Development and Qualitative Evaluation of R-Speak: Acceptability and Usability of a Smartphone App System Using AI to Enhance Communication in People With Expressive Aphasia

Abdel-Karim Al-Tamimi, Jacob Andrews, Jacqueline Benfield, Cath Sweby, Chris Gilmartin, Rebecca Lindley, Diane Trusson, Molly Dziunka, Dee Webster, Kathryn Radford

Submitted to: Journal of Medical Internet Research
on: February 13, 2026

Table of Contents

Original Manuscript **5**
Supplementary Files **30**
Figures **31**
Figure 1 ..... **32**
Figure 2 ..... **33**
Figure 3 ..... **34**
Figure 4 ..... **35**
Figure 5 ..... **36**
Figure 6 ..... **37**
Figure 7 ..... **38**

Development and Qualitative Evaluation of R-Speak: Acceptability and Usability of a Smartphone App System Using AI to Enhance Communication in People With Expressive Aphasia

Abdel-Karim Al-Tamimi[1, 2*] MSc, BSc, PhD; Jacob Andrews[3] MA, PhD; Jacqueline Benfield[4, 5*] BSc, MSc, PhD; Cath Sweby[6] BSc, MSc; Chris Gilmartin[7, 8]; Rebecca Lindley[5]; Diane Trusson[5]; Molly Dziunka[5]; Dee Webster[9, 10]; Kathryn Radford[3, 5*] PhD

[1] School of Computing and Digital Technology Sheffield Hallam University Sheffield GB
[2] Computer Engineering Department Yarmouk University Irbid JO
[3] Biomedical Research Centre University of Nottingham Nottingham GB
[4] Derbyshire Community Health Services NHS Trust Chesterfield GB
[5] Centre for Rehabilitation and Ageing Research University of Nottingham Nottingham GB
[6] Northern Care Alliance NHS Foundation Trust Manchester GB
[7] Nottingham University Hospitals NHS Trust Nottingham GB
[8] Mental Health and Clinical Neurosciences Academic Unit University of Nottingham Nottingham GB
[9] Sheffield Teaching Hospitals NHS Foundation Trust Sheffield GB
[10] Health Sciences School Human Communication Sciences University of Sheffield Sheffield GB
[*] these authors contributed equally

Corresponding Author:
Abdel-Karim Al-Tamimi MSc, BSc, PhD

School of Computing and Digital Technology
Sheffield Hallam University
City Campus, Howard Street
Sheffield
GB

Abstract

Background: Aphasia, an acquired language disorder impacting the ability to understand and produce language, greatly impacts effective communication. Large language models (LLMs) like GPT-5 offer potential to support communication by generating human-like sentences and coherent speech and subsequently enhance functional communication for individuals with aphasia.

Objective: Co-produce a system using LLMs to support communication and explore potential utility and acceptability in people with mild-to-moderate aphasia.

Methods: : Using the Double Diamond approach: Phase 1: Discover and define; Stroke survivor PPI group (n=5) and research team used MoSCoW prioritisation to develop and prioritise ideas and co-design a software solution (R-SPEAK) to augment verbal communication. Phase 2: Develop and demonstrate; eight LLM's were evaluated for interpretation using existing datasets from AphasiaBank, ratified by team members. The best-performing model was used for prototype development. Prototype testing was undertaken with 4 people with aphasia (PwA) and 1 carer using semi-structured interviews. A healthcare professional (HCP) focus group (n=6) evaluated the concept and prototype. The topic guide was informed by, and themes from thematic analysis were mapped onto the Technology Acceptance Model (TAM). Participants rated usability with the System Usability Scale (SUS).
Phase 3: Refine and resign. To increase the processing speed, we systematically evaluated 12 lightweight open-weight LLMs (0.5B–3.8B) on interpreting real aphasic speech, using clinician-curated dialogues and an LLM-as-a-judge framework assessing relevance, faithfulness, and completeness.

Results: Initially Mixtral (8x7b), was the best-performing LLM for aphasic utterances, and was utilised for the prototype. PwA rated R-SPEAK as good using the SUS (mean 75). Themes extracted from qualitative data mapped across all three TAM constructs. Attitude towards using; PwA had high hopes whilst clinicians demonstrated more caution about its benefits.

Perceived ease of use; participants found it easy to use but it may be more challenging for those with other post stroke impairments or more severe aphasia and training might be needed. Perceived usefulness: R-SPEAK could be useful in many scenarios and has potential to improve independence for PwA. Recommendations for development included improved accuracy, speed and modifications to the interface according to the individual's needs. Further refinement demonstrated that Qwen (2.5:3b) achieved the strongest overall performance with high faithfulness and sub-second latency, while models under 1.5b parameters showed pronounced hallucination issues, indicating a lower bound on model capacity for reliable clinical speech interpretation.

Conclusions: Our co-designed R-SPEAK prototype was considered acceptable to patients. Next steps involve ongoing refinement, development of a phone-based app for feasibility testing in a larger and broader cohort of people with mild-to-moderate aphasia.

