases: the initial stage encompassed engaging stakeholders to understand potential design challenges; the second stage focused on the design of 'Ana', a chatbot for dementia caregivers; the third stage assessed the chatbot through interviews and a caregiver satisfaction survey and the fourth stage gathered contextual insights through a field study for culturally-sensitive recommendations for 'Ana'. The findings reveal that caregivers seek immediate access to information on handling behavioural symptoms and a platform for emotional release. Moreover, 'Ana' was tested in two configurations—one employed predefined conversation patterns, while the other harnessed generative AI for more dynamic responses. 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. Lastly, the field observations challenge the practicality of offering care navigation due to the lack of resources available in the health system and highlight that 'Ana' needs to be accessible for the varying educational and technological literacy levels present in Peru. However, we found that WhatsApp could be leveraged to overcome the accessibility challenges due to its pervasiveness in Peruvian society and that Community Health Workers could play a vital role in dementia care interventions.

RESEARCH HIGHLIGHTS

- A chatbot offers a cost-effective way to provide Peruvian caregivers with reliable dementia care information.
- Caregivers preferred Ana-V2 (GAI-powered) for its empathy, understanding, and emotional support, despite the risks of inaccurate or insensitive information.
- There was a statistically significant preference of Ana-V2 over Ana-V1 in understanding, empathy, and knowledgeability, suggesting GAI enhances user experience.
- Field observations highlight the need for 'Ana' to accommodate varying literacy levels and leverage WhatsApp for accessibility.

Keywords: *dementia; Chatbot; human-centred design; generative AI; user studies; field studies*

1. INTRODUCTION

Dementia is a syndrome characterized by the 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). A significant number of caregivers support patients in managing their cognitive and behavioural symptoms (Custodio and 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 ones even more challenging (Custodio, 2016). The challenges faced by people with dementia (PWD) and their caregivers is exacerbated by the fragmented and segmented health system that constrains the state's capacity to deliver high-quality health care to all—which calls for inclusive multi-cultural technological interventions for vulnerable populations (Carrillo-Larco et al., 2022).

A review of the literature found only one study that addresses interventions for caregivers in Peru (Guerra et al., 2010). Here, the authors evaluated an adaptation of the intervention: 'Helping carers to help', which focused on the education and training of

Received: November 28, 2023. **Revised:** August 25, 2024. **Accepted:** October 3, 2024

caregivers. The intervention decreased strain after six months but 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 PWD in high-income countries (HIC) (Aravena 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 is 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 recognized 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 as 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). However, the stochastic nature of GAI means its responses can be unpredictable and potentially unsafe, particularly in health-related applications where high levels of control and accuracy are paramount. It is, therefore, important to engage stakeholders to understand the risks of its availability in the context of dementia in an LMIC.

Developing healthcare interventions without in-depth contextual considerations also introduces significant risks, particularly when cultural sensitivities are overlooked. Such oversights can lead to profound adverse outcomes in the communities where they are deployed. For example, during the 1990s, Peru implemented a family planning program aimed at reducing the national birth rate (Carranza Ko, 2021). This initiative sparked controversy and backlash due to reports of involuntary sterilizations among indigenous women. The program failed to account for the linguistic barriers, as many women were Quechua speakers, and did not recognize the cultural significance of large families within Andean communities. The government's top-down approach, lacking in cultural and ethical foresight, resulted in widespread human rights violations. This underscores the necessity of tailoring the development of interventions like the chatbot to the specific cultural, ethical and practical realities of the target setting. Therefore, it is crucial to engage stakeholders in the design and place the local context at the forefront. Such activities described here aim to support a preliminary assessment of the chatbot's potential as a responsible and culturally sensitive tool for dementia caregivers in Peru.

The remainder of this paper is structured accordingly: Section 2 introduces the context of our study and the literature on related studies. Section 3 describes early stakeholder engagement (online) which helped identify some of the issues. These led to the design and development of two versions of the chatbot described in Section 4, and a quantitative evaluation of both versions described in Section 5. Peru's health system is complex, so we were aware that community engagement was required to investigate adoption and contextual considerations for the chatbot. The community engagement and its results are described in Section 6. Section 7 concludes and describes future steps for the project.

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, technology-enabled innovation, delivered by the community. IMPACT is a 4-year project funded by UK's NIHR and conducted by researchers in Peru and the UK, involving participants from four Peruvian cities: Lima, Tumbes, Huancayo and Iquitos. The study used design research methods as part of Community Engagement and Involvement (CEI) activities. IMPACT plans to reuse and culturally adapt aspects of the CE. Since the CE requires many trained CTNs not available in Peru, this study investigates the potential of automating some aspects of the intervention with a chatbot.

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). Examples include ICT devices, assistive technology, mobile health applications and conversational agents (Kruse et al., 2020; Topo, 2008). 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. The authors 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 facilitators were caregivers' desire for AT and increased independence for PWD. Conversely, the main barriers were the PWD rejection of AT and the monetary cost of AT (Kruse et al., 2020). This suggests that caregivers are willing to adopt technology-based interventions, but these should be cost-efficient to enable adoption. Thus, investigating low-cost interventions, such as a chatbot, is particularly important for caregivers in low resource settings, such as Peru.

2.2.2. *Conversational agents in healthcare*

Conversational agents are being increasingly utilized 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 summarized 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 seven caregivers who identified forgetfulness and repetitive questioning as PWD's most challenging behaviours, 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 personalization, 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 online semi-structured interviews with dementia caregivers and dementia experts in Peru. The purpose was to inform the design of a chatbot. The 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; International Telecommunication Union, 2021). The caregiver interviews explored participants' challenges, needs and wants, as well as their perception, requirements and concerns regarding a chatbot. The expert interviews explored 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. In contrast, 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 seven caregivers to participate in the interviews. Six participants (86%) were female, and one participant was male (14%), 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 a representative sample of our intended demographic (Custodio, 2016). All participants are residents of the capital city (Lima), a factor often associated with a higher socio-economic status relative to the broader population, particularly in comparison with inhabitants of rural areas. All participants had successfully completed secondary education and their median age was 57 years old.

We also recruited four dementia experts, including 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 project-at-large, 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.

3.1.3. *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 and 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*

We grouped the findings from the caregivers' interviews into three themes: 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 wanted 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, all except one caregiver admitted to neglecting their own well-being due to their caregiving role.

The caregivers had mixed perceptions about a chatbot. These negative perceptions stemmed from distrust in the Peruvian health system, due to the system's inefficiency, unfairness and low-quality 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 and 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 generated from these interviews 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) Support a sense of being heard and understood.
(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*

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

3.2.5. *Conversational design requirements*

(i) Provide a menu of options for users to select features of the chatbot.
(ii) Personalize responses by integrating personal information.
(iii) Incorporate emojis in the chatbot's responses to create a friendly tone.
(iv) Employ simple vocabulary avoiding medical jargon.

4. DESIGN AND DEVELOPMENT

As our project evolved, chatbots that use Generative AI started being developed, and we aimed to understand how it could be used to address the design requirements above. GAI could address some requirements (e.g. promote the sense of empathy and patience) while challenging others (e.g. trustworthiness). The second phase aimed to explore these tensions. 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 was anticipated that a generative AI approach (V2) could be more engaging but less accurate (Table 1).

The chatbots were developed in Voiceflow, a low-code 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 background-affect-trouble handling-empathy (BATHE), were applied in the conversational design to demonstrate empathy, patience and support (Cayley, 2018; Lieberman and Stuart, 1999).

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 and De Angeli, 2012).

TABLE 1. Comparison (e.g. features, ML models) of Ana-V1 (non-GAI chatbot) and Ana-V2 (GAI chatbot).

	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
Knowledge base	Yes	Yes
Verified responses	Yes	No

TABLE 2. Preview of the Ana database showing three examples of the question-answer pairs.

Question	**Answer**	**Source**
What is dementia?	The term dementia is defined as a clinical syndrome caused by a cognitive dysfunction.	IPN manual
What strategies can I use if the person repeats the same questions?	Remember that the person with dementia is not doing this on purpose, use a calm tone of voice, establish a consistent daily routine and give them simple tasks to distract them.	CE materials
How can I calm the PWD when they experience hallucinations?	You can calm the person by telling them not to worry and that you are there to help. Try to avoid debating about what the person sees or hears, because it is real for them. Try distracting them with activities they like or find comforting.	CE materials

The chatbot was designed to make use of a database with verified and trustworthy responses for users, which had to be created and populated. The resources used to populate this database were the manual 'Recognizing symptoms of dementia' by the IPN and the Behaviour Management Protocol and Educational Materials in Spanish from the CE. The CE materials were reviewed and only behavioural symptoms management advice that could also be applied in the Peruvian context were used for the database. To elaborate, any advice that involved helplines, support groups or contact details for caregivers in the US and recommendations that involved medication or connection to the health system were excluded from the Ana database. The final content and answers in the database were verified by the Peruvian dementia experts.

The Ana database (preview in Table 2) is comprised of validated predefined pairs of questions and answers about dementia, its symptoms and care management strategies for a set of behavioural symptoms (anxiety, delusions and hallucinations, sleep disturbances and wandering). It also contains 'documents', which are extracts (around 500 words) from the manual by the IPN and the CE materials from which the questions and answers came from.

It is important to note that the Ana database is not comprehensive since it was only populated with two sources and reviewed by two experts. The purpose of the database was to prototype the software architecture of the chatbot and investigate the differences between a chatbot employing GAI and one that does not. Developing this lo-fi prototype of the database showed us that the process to developing the database that would be required for a chatbot like Ana to be useful is an intensive data collection exercise that amounts to its own investigation. To populate a useful database, more sources should be used, and a validation process needs to be established for the data that will populate the question-answer pairs for advice to caregivers. Furthermore, this validation process should ensure that the data is culturally sensitive, so it is relevant for the context of Peruvian caregivers.

The conversation with Ana always begins with a greeting. Then, if the user is new, it follows with an introduction from Ana as a nonhuman 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. Additionally, the 'vent' feature could not be implemented for V1 because this feature is for caregivers to write 'anything', so GAI is required to reply to those open-ended messages. Without GAI, the chatbot responses would be the same regardless of the caregivers' messages—and in the context of 'emotional venting' receiving the same response repeatedly could have been emotionally frustrating and triggering for caregivers. Therefore, we think it would have not been useful to develop such static prototype of the 'vent' feature for V1 and ask caregivers to interact with it.

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 by answering their questions on these subjects. Ana personalizes the responses by including known data about the user in those responses such as their name and the person they are caring for. For instance, if it is replying to a

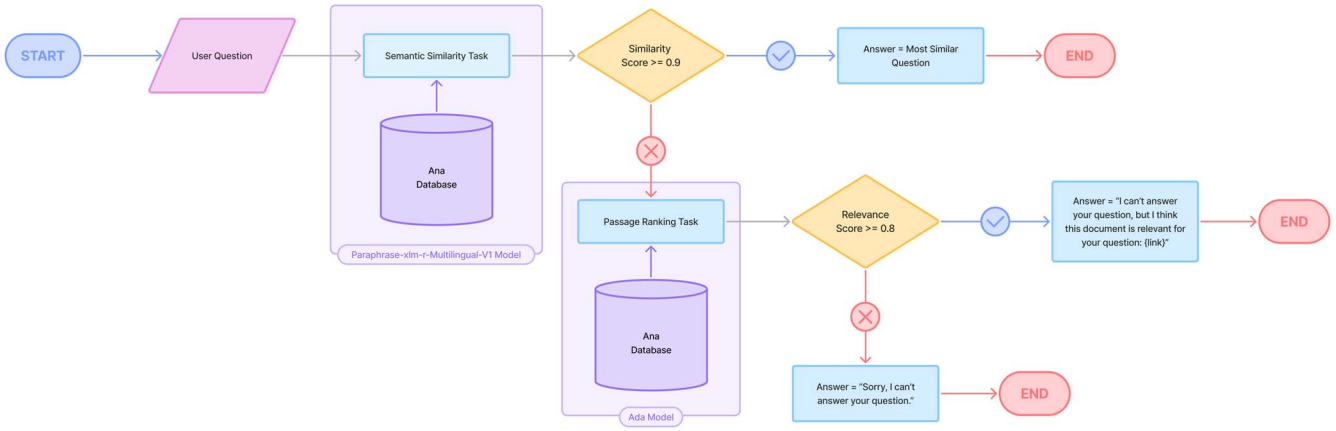

FIGURE 1. Ana V1—'Ask Anything' feature back-end system process.

user who is caring for their mother, the chatbot would phrase the response using phrases like 'your mother' when speaking about the PWD instead of using a general denomination. After responding to the user's question, Ana asks if the response has answered the question. If the user indicates it has not 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.