(JMIR Preprints 13/02/2026:93518)
DOI: https://doi.org/10.2196/preprints.93518

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✔ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✔ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http
No. Please do not make my accepted manuscript PDF available to anyone. I understand that if I later pay to participate in <a href="https:/

Original Manuscript

Original Paper

Development and Qualitative Evaluation of R-Speak: Acceptability and Usability of a Smartphone App System Using AI to Enhance Communication in People With Expressive Aphasia

Abdel-Karim Al Tamimi[1,10*], PhD; Jacob Andrews[2], PhD; Jacqueline Benfield [3,7]*, PhD; Cath Sweby [4] ; Chris Gilmartin [5,6] , MD; Rebecca Lindley [7]; Diane Trusson [7], PhD; Molly Dziunka [7], MD; Dee Webster[8, 9]; Kathryn Radford [2, 7]*, PhD

[1] Sheffield Hallam University, Sheffield, United Kingdom,
[2] Nottingham Biomedical Research Centre, Nottingham, United Kingdom,
[3] Derbyshire Community Health Services NHS Trust, Derby, United Kingdom,
[4] Northern Care Alliance NHS Foundation Trust, Manchester, United Kingdom,
[5] Nottingham University Hospitals NHS Trust
[6] Mental Health and Clinical Neurosciences Academic Unit, School of Medicine, University of Nottingham, United Kingdom
[7] Centre for Rehabilitation and Ageing Research, University of Nottingham
[8] Sheffield Teaching Hospitals NHS Foundation Trust, Sheffield, United Kingdom
[9] Human Communication Sciences, Health Sciences School, University of Sheffield, Sheffield, United Kingdom
[10] Yarmouk University, Irbid, Jordan
[*] these authors contributed equally

Abstract

Background: Aphasia, an acquired language disorder impacting the ability to understand and produce language, greatly impacts effective communication. Large language models (LLMs) like GPT-5 offer potential to support communication by generating human-like sentences and coherent speech and subsequently enhance functional communication for individuals with aphasia.
Objective: Co-produce a system using LLMs to support communication and explore potential utility and acceptability in people with mild-to-moderate aphasia.
Methods: Using the Double Diamond approach: Phase 1: Discover and define; Stroke survivor PPI group (n=5) and research team used MoSCoW prioritisation to develop and prioritise ideas and co-design a software solution (R-SPEAK) to augment verbal communication. Phase 2: Develop and demonstrate; eight LLM's were evaluated for interpretation using existing datasets from AphasiaBank, ratified by team members. The best-performing model was used for prototype development. Prototype testing was undertaken with 4 people with aphasia (PwA) and 1 carer using semi-structured interviews. A healthcare professional (HCP) focus group (n=6) evaluated the concept and prototype. The topic guide was informed by, and themes from thematic analysis were mapped onto the Technology Acceptance Model (TAM). Participants rated usability with the System Usability Scale (SUS).
Phase 3: Refine and resign. To increase the processing speed, we systematically evaluated 12 lightweight open-weight LLMs (0.5B–3.8B) on interpreting real aphasic speech, using clinician-curated dialogues and an LLM-as-a-judge framework assessing relevance, faithfulness, and completeness.
Results: Initially Mixtral (8x7b), was the best-performing LLM for aphasic utterances, and was utilised for the prototype. PwA rated R-SPEAK as good using the SUS (mean 75). Themes extracted from qualitative data mapped across all three TAM constructs. Attitude towards using; PwA had high

hopes whilst clinicians demonstrated more caution about its benefits. Perceived ease of use; participants found it easy to use but it may be more challenging for those with other post stroke impairments or more severe aphasia and training might be needed. Perceived usefulness: R-SPEAK could be useful in many scenarios and has potential to improve independence for PwA. Recommendations for development included improved accuracy, speed and modifications to the interface according to the individual's needs. Further refinement demonstrated that Qwen (2.5:3b) achieved the strongest overall performance with high faithfulness and sub-second latency, while models under 1.5b parameters showed pronounced hallucination issues, indicating a lower bound on model capacity for reliable clinical speech interpretation.
Conclusions: Our co-designed R-SPEAK prototype was considered acceptable to patients. Next steps involve ongoing refinement, development of a phone-based app for feasibility testing in a larger and broader cohort of people with mild-to-moderate aphasia.

Keywords: Aphasia; AAC; digital health; communication; health technology; GenAI; LLM; Human-AI Interaction; Affective computing.

Introduction

Aphasia is an acquired communication disorder and affects around 41% of UK stroke survivors (Mitchell et al., 2021) [1]. It can also occur in other acquired neurological conditions such as brain tumours and fronto-temporal dementias, particularly Primary Progressive Aphasia. There is a broad phenotype for aphasia, where some individuals are unable to say single words or understand language at all, others have more mild difficulties such as reduced sentence complexity or word-finding errors which may be sound-based (phonological) or meaning-based (semantic) [2, 3]. Speech production may be more effortful, with a reduction in speech rate, or with changes in intonation and prosody, leading to an overall reduction in verbal fluency and efficiency. Aphasia can affect both spoken and written language with wide-ranging, life-changing consequences [4]. All severities of aphasia can profoundly affect a person's life, including their ability to access rehabilitation, engage in daily activities, maintain relationships, and participate in society, including returning to work and leisure [5]. This can often lead to social isolation, low mood, and reduced quality of life [6, 7] . Whilst rehabilitation is important to help people living with aphasia (PwA) regain as much language as possible, individuals often continue to experience aphasia and its impact over the longer term and are faced with the challenge of learning to live with aphasia using their preserved language.

Augmentative and alternative communication (AAC) options, such as digital or paper-based word, picture or letter grids for people with aphasia, aim to improve communication of basic needs [8]. These systems often have a limited vocabulary and require pre-programming by a clinician or carer. Furthermore, the purpose of communication is much more than expressing basic needs and people with communication difficulties are more dependent on co-constructed interactions with a communication partner [9].