For Ana-V1, the back-end conducts two sequential NLP tasks: semantic similarity and passage ranking (shown in Fig. 1). The semantic similarity task finds the most similar verified question in the Ana database to the question asked by the user, using the Paraphrase-xlm-r-Multilingual-V1 model from Sentence Transformers (Reimers and Gurevych, 2019). 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 using the Ada model 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 conducts three sequential NLP tasks: semantic similarity, passage ranking, and text generation (shown in Fig. 2). 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 'vent' feature aims 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 GAI (shown in Fig. 3). The responses are generated with the GPT-3.5 model, consistent with related 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 personalized 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 and 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¿. 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'.

5. EVALUATION

5.1. Method

5.1.1. *Semi-structured interviews*

To understand caregivers and experts' perceptions about the chatbots, semi-structured interviews were conducted. The same eligibility criteria as for the Stakeholder Engagement were applied for caregivers because the participants agreed to be re-contacted

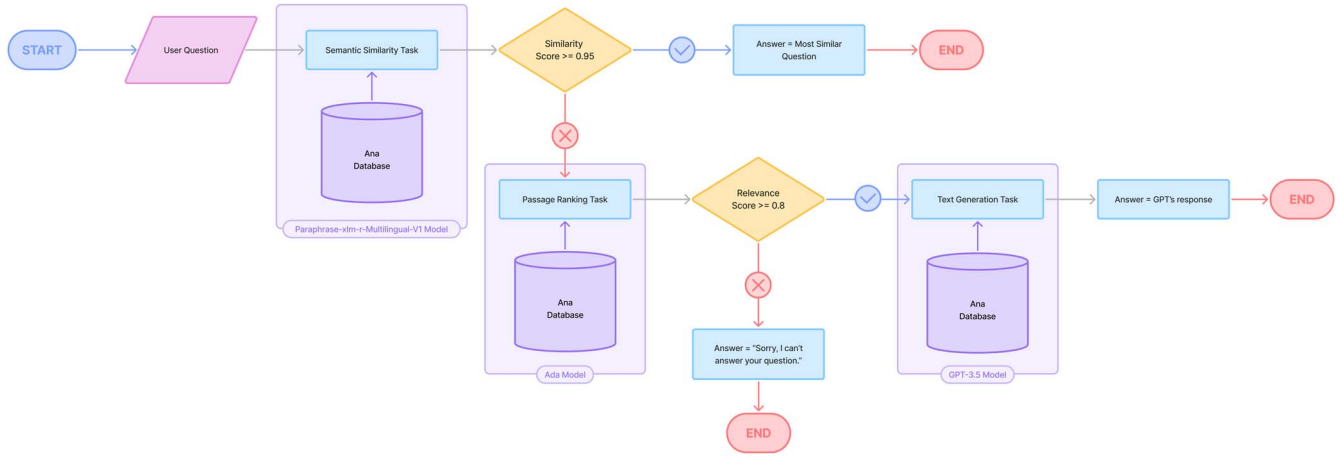

FIGURE 2. Ana V2—'Ask Anything' feature back-end system process.

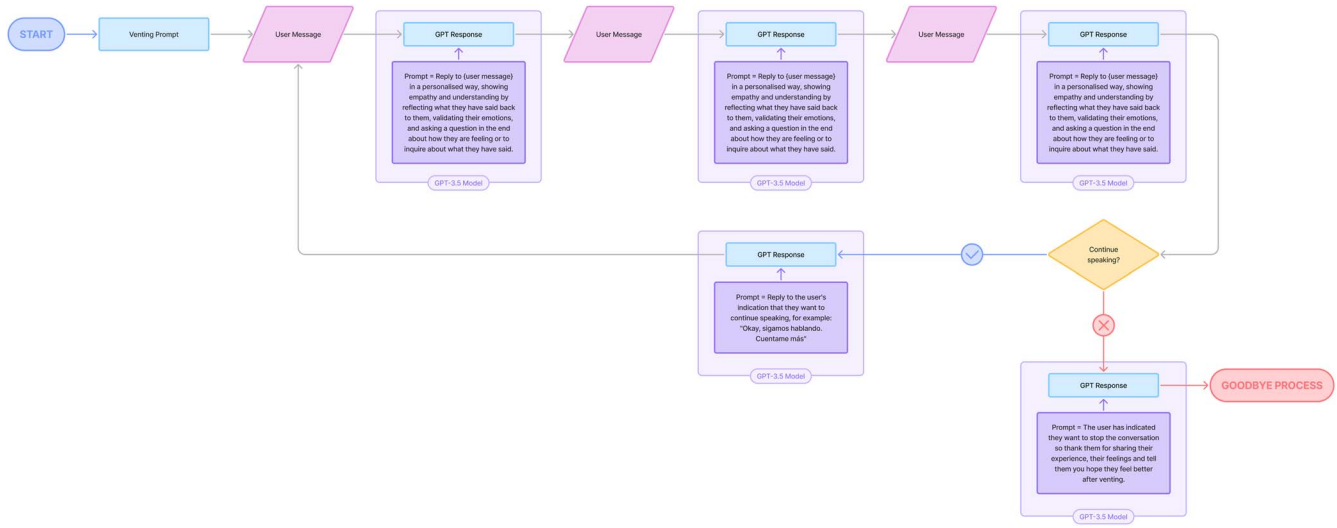

FIGURE 3. Ana V2—'Vent' feature back-end system process using the GPT-3.5 Model.

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 any question about dementia, its symptoms or the following behavioural symptoms: anxiety, delusions, hallucinations, sleep disturbances and wandering. They were not given instructions on how to phrase the questions, instead if they needed or wanted help they were given two scenarios: 'Imagine the person you are caring for is waking up every night to use the bathroom and you do not know how to help them – how would you use the chatbot in this situation?' and 'Imagine the person you are caring for has started to hallucinate and you do not know how to manage this symptom – what would you ask the chatbot?'. Then, they were told to engage with Ana-V2 and repeat the same questions. Finally, they used the 'vent' feature in Ana-V2 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 analyzed thematically.