Artificial Intelligence (AI) may offer a solution and has been used to augment non-impaired speech in several ways, including predictive texting [10] , and speech-to-text technologies [11] , language translation and interpretation [12], speech synthesis [13] and voice cloning [14].

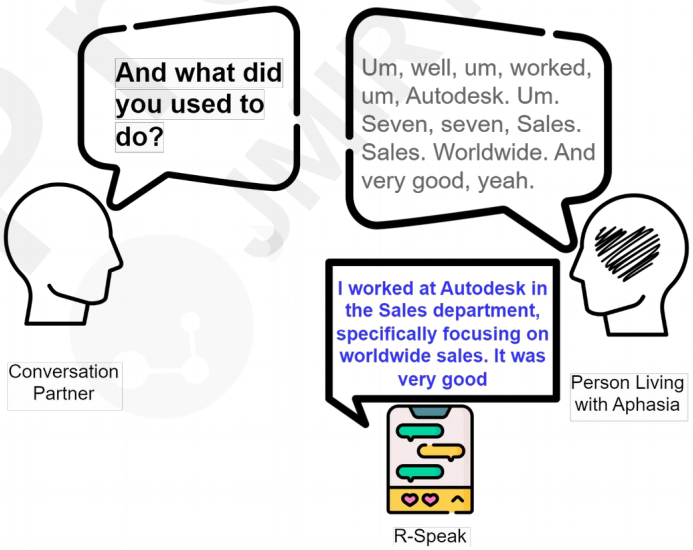
Text-to-speech (TTS) technologies draw on recent advances in natural language processing (NLP) techniques, and the introduction of large language models (LLMs) like Google Gemini, Open AI's GPT-3/4, and Meta's Llama. LLMs operate on user-provided prompts and can generate coherent paragraphs of text [15]and predict and generate human-like sentences by learning from vast textual

datasets. While LLMs differ in efficiency and accuracy, prompt engineering (the art of crafting effective prompts) can yield varying results [16].

Aphasia often includes agraphia (written language impairment) and cannot utilise text to speech technologies to augment their communication [17]. By incorporating automatic speech recognition (ASR) technology into these systems, this approach enables the removal of barriers to accessing technology for those with communication difficulties and in the development of personalised AAC devices to support communication. However, ASR have been trained on a largely non-impaired speech dataset, and accuracy in interpreting impaired speech is much lower [18]. Much work is being done to improve ASR for dysarthric speech (motor speech disorder) by training LLMs with disordered speech [19], but this has not been explored for people with linguistic impairments such as aphasia.

Moreover, AI's ability to learn from the daily activities and contextual data of the individual with aphasia means it can offer bespoke solutions tailored to the individual's needs. By learning the user's behaviour, mistakes and preferences, these tools can adapt to offer more accurate or efficient communication support over time.
Leveraging LLMs in this way offers the potential to reconstruct aphasic speech and translate it into spoken language, improving comprehensibility and communication. As demonstrated in (Figure 1), our proposed *Revolutionising Speech Enhancement in Aphasia Using Knowledgeable-AI* or (R-SPEAK) technology has the potential to support clearer communication, helping the person with aphasia regain control over their personal communication interactions, increasing autonomy, reducing social isolation and improving their quality of life. It has the potential to revolutionise the lives of people living with aphasia, caregivers, and healthcare professionals involved in their rehabilitation.

Figure 1. Example of how R-SPEAK converts real-world aphasic speech into comprehensible speech.



Aim

Working with people with aphasia and stroke clinicians, we aimed to co-design and test a prototype for a portable device (e.g. mobile device) to enhance the clarity and fluency of verbal expression of people living with mild-to-moderate expressive aphasia (PwA).

Study Objectives

- To co-design and develop a prototype portable device with people with aphasia and stroke clinicians that incorporates large language models to enhance the clarity and fluency of verbal expression in mild-to-moderate expressive aphasia.
- To evaluate the technical effectiveness and usability of the prototype device through testing with people with aphasia in real-world communication scenarios, assessing its impact on communication clarity and fluency.
- To assess the acceptability, perceived utility, and implementation feasibility of the device among people with aphasia and healthcare workers, including identification of barriers and facilitators for adoption.

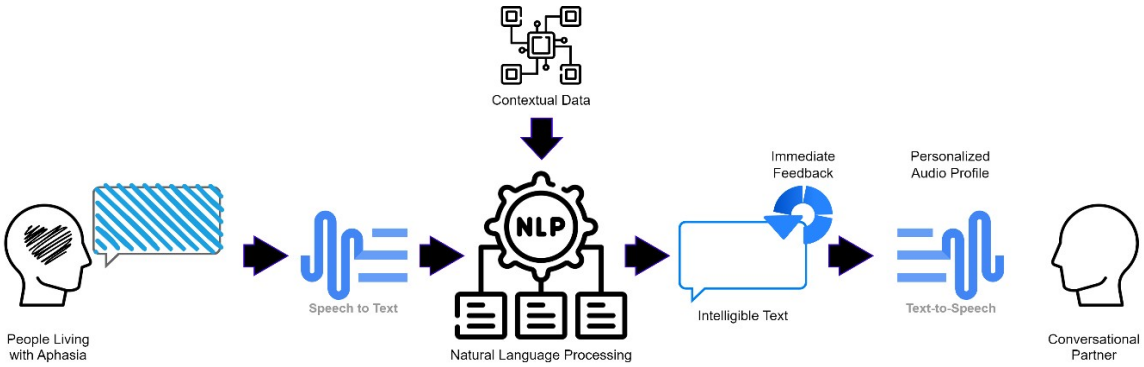
Methods

This study employed a mixed-methods participatory design approach using the Double Diamond methodology to co-design and evaluate a prototype of a communication support device for people with mild-to-moderate expressive aphasia [20]. Phase 1 (Discover and Define) involved a stroke survivor Patient and Public Involvement (PPI) group (n=5) working collaboratively with the research team to identify communication needs and co-design requirements. Using MoSCow prioritisation methodology [21], the group developed and prioritised the key features to inform the design of R-SPEAK, a software solution designed to augment verbal expression in people with aphasia. This phase utilised user-centred design principles to define the core functionalities and user interface requirements for the prototype device.
Phase 2 (Develop and Demonstrate) comprised technical development and evaluation components. Eight large language models were systematically evaluated for their ability to interpret aphasic utterances using existing transcripts from AphasiaBank [22], with performance ratified by team members to select the optimal model for prototype integration. The resulting R-SPEAK prototype underwent usability testing with four PPI group members through semi-structured interviews, followed by concept and prototype evaluation via a healthcare professional focus group (n=6). All qualitative data were analysed using inductive thematic analysis within the Technology Acceptance Model (TAM) framework [23] to assess perceived utility, ease of use, and acceptance of the system.

Concept

Our methodology capitalises on the recent advancements in natural language processing (NLP), particularly the emergence of Large Language Models (LLMs), to enhance the speech comprehension of individuals living with aphasia (PwA). Our concept is for a system that utilises LLMs' language-understanding (LU) capabilities to build upon the utterances of people with mild-moderate aphasia to produce coherent speech. The process begins by recording the speech of PwA and converting it into text using speech-to-text or automatic speech recognition (ASR) technology, as shown in (Figure 2). We utilise this textual data and the speech of the conversation partner/interlocutor to craft precise prompts for LLMs.

Figure 2. Overview of the R-SPEAK system workflow for enhancing communication in people with aphasia (PwA).

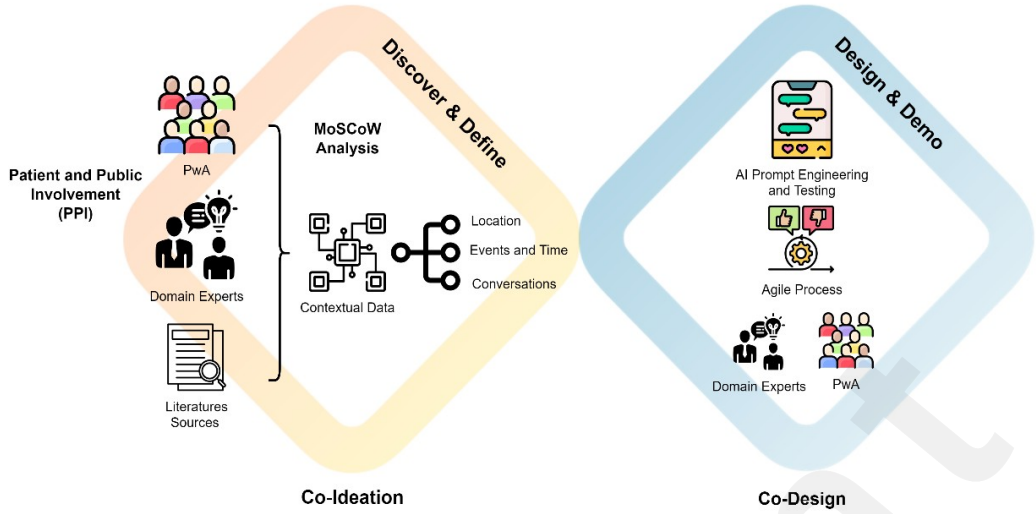


The LLM analyses the prompt-provided data, endeavouring to generate coherent interpretations of the speech by completing missing words, inferring unclear terms, and eliminating unnecessary filler words. The generated text is presented to the user, who selects the most suitable transformation that effectively conveys their intended meaning, with the option to edit and manually adjust the text as needed. Subsequently, the chosen text is transformed into spoken words using text-to-speech (TTS) technology, with the option to use a synthesised voice personalised to the patient's audio profile. The resulting clear, coherent sentences can be easily understood by the person's conversation partner. The modular design of our system provides several key advantages: it allows for personalisation to individual user needs, can continuously improve as language models advance, integrates well with existing assistive technologies, and focuses on locally deployable LLMs with the option to access cloud-based LLMs through internet connectivity when needed.

Design

Our research methodology employed the Double Diamond design approach, a systematic framework that structures the design process into four stages across two phases, as shown in (Figure 3) [20]. Phase 1 encompassed the Discover and Define stages, systematically exploring communication challenges faced by people with aphasia and identifying specific user needs through patient and public involvement (PPI). This involved divergent exploration of the problem space followed by convergent synthesis to define clear design requirements.
Phase 2 comprised the Develop and Demonstrate stages, focusing on iterative development of the R-SPEAK technological solution and subsequent validation through user testing. Throughout both phases, participatory co-design principles ensured meaningful involvement of people with aphasia and stroke clinicians through workshops, interviews, and focus groups conducted across multiple iterative stages. This collaborative approach ensured the technological solution remained grounded in real-world user needs and clinical practice requirements. All qualitative findings are reported in accordance with the Consolidated Criteria for Reporting Qualitative Research (COREQ) [24] checklist to ensure methodological transparency and rigour.

Figure 3. Double Diamond approach to prototype development.