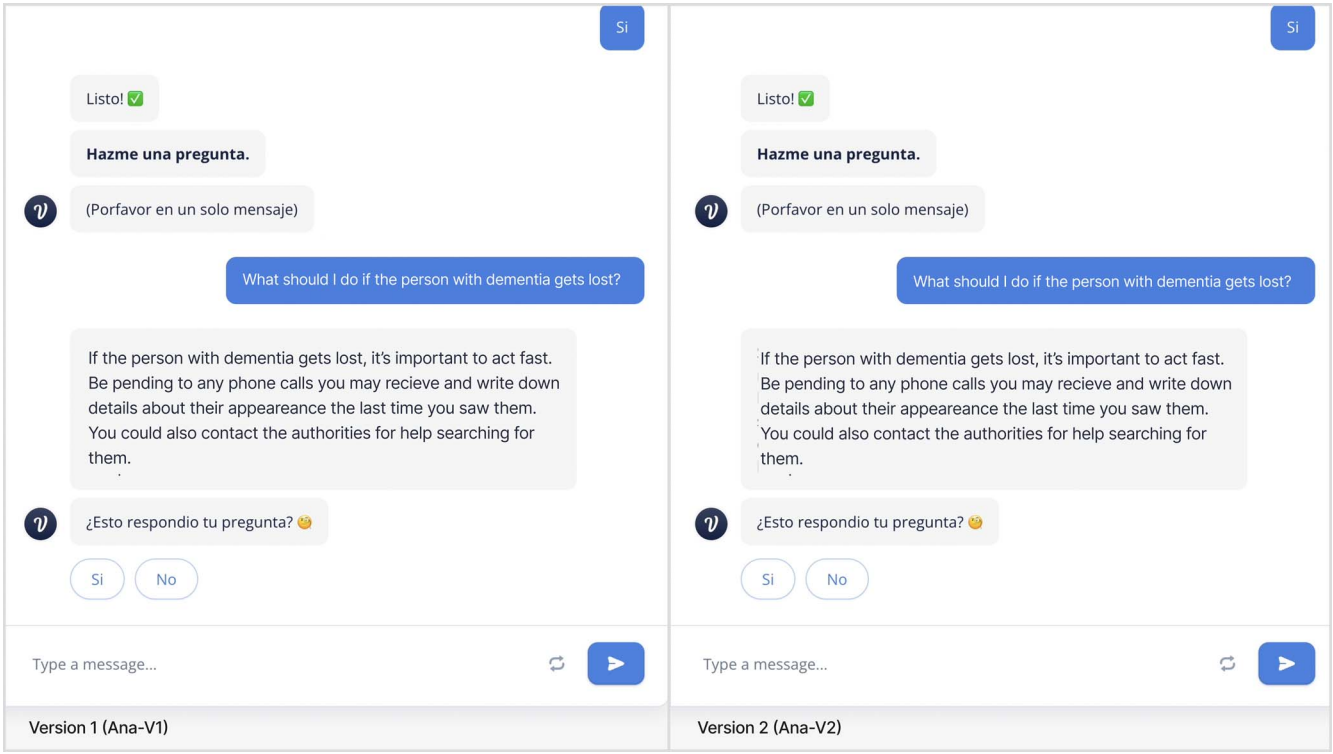

FIGURE 4. Sample interactions of a user asking a question that is in the Ana database to Ana-V1 (left) and Ana-V2 (right) showing the same response by both chatbots. Note: The questions and answers have been translated.

5.1.2. *Participants*

All seven participants were women (100%), which is not an accurate representation of the caregiver population. 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.

5.1.3. *Statistical analysis*

Quantitative data was collected to determine whether there is a statistical difference in the chatbots' quality ratings. Each user rated the chatbots on a scale of 1–5. Due to the small sample size, we could not reliably determine the distribution of the data, a requirement for parametric tests (e.g. *t*-tests require data to be normally distributed). Therefore, we conducted the Wilcoxon signed-rank test. The effect sizes from the test were employed to determine the magnitude of these differences, providing insight into their practical significance. The test was conducted using the NumPy and SciPy libraries in Python (v.3.10.9).

5.2. Results and discussion

5.2.1. *User interaction findings*

This section summarizes the caregivers' interactions with Ana during the evaluation interviews. Most participants ($n = 6$) asked 2–3 questions to Ana-V1, all participants repeated their same questions to Ana-V2, and no participant asked a follow-up question. Figures 4 and 5 compare the responses of Ana-V1 and Ana-V2 to the same questions. Figure 4 shows the interaction when the question was answered with the Ana database and Fig. 5 shows when the question is not in the Ana database and therefore, Ana-V1 answers with a signpost to a relevant and document and Ana-V2 generated the answer with that same document.

5.2.2. *Questionnaire findings*

Table 3 shows that Ana-V2 has higher average satisfaction scores for all questions. All caregivers scored Ana-V2 equal or higher than Ana-V1, suggesting 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 a 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).

5.2.3. *Statistical findings*

The Wilcoxon signed-rank test shown in Table 4 revealed significant differences in caregiver perceptions for understanding, empathy, knowledgeability, dependability and support when interacting with the two versions of Ana, with significant results in over half of the metrics assessed (Questions 1, 2, 5, 6 and 8, $P < 0.05$). The negative rank-biserial correlations in these areas suggest a substantial preference for Ana-V2 over Ana-V1, suggesting that GAI improves the user experience in these areas. Conversely, trust,

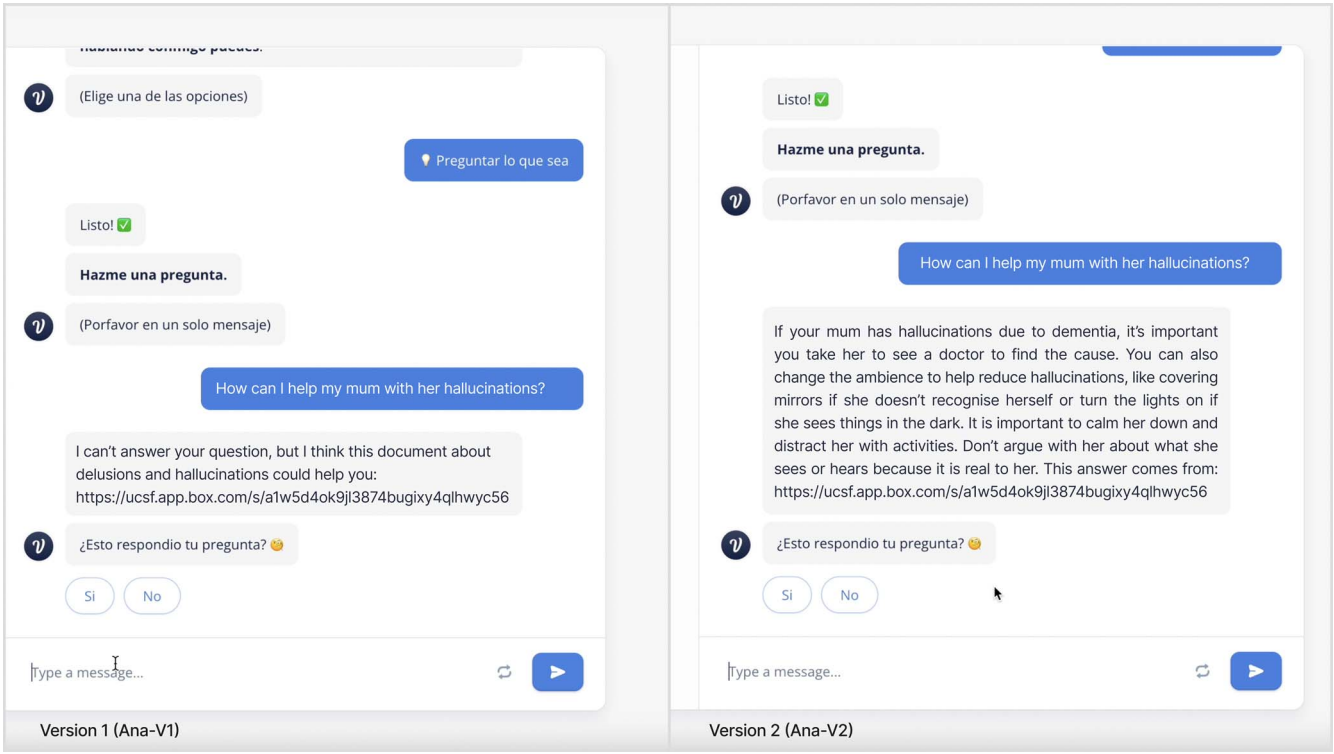

FIGURE 5. Sample interactions of a user asking a question that is not in the Ana database to Ana-V1 (left) and Ana-V2 (right) showing the difference in responses. Note: The questions and answers have been translated.

TABLE 3. Average user satisfaction rating ($n = 7$ users) per chatbot version on a 1–5 scale (1 = 'no, never', 2 = 'no, rarely', 3 = 'maybe sometimes', 4 = 'yes, quite frequently' and 5 = 'yes, nearly always').

Question	Ana V1 Av. score	Ana V2 Av. score	Human CTN Av. score
Did you feel understood by the chatbot?	2.7	4.4	4.84
Did you feel that the chatbot was empathetic?	2.9	4.6	-
Do you feel that you can trust the chatbot?	4.4	4.9	4.91
Do you feel that the chatbot gives you confidence in your ability to be a caregiver?	3.3	4.6	-
Do you feel the chatbot is knowledgeable?	3.0	4.3	4.58
Do you feel that you can depend on the chatbot?	3.0	4.0	-
Do you feel that the chatbot is useful for you?	3.7	4.6	-
Do you feel supported by the chatbot?	3.3	4.3	4.71

Table 4. Results of the Wilcoxon signed-rank test between the median user-satisfaction scores for both chatbots considering *P*-values at or below 0.05 as a statistically significant difference.

Question	Wilcoxon signed-rank test statistic	*P*-value	Rank-biserial correlation
Did you feel understood by the chatbot?	0	0.02*	−1.00
Did you feel that the chatbot was empathetic?	0	0.03*	−0.96
Do you feel that you can trust the chatbot?	0	0.18	−0.46
Do you feel that the chatbot gives you confidence in your ability to be a caregiver?	0	0.07	−0.79
Do you feel the chatbot is knowledgeable?	0	0.02*	−0.96
Do you feel that you can depend on the chatbot?	0	0.04*	−0.89
Do you feel that the chatbot is useful for you?	0	0.11	−0.64
Do you feel supported by the chatbot?	0	0.04*	−0.89

caregiving confidence and perceived usefulness (Questions 3, 4 and 7, $P > 0.05$) did not differ significantly between the chatbots, so there is no evidence that GAI positively or negatively influences the user experience in these aspects.

Overall, the consistent direction (negative value) of rank differences across all metrics highlights a systematic caregiver preference towards Ana-V2, suggesting the need for further investigation into the impact of GAI in conversational agents

in healthcare. However, considering the small sample size, the insights underscore the need for larger-scale studies to validate these preliminary findings.

5.2.4. *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 is 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.

The positive reception of the 'vent' feature in Ana-V2 introduces ethical and safety concerns that need to be addressed as the chatbot is further developed. An ethical dilemma identified during the interviews arises from the use of GAI to enhance Ana-V2's human-like responses, particularly to simulate an emotionally supportive 'friend'. The potential for users to misconstrue interaction with a machine as a human interaction presents unmeasurable risks, so it is essential to strike a balance between emulating human responses and maintaining user awareness that the interaction is with a chatbot, not a person. Hence, optimizing user experience must be balanced with transparency to ensure user safety.

Furthermore, as described in Section 4.1.2, the 'vent' feature operates on a loop of responses from ChatGPT, designed to offer emotional support. The inherent unpredictability of ChatGPT and GAI technologies poses safety risks, particularly if the chatbot generates insensitive or harmful responses during sensitive discussions, such as those concerning mental health. For instance, conversations that delve into topics involving self-harm or suicidal ideation could inadvertently elicit chatbot responses that negatively influence the user's mental well-being. Therefore, it is paramount to engineer robust safeguards into the chatbot's architecture to address and mitigate these GAI-associated risks. These findings underscore the necessity for further exploration into the safe application of GAI, potentially through the lens of Responsible AI research, to harness its advantages while minimizing its potential hazards.

Concerning the welcome process, participants' favourite aspects were having the opportunity to ask 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. COMMUNITY ENGAGEMENT AND ADOPTION

Technologies that have been developed in HIC cannot be introduced in LMIC without studying their impact in the new context. In our case, conversational agents and GAI are new technologies that have been rapidly advancing in HIC but are just beginning to emerge in LMIC. Thus, Peruvian communities are likely to be less familiar with these technologies than UK or US. To neglect a consideration of the cultural context in which Ana would be deployed would not only be impractical and inefficient but also unethical. For this reason, it was important for us to investigate the potential contextual and adoption issues in-person. We were able to do this through a field study for the project-at-large.

For the project-at-large, the first and third author conducted a field study in Peru focused on the co-design of a mobile Health app for dementia diagnosis. This app was designed to be used by Community Health Workers (CHWs) in Peru, so the aim of the field study was to engage the CHWs to learn about their work and their needs and wants regarding the app. Even though the focus of the field study was not a chatbot for dementia caregivers, the activities conducted by the authors enabled them to gather contextual observations and insights to identify challenges and opportunities for a dementia-focused chatbot in Peru.