Phase 1: Discover and define

As part of the Discover and Define phase of the Double Diamond design process, Patient and Public Involvement (PPI) was integral to identifying communication challenges and clarifying user needs.

Patient and Public Involvement (PPI)

We convened a PPI group consisting of people with experience of living with aphasia and their carers. We recruited individuals with aphasia through established networks, including members of the research team's existing contacts and participants from stroke survivor groups such as the Nottingham Stroke Partnership Group and Speakeasy (Stockport), following an initial presentation and invitation to participate.

We involved the PPI group in several ways: 1) to inform system development, 2) to support recruiting for interviews to evaluate the system for usability, accessibility, acceptability and usefulness, and 3) to review participant-facing materials to ensure accessibility for people with aphasia before use.

We held two online PPI group meetings; one with people with aphasia and their carers, and another with the research team who were also healthcare professionals with experience of working with PwA (KR, JA, CW, CG, JB). During the meetings, we presented the concept both verbally and with simple graphics. We used yes/no, closed and open questions to facilitate discussion and gather feedback on the proposed system.

Co-ideation; we used MoSCow prioritisation (Must Have, Should Have, Could have, Will not Have) [21] to systematically evaluate ideas and prioritise the features most critical to people with aphasia (PwA). This approach guided our understanding of which elements were essential for ensuring usability, accessibility, acceptability, and practical utility in real-world contexts. Through this structured process, we also identified the types of contextual data required to enhance the accuracy and relevance of the app’s outputs.

To facilitate effective communication and engagement with participants, we used visual aids such as images, keywords, and simplified informatics to help convey complex ideas. Group sessions were led by facilitators experienced in working with individuals with aphasia, who ensured sessions were inclusive by allowing extended response times and using yes/no or closed-ended questions to accommodate expressive language difficulties.

Phase 2: Design and Demo

App design (System frontend)

We concentrated our efforts on designing an intuitive user interface and refining the underlying

prompt engineering approach. Our computer scientist (AA) developed a low-fidelity prototype of the application's frontend using Proto.io, a flexible prototyping platform well-suited for rapid iteration. This prototype was informed by insights gathered through engagement with Public and Patient Involvement (PPI) groups and close collaboration with clinical team members. Their feedback played a critical role in shaping both the visual layout and functional requirements of the interface, ensuring that it is user-friendly, accessible, and aligned with real-world needs in clinical settings.

Technical development (System Background)

To develop a functional prototype capable of reconstructing intelligible speech, one team member (MD) extracted 30 question-and-answer pairs from AphasiaBank [22], a repository of transcribed conversations with individuals living with aphasia. This dataset includes contributions from individuals with mild to moderate aphasia and is widely used in aphasiology research. The selected dialogue pairs were converted from the Codes for the Human Analysis of Transcripts (CHAT) format [22] into plain text using a bespoke conversion tool developed by (AA).

These transcribed question–response pairs were then processed using six open-source, locally deployable large language models (LLMs): Mixtral, Gemma, Qwen, Llama 3, Phi-3, and WizardLM. These models were chosen based on their advanced reasoning capabilities and demonstrated adherence to user instructions. Several zero-shot prompting strategies were tested, in which the prompts included contextual information about aphasia and the characteristic differences between aphasic and typical speech patterns.

The outputs generated by the LLMs were independently evaluated by five members of the research team (CG, JA, JB, RL, and AA), all with clinical and/or technical expertise in LLMs. Evaluators assessed each LLM's ability to interpret and reconstruct the intended meaning behind the original aphasic responses. For each question–answer pair, the team selected the most accurate LLM-generated response. Based on this expert consensus, the highest-performing LLM, in this case Mixtral, was identified and subsequently integrated into the development of the R-SPEAK prototype.

Interviews and focus groups

Development of the topic guide

The aim of the qualitative interviews and focus group was to understand HCPs and People with aphasia's views and perception of the R-SPEAK prototype, to gather insights on its functionality, and understand potential benefits and challenges. Two topic guides informed by the TAM [23] were developed to address the different perspectives of HCPs and people with aphasia. Developed by Davis [23], the TAM is used to shed light on the processes involved in the acceptance of technology. TAM suggests that attitudes towards using innovative technology are based on perceptions of both usefulness and ease of use [25].

The topic guides were developed by RL, a female rehabilitation psychology researcher with extensive qualitative research experience with people with stroke and HCPs. The guides were reviewed and edited based on feedback from the wider research team with SLT expertise (JB), extensive qualitative research experience (KR, JA), as well as expertise related to AI (AA).

Recruitment

Recruitment took place between July to August 2024, with interviews and the focus group conducted in September 2024.

Interviews

People with aphasia (all severities) were recruited via known contacts of the research team and aphasia support groups (Speakeasy), using convenience and snowball sampling [26]. Known contacts were approached by email or phone and an aphasia friendly recruitment poster was shared to potential participants in aphasia support groups, inviting them (or their carers) to contact the

research team if interested. Those expressing interest in the interviews were sent an information sheet and consent form by post or email, depending on preference and asked to complete and return the consent form (using a Microsoft form if online, or using a (provided) stamped addressed envelope if via post) within 2 weeks of receiving the information sheet. Participants were offered an opportunity to meet, either in person or online, with a member of the research team, to have the information sheet and consent form read aloud with aphasia friendly pictorial participant information to assist understanding, and completion of the consent form where informed consent was demonstrated.

Healthcare Professionals Focus Group

Stroke clinicians were recruited through the research networks of the members of the research team. Participant information sheets were sent to those interested and informed consent was collected using an online consent form.