For our study, the objectives of these field activities were (i) to understand through first-hand experience the context where Ana would potentially be deployed and (ii) to further evaluate whether Ana can be a suitable intervention to support dementia caregivers.

6.1. Method

6.1.1. *Multi-method stakeholder engagement*

The field study took place in four Peruvian cities: Lima (capital), Tumbes (seaside), Huancayo (Andes) and Nauta (Amazon rainforest). The same four-day plan was followed in all the sites to undertake active Community Engagement and Involvement (CEI), which involved the integration of mixed methods from Participatory and Human-Centred Design.

All participants were previously contacted to introduce the project-at-large, the aim of the field study and to explain the four-day plan. To be involved in the study, all participants signed a consent form, which also consented for the audio and video recording of all the activities that took place.

The first day consisted in the field researchers exploring the site to gather observational data and in building a rapport with the local facilitators and CHWs. The second and third day consisted in training workshops for the CHWs about dementia diagnosis conducted and focus groups regarding the CHWs work, their motivation and their perceptions, needs and wants regarding the diagnosis tool and technology. The focus group was led using a discussion guide while allowing the free flow of conversation, which enabled us to uncover important observations for our work with a dementia-focused chatbot. The fourth day consisted in the CHWs using the app to evaluate an older adult in their community and semi-structured interviews with each CHW about their experience.

It is important to reiterate that the discussions we had with stakeholders (e.g. CHWs and older adults) during this field trip were focused on a dementia diagnostic tool, which is unrelated to the chatbot (Ana), so the concept of Ana was never presented or explicitly discussed. Instead, the discussions about the diagnostic tool exposed relevant insights about the context and the local communities for the deployment of a tech-based intervention, such as the chatbot.

The data was collected through observations from researchers in the field and through audio and video recordings of the focus group discussions and semi-structured interviews. The recordings were transcribed and translated.

6.1.2. *Participants*

The field study engaged four groups of participants: CHWs, local facilitators, older adults who underwent evaluation and the relatives of these older adults.

The number of CHWs involved for each site were: 10 in Lima, 7 in Tumbes, 6 in Nauta and 5 in Huancayo. In Lima, Tumbes and Huancayo, they were all women (100%) and in Nauta 2 (33%) were women and 4 were men (67%). Overall, 24 (86%) were women and 4 (14%) were men. It is important to consider the possible gender bias this pool of participants may have.

The number of older adults recruited were: 4 in Lima, 3 in Tumbes, 6 in Nauta and 5 in Huancayo. Overall, 8 were women (44%) and 10 were men (56%). All older adults were over 65 years old. One relative per older adult was recruited that results in the same number of relatives involved in each site: 4 in Lima, 3 in Tumbes, 6 in Nauta and 5 in Huancayo. The relatives were 16 women (89%) and 2 men (11%). All relatives recruited were over 18 years old.

The local facilitators involved were: 1 CHW supervisor in Lima, 1 CHW supervisor in Huancayo, 2 psychologists in Nauta and 1 healthcare professional in Tumbes.

In total, the field study is comprised of 28 CHWs, 5 local facilitators, 18 older adults and 18 relatives.

In each city, participants were recruited from the same district, limiting their ability to represent the entire region due to the social, cultural and economic disparities between different districts within the same city.

6.2. Results and discussion

Even though all conversations were focused on the dementia diagnostic tool, we delved into topics surrounding technology, so we learned about stakeholders' experience, attitudes and perceptions towards digital products, which is relevant for the development of a chatbot for these communities. Similarly, even though the CHWs represent a small part of the population, we were able to learn about the caregivers and stakeholders in the regions who would interact with the chatbot, as well as the context in which the chatbot would be used, through our conversations with them.

6.2.1. *Challenges*

CHWs are volunteers and work with the local public health centres, hospitals or municipalities. Their health expertise depends on the health priorities of their regions, for example, CHWs in Nauta are knowledgeable about Malaria because it is a prevalent disease in Amazonian regions. As a result, the CHWs had little to no knowledge about dementia—most had never heard about dementia before. This shows that the government will face health literacy challenges when deploying dementia care interventions in Peru.

Additionally, CHWs in Huancayo were part of a network, Red Amachay, created to support older adults during the pandemic. A main service of the network was to signpost older adults to health services, so offering them care navigation. However, CHWs highlighted that due to the lack of resources in the health system and the burden experienced by healthcare professionals, the older adults almost never reached the service they needed. The lack of services available for older adults and for mental health was a recurrent observation from CHWs in all sites. This presents a critical challenge for a chatbot offering care navigation, because how could it offer care navigation if services are unable to attend the patients referred or if these services do not exist? Answering this question is key for the future development of Ana and means that the feature of care navigation must be thoroughly evaluated.

Another critical observation is that, on average, CHWs in Lima had higher levels of education than CHWs in the other three regions and this was reflected in a difference in literacy. Compared to CHWs in Lima, CHWs in Nauta, Huancayo and Tumbes took longer to understand concepts of dementia diagnosis and found reading out loud challenging (seen through stuttering, mispronunciation and misreading). This not only showcases the disparities between the capital city and the rest of the country, but also implies that there is a vast difference in literacy within Peruvian communities. Therefore, to ensure the chatbot is accessible it is crucial for Ana to be designed for people with varying literacy levels, suggesting further conversational design investigation.

Similarly, the field study revealed that there is also a difference in technological literacy between CHWs from different regions. For example, CHWs from Lima and Huancayo felt more comfortable using a tablet than CHWs from Nauta and Tumbes, who

expressed feelings of 'fear' when presented with a tablet. When the idea of a chatbot was introduced, CHWs in Lima were familiar with the concept and even had previous experience with some commercial chatbots, whereas CHWs in the other regions were not. This highlights the varying technological literacies we must consider when designing Ana as this will affect user's perception and experience, especially if the notion of a chatbot is new to them. Hence, future research should explore how the chatbot should be developed responsibly for communities where this is new technology to ensure we are not disrupting these communities and their cultural values.

Furthermore, we observed technical challenges for technological interventions in Peru. In most regions besides Lima, the climate and the existing infrastructure make cell phone service and internet connection unstable. For instance, in Amazonian regions there is almost no internet connection or cell phone service during torrential rains. In San Jacinto, a district in Tumbes, we found that virtually no one had Wi-Fi, not even the local health centre, so instead, people use their smartphone's hotspot for internet connection. However, the high temperature in Tumbes makes the smartphones overheat, which limits the internet connection during hot days. These are limitations from the Peruvian context that must be considered in the further development of the chatbot to ensure its technical feasibility for regions outside of Lima.

6.2.2. *Opportunities*

One of the most relevant observations is that a majority uses WhatsApp daily as their main messaging application, and this is consistent across the four sites we visited and across all groups of participants. Hence, almost everyone had their own phone with mobile data. We found that only two CHWs (6%) did not have their own phone and used their relative's phone's instead to communicate with others through WhatsApp, while the rest (94%) did. This means that the adoption of a chatbot in WhatsApp would be easier than in any other messaging platform due to the population's familiarity with it. However, we must also consider the privacy issues that may arise from family members sharing mobile devices since the chatbot could contain sensitive, personal and medical information.

Moreover, the prevalent use of WhatsApp among the population indicates that deploying Ana on this platform could address the issue of varying technological literacy levels. For example, when the diagnostic app was introduced to CHWs, those with lower technological proficiency experienced difficulties using it and displayed resistance, showing an obstacle to the app's usage and acceptance. These CHWs were required to acquire new skills and knowledge to effectively utilize the app, and this demand on their time and effort contributed to their reluctance. Conversely, a chatbot integrated within WhatsApp would likely not face such impediments, given the widespread familiarity with the application. In addition, WhatsApp can adapt to personal preferences (e.g. changing font size) and has additional accessibility features (e.g. reading out loud, voice input). This makes the chatbot more accessible as it broadens its reach of engagement across different levels of technological literacy.

7. CONCLUSION

We engaged with stakeholders that currently or could in the future, provide support to PLWD. Interventions used in HIC include trained care team navigators that are not available in LMIC like Peru. We explored if an adaptation of such interventions could include a chatbot, potentially one that uses Generative AI.

The study reveals that a chatbot can offer care navigation and emotional support to caregivers in Peru, and that the chatbot employing GAI, Ana-V2, achieves a higher caregiver satisfaction than the chatbot that does not employ GAI, Ana-V1.

The chatbots were evaluated by stakeholders through interviews and a satisfaction questionnaire based on eight success metrics. When comparing average scores, Ana-V2 outperformed Ana-V1 in every success metric. Ana-V2 scored 4 or over for all metrics on a 1–5 scale and approximated the scores of human CTNs. In contrast, Ana-V1 scored over 3.7 for only 1 metric, making it less successful in supporting caregivers. To determine the statistical significance of the differences in scores, we conducted the Wilcoxon signed-rank test. The test highlighted a significant statistical preference for Ana-V2 over Ana-V1 regarding understanding, empathy, knowledgeability, dependability and support, suggesting that GAI improves the user experience in these five aspects. On the other hand, the test found that there was no statistically significant difference in the scores between Ana-V1 and Ana-V2 regarding trust, usefulness and improving confidence, implying that GAI does not improve the user experience in these aspects. 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.

Moreover, the subsequent field study highlighted contextual challenges and opportunities for a dementia-focused chatbot like Ana. The challenges found are the lack of public awareness regarding dementia, the lack of local resources available for older adults and mental health, the varying educational and technological literacy levels, and the technical limitations in coastal and Amazonian regions due to extreme weather. On the other hand, we found that deploying Ana within WhatsApp presents an opportunity to overcome accessibility challenges due to the platform's widespread use in all Peruvian regions. These insights imply that even though there are contextual and technical challenges that need to be investigated further to evaluate the feasibility of a chatbot like Ana, there are also opportunities that could be leveraged to overcome these challenges and make Ana culturally sensitive to its Peruvian context.

Furthermore, the study results should be considered alongside its limitations. The most significant limitation is that all study participants for the first three phases are from Lima, the capital city. As said by one of them: 'Lima is not Peru', so these parts of 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 for all phases of the study from different regions and cultural communities to account for the diversity and disparities in Peru (World Bank, 2023).

Another important limitation is the number of participants and the small number of interactions they had with the chatbots during the evaluation. The time-consuming nature of a caregiving role meant that participants' availability was limited so the time where they could interact with the chatbots could not be longer than 1 h, resulting in few interactions. Ana was not evaluated thoroughly enough to provide a generalizable perception of Peruvian caregivers towards a dementia-care chatbot.

Additionally, all participants interacted with Ana-V1 first and then with Ana-V2. Even though this was done for participants

to get a sense of the possible questions that Ana-V1 could not explicitly answer so then they could repeat the same questions to Ana-V2, this sequence influenced participants' perceptions and limits the validity of the study. To tackle these limitations of the evaluation, a larger number of participants could be recruited so they could be split into two groups, where one group only interacts with Ana-V1, and the other group only interacts with Ana-V2 to investigate their perceptions without the influence of testing the other version and to gather greater input from the target users.

Furthermore, 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.

7.1. 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 ChatGPT, 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.

7.2. 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 participants of the project-at-large.

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 involve stakeholders such as caregivers, CHW, neurologists, psychologists, public health experts and legal advisors.

ACKNOWLEDGEMENTS

We are grateful to the Peruvian caregivers, community health workers and communities who make our research possible. We want to acknowledge the work, contribution and collaboration from all the members of the IMPACT Health Study group. The IMPACT Health project can be followed at https://impact-salud.org/.

FUNDING

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

DATA AVAILABILITY

The data and code underlying this article cannot be shared publicly to protect the privacy of the individuals who participated.

References

Aravena, C. J., Spencer, S. M., Verdugo, H. S., Flores, C. D., Gajardo, J. J. and Albala, B. C. (2016) Calidad de Vida en cuidadores informales de personas con demencia: una revisión sistemática de intervenciones psicosociales. *Revista Chilena de Neuro-Psiquiatría*, **54**, 328–341. https://doi.org/10.4067/s0717-92272016000400008.

de Arriba-Pérez, F., García-Méndez, S., González-Castaño, F. J. and Costa-Montenegro, E. (2022) Automatic detection of cognitive impairment in elderly people using an entertainment chatbot with natural language processing capabilities. *J. Ambient. Intell. Humaniz. Comput.*, **14**, 16283–16298.

Ayers, J. W. *et al.* (2023) Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Intern. Med.*, **183**, 589–596. https://doi.org/10.1001/jamainternmed.2023.1838.

Bernstein, A. *et al.* (2019) The role of care navigators working with people with dementia and their caregivers. *J. Alzheimers Dis.*, **71**, 45–55. https://doi.org/10.3233/jad-180957.