Materials

To give participants an understanding of the tool an aphasia friendly power point presentation was prepared for the interviews/focus group.
For the interviews with PwA, the web-based R-SPEAK prototype allowed it to be shared on the participants screen during the interviews. The use case for exploration was booking a holiday (i.e. we aimed to test if the system could be helpful in this use case). It consisted of eight pre-programmed questions that had been co-designed by the research team and PPI group members and included questions such as “Where would you like to go on holiday”. When the participant was ready they could use their mouse to click to start recording whilst they answered and clicked again to stop. The prototype then produced an output, displayed on the screen to the interviewer and interviewee.
The SUS was used for objectively measuring perceptions of usability [27]. It consists of 10 statements such as “I think I would like to use this system frequently” and five response options from strongly disagree to strongly agree. Scores over 70 demonstrate system acceptability. A slide deck was prepared in advance with one question per slide for sharing with the interview participants.

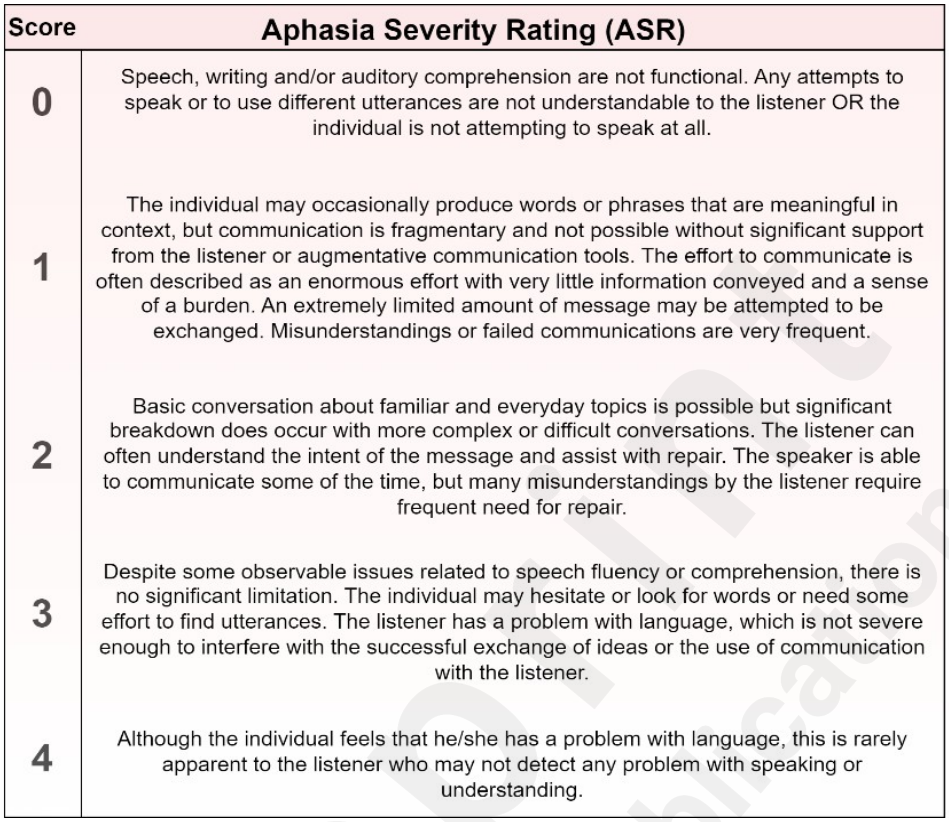
Procedure

Semi-structured interviews and the focus group were conducted by SLT and post-doctoral fellow (JB) and the research fellow (RL) via MS Teams or Zoom. In the case of the interviews, participants were offered in-person interviews in their own homes if preferred. Potential participants were advised a carer or an SLT could attend to support their communication.

At the beginning of both the interviews and focus group the participants were introduced to the interviewers and shown the introduction presentation with a demonstration of R-SPEAK. Participants were asked a series of semi-structured questions on usability, accessibility, acceptability, and usefulness of the system in line with the TAM [23].

In addition, during the interviews the participants tested the prototype by answering the prepared questions and receiving the LLM generated output for the eight questions. To better understand the prototype in relation to the target population participants were asked about their communication difficulties and experience of using technology and communication apps. Participants’ aphasia severity was rated by an SLT (JB) by listening to the interview recording and applying the Aphasia Severity Rating scale (Figure 4) [28] which enables a score of 0, 1, 2, 3 or 4 to be assigned, where 0 = No functional communication and 4=very mild impairment that may be undetectable to the listener. At the end of the interview participants were asked to rate prototype usability using the SUS. Interviews lasted under one hour and the focus group 1.5 hours.

Figure 4. Aphasia Severity Rating (ASR) Scale. This scale provides a structured index of aphasia severity, ranging from complete language impairment (0) to minimal or undetectable difficulties (4), based on speech, writing, and comprehension abilities. Adapted from [28].



Ethics and Payment

The study was approved by the Research Ethics Committee at Sheffield Hallam University (approval number ER64821246). All participants provided written informed consent before taking part. Interview participants were paid £50.00 for taking part.