Brahnam, S. and De Angeli, A. (2012) Gender affordances of conversational agents. *Interact. Comput.*, **24**, 139–153. https://doi.org/10.1016/j.intcom.2012.05.001.

Braun, V. and Clarke, V. (2006) Using thematic analysis in psychology. *Qual. Res. Psychol.*, **3**, 77–101. https://doi.org/10.1191/1478088706qp063oa.

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

Carranza Ko, Ñ. (2021, February 19) *Analysis | Peru 's government forcibly sterilized Indigenous women from 1996 to 2001, the women say. Why?* Washington Post, Washington, D.C.. Retrieved from https://www.washingtonpost.com/politics/2021/02/19/perus-government-forcibly-sterilized-indigenous-women-1996-2001-why/.

Carrillo-Larco, R. M. *et al.* (2022) Peru – Progress in health and sciences in 200 years of independence. *The Lancet Regional Health – Americas*, **7**, 100148. https://doi.org/10.1016/j.lana.2021.100148.

Cayley, W.E. (2018) Four evidence-based communication strategies to enhance patient care. *Fam. Pract. Manag.*, **25**, 13–17. Retrieved from https://www.aafp.org/pubs/fpm/issues/2018/0900/p13.html.

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

Custodio, N. (2016) Vivir con demencia en Perú: ¿El sistema de salud está enfrentando la sobrecarga? *Revista de Neuro-Psiquiatría*, **79**, 1–2. Retrieved from http://www.scielo.org.pe/scielo.php?script=sci_arttext&pid=S0034-85972016000100001.

Custodio, N. and Montesinos, R. (2022) *Reconociendo los sintomas de pacientes con demencia*. Instituto Peruano de

Neurociencias (IPN), Lima, Peru. https://drive.google.com/file/d/1SeucsILmYpf9-oNiET0syjQIhOvRFjE9/view?fbclid=PAZXh0bgNhZW0CMTEAAab7VLZGB2o6GDRRwRl1VizpCBODONYIN4D_CJ7u3tVGmDrHtoKVbsA1gsQ_aem_tfhjjf8-QglvM8XntVNfIQ

Custodio, N., García, A., Montesinos, R., Escobar, J., & Bendezú, L. (2008) Prevalencia de demencia en una población Urbana de Lima-Perú: estudio puerta a puerta. *Anales de La Facultad de Medicina*, **69**, 233–238. Retrieved from http://www.scielo.org.pe/scielo.php?script=sci_arttext&pid=S1025-55832008000400003.

Custodio, N., Lira, D., Herrera-Perez, E., del Prado, L. N., Parodi, J., Guevara-Silva, E., Castro-Suarez, S., Mar, M., Montesinos, R. and Cortijo, P. (2014) Informal caregiver burden in middle-income countries results from memory centers in Lima - Peru. *Dementia & Neuropsychologia*, **8**, 376–383. https://doi.org/10.1590/s1980-57642014dn84000012.

Guerra, M., Ferri, C. P., Fonseca, M., Banerjee, S. and Prince, M. (2010) Helping carers to care: the 10/66 dementia research group's randomized control trial of a caregiver intervention in Peru. *Rev. Bras. Psiquiatr.*, **33**, 47–54.

Jiménez, S., Favela, J., Quezada, A., Ramachandran, R. and Juárez-Ramírez, R. (2022) Towards conversational agents to support informal caregivers of people with dementia: challenges and opportunities. *Program. Comput. Softw.*, **48**, 606–613. https://doi.org/10.1134/s036176882208014x.

Kruse, C. S., Fohn, J., Umunnakwe, G., Patel, K. and Patel, S. (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**, 278. https://doi.org/10.3390/healthcare8030278.

Laranjo, L. *et al.* (2018) Conversational agents in healthcare: a systematic review. *J. Am. Med. Inform. Assoc.*, **25**, 1248–1258. https://doi.org/10.1093/jamia/ocy072.

Lieberman, J. A. and Stuart, M. R. (1999) The BATHE method. *The Primary Care Companion to The Journal of Clinical Psychiatry*, **1**, 35–38.

Parodi, J., Montoya, J. C., Rojas, D. and Morante, R. (2011) Factores de Riesgo Asociados al Estrés Del Cuidador Del Paciente Adulto mayor. *Rev. Asoc. Colomb. Gerontol. Geriatr.*, **25**, 1504–1514.

Possin, K. L., Merrilees, J. J., Dulaney, S., Bonasera, S. J., Chiong, W., Lee, K. *et al.* (2019) Effect of collaborative dementia care via telephone and internet on quality of life, caregiver well-being, and health care use. *JAMA Intern. Med.*, **179**, 1658–1667. https://doi.org/10.1001/jamainternmed.2019.4101.

Rampioni, M., Stara, V., Felici, E., Rossi, L. and Paolini, S. (2021) Embodied conversational agents for patients with dementia: thematic literature analysis. *JMIR MHealth and UHealth*, **9**, e25381. https://doi.org/10.2196/25381.

Reimers, N., & Gurevych, I. (2019) Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3982–3992. Association for Computational Linguistics, Hong Kong, China. https://doi.org/10.18653/v1/d19-1410.

Ruggiano, N., Brown, E. L., Li, J. and Scaccianoce, M. (2018) Rural dementia caregivers and technology: what is the evidence? *Res. Gerontol. Nurs.*, **11**, 216–224. https://doi.org/10.3928/19404921-20180628-04.

Ruggiano, N., Brown, E. L., Roberts, L., Suarez, C. V. F., Luo, Y., Hao, Z. and Hristidis, V. (2021) Chatbots to support people with dementia and their caregivers: systematic review of functions and quality. *J. Med. Internet Res.*, **23**, e25006. https://doi.org/10.2196/25006.

Seah, C. E. L., Zhang, Z., Sun, S., Wiskerke, E., Daniels, S., Porat, T. and Calvo, R. A. (2022) Designing mindfulness conversational agents for people with early-stage dementia and their caregivers: thematic analysis of expert and user perspectives. *JMIR Aging*, **5**, e40360. https://doi.org/10.2196/40360.

Topo, P. (2008) Technology studies to meet the needs of people with dementia and their caregivers. *J. Appl. Gerontol.*, **28**, 5–37.

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

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

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

World Health Organization & International Telecommunication Union. (2021) *Be healthy, be mobile. A handbook on how to implement mDementia*. WHO, Geneva. https://www.who.int/publications/i/item/9789240019966.

Interacting with Computers, 2025, **37**, 430–443
https://doi.org/10.1093/iwc/iwae057
Article