Data Analysis

The SUS raw scores were summed and converted into scores out of 100 as described by Brooke 1986 [27]. Individual and mean scores across the group are presented visually. Interviews and focus groups were recorded in either (.mp3, .wav, or .mp4) format using encrypted computers owned by the University of Nottingham. They were then transcribed and analysed independently by two researchers (JB, DT) using NVIVO and an inductive approach of data familiarisation and development of themes [29, 30]. To enhance reliability, themes from both groups were discussed and aligned into one combined construct before being mapped onto the TAM constructs (i.e. attitude towards using technology, perceived ease of use, and perceived usefulness) [23]. Recordings and anonymised transcripts were stored on a secure University of Nottingham password protected server.

Results

Phase 1 Discover and define

Patient and Public Involvement (PPI)

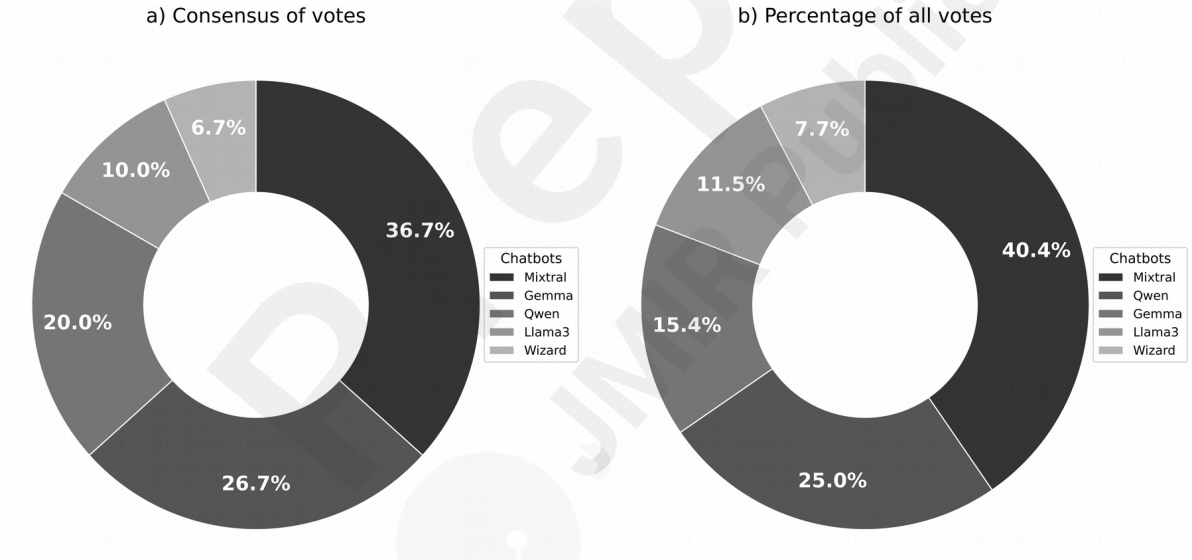
Three stroke survivors and one carer were co-opted to the R-SPEAK PPI group and were involved in the two co-ideation groups during app development.

Phase 2 Develop and Demo

App design and technical development

The LLMs were rated by five mixed experts. Mistral and Llama 2 were excluded from consideration early due to their poor performance on basic language testing. Mixtral outperformed all other LLMs, followed by Gemma and then the lightweight Qwen (Figure 5). This consensus highlights Mixtral as the most effective model in interpreting responses from individuals living with aphasia among tested models.

Figure 5. Expert evaluation of LLM performance through consensus and aggregate voting. (a) Consensus votes represent cases where all five experts unanimously selected the same model as superior. (b) Percentage of all votes shows the distribution across all individual expert votes.



Interviews and focus groups

Four stroke survivors with aphasia were involved in prototype testing and interviews. Three were assessed as having moderate aphasia with one having milder impairment on the Aphasia Severity Rating scale. Characteristics of participants can be found in (Table 1).

All had some previous experience with technology, including computers and smartphones which they used for emails and online meetings but none of the group used AAC apps or had previous experience with AI.

Five SLTs and one physiotherapist (PT) were recruited for the focus group. All members had experience of working with people who had aphasia and across different clinical settings. (Table 2)

details their experience.

Table 1. Characteristics of participants

	Self-reported difficulties with communication	Ability to use technology	Experience with AAC communication apps	Aphasia Severity Rating (by SLT)
P1	Word finding difficulties and reading and writing difficulties with longer text	Smart phone and computer for short emails, texts & video calls	None	3 - mild
P2	Able to communicate but talking is difficult, especially numbers. Reading is okay	Computer and smartphone	Used apps for speech therapy exercises.	2 - moderate
P3	Difficulties getting words out and reading more than single sentences	Computer, ipad and smartphone and uses for online meetings, accounts, emails	None	2 - moderate
P4 & C1	Reduced fluency, difficulty getting out words and reading, often missing key words	Has a smartphone and laptop. Needs support with emails, getting onto video calls	None	2- moderate

Table 2. Characteristics of HCP participants [a]

Participant	Details
SLT 1	Works in neuro rehabilitation centre with acquired brain injury patients
SLT 2	Advanced clinical practitioner. Stroke specific. Lots of patients with aphasia every day.
SLT 3	Works in a community neuro rehabilitation team with acquired brain injury and stroke patients.
SLT 4	Works in stroke rehabilitation. Regularly sees people with aphasia, dysarthria and apraxia of speech.
SLT 5	Works in stroke rehabilitation wards. Worked with a lot of people who have aphasia.

PT 1	Works with SLT5 on acute and chronic stroke and outpatient wards.

[a] Abbreviations: SLT = Speech and Language Therapist; PT = Physiotherapist

Analysis of the interviews revealed similarities and differences in perceptions of R-SPEAK from the perspectives of people with aphasia (and carers) and HCPs. Themes are presented under the constructs of the TAM framework [23] themes are summarised in (Table 3).

Attitudes towards using

High hopes vs Scepticism or somewhere in between.

Overall attitudes varied between the PwA and HCPs. PwA and carers had high hopes that this was a life-changing technology, but clinicians were more sceptical about it being groundbreaking as shown in (Textbox 1).

Textbox 1. Attitude towards using: Mixed reactions: excitement tempered by realistic concerns

“I think it's marvellous really. Yeah, really good.” (P2)
“I'm really excited about this, aren't you?” (C1)
“Don’t want to give people false hope, thinking that this is going to solve everything, because actually conversation has so many more layers to it than just the language. So I’m a bit worried about people thinking that it's going to solve things” (SLT3).

There was cautious optimism by both groups, HCPs suggested that it could be good to have another tool in the tool box and may be less stigmatised than other AAC aids in use but both groups want to see R-SPEAK develop and be used in action to make their minds up about it, as shown in (Textbox 2).

Textbox 2. Attitude towards using: Cautious openness as potential tool, not cure

“I don't think I I've got enough knowledge from it to say whether it would be really good or yeah, because I think it needs to some any because that's a prototype and it only one scenario and it wasn't very long.” (P1)
People might be more willing to use it than things that look potentially stigmatised, like some of the older AACs (SLT4)
“It’s sort of keeping a bit open to trying it, but not presenting it as an answer to aphasia. So we might see it as one more tool that would work with some people” (SLT4).

False assumptions about AI.

HCPs expressed concern about using AI in healthcare apps but specifically about security of PwA’s personal information. Some PwA might not be able to make informed choices about using the technology if it is based on reading data sharing agreement text, as shown in (Textbox 3).

Textbox 3. Attitude towards using: AI controversy as potential engagement barrier

“AI in general has got a bit of controversy as well. So I also wonder if that might be a barrier to some people engaging” (SLT1).

However, as shown in (Textbox 4), the PwA we interviewed were not concerned at all.

Textbox 4. Attitude towards using: Acceptable with personal control and opt-out option

“In terms of sort of using the AI of in this sort of yeah, I think it would be OK. But but sort of having an app on your phone that you can control that I think it's OK. Yeah, yeah, yeah, I can turn it off anyway.” (P1)

Costs could be prohibitive

HCPs voiced concerns about costs to PwA given their experience of other apps and software options. However, as shown in (Textbox 5), this was not raised as a concern by PwA who we interviewed.

Textbox 5. Attitude towards using: Pricing model concerns and inequality implications

“Do you pay a tenner and you have it forever? Or are you having to pay like £15 a month forever? […] obviously the more it's charging, the more inequality we get with it” (SLT4).

Perceived Ease of Use

Dependent on type and severity of aphasia.

Both PwA and HCPs acknowledged that there may be people who would find R-SPEAK more difficult to use or who would not be able to use it all due to their aphasia. They gave specific examples, as shown in (Textbox 6), like people who were unable to read or read quickly enough, people with auditory comprehension difficulties, those with jargon aphasia where words may be unrelated to the target, and those with no, or very little, expressive language or with apraxia of speech.

Textbox 6. Perceived ease of use: Reading requirements and speech complexity limit accessibility

“You have to read. You have to read the text and say that's the place. And then, yes, some people can't read.” (P3)
“A lot of our patients that have jargon speech, some of it's just non words. I don't think that would be helpful for the AI if what they're saying isn't even real words […] but even with real words a lot of it's completely out of context, I think it'd be difficult for the AI to learn from that. So I think those patients should find that hard to use” (SLT5).

Other neurological impairments may pose challenges.

Both groups, as shown in (Textbox 7), reported that other stroke symptoms such as unilateral motor weakness, visual impairment or cognitive difficulties affecting memory, self-monitoring or, initiation might make using R-SPEAK difficult or impossible.

Textbox 7. Perceived ease of use: Physical and cognitive demands pose usability challenges

“Yeah, some people, it's disabilities, so they can't use their left arm or their right arm, and so they have to tap it.” (P3)
You need to be able to self-monitor the output and say whether it's right or wrong and how much output there is. Can you cognitively cope with that burden? (SLT2).

Using technology is easier for some people.

The PwA we interviewed found the tech quite easy themselves. Both groups, as shown in (Textbox 8), thought that some people would be able to use R-SPEAK easily, perhaps those already using their smartphones and messaging apps. However, both groups thought that it might be more challenging for some who were less tech savvy or not used to using smartphones.

Textbox 8. Perceived ease of use: Prior tech familiarity determines ease of use

"So probably people who were using phones to some extent before acquiring aphasia" (SLT4)
"I think it's quite, quite easy to use and I think younger people will be even more easy to do use." (P1)
"And I think although he would really benefit from something like this because his speech, he's hesitant. (But) he's not confident with technology, neither would he have the right equipment." (C1)

Try before you buy

Given that some people may find R-SPEAK more difficult to use for a range of reasons and with potential cost implications, HCPs, as shown in (Textbox 9), voiced the need for a 'try before you buy' option.

Textbox 9. Perceived ease of use: Need for free trial before purchase commitment

My first thought was cost […] of a lot of the people I'm seeing, the first thing they think is actually they don't want to pay for something until they know it's worth it, and whether there's a free trial […] how long would people get to try it out? Because it's quite rare that I come across people who are happy to pay for something straight away without trying it (SLT3).

Some training needed.

Both groups, as shown in (Textbox 10), felt that PwA would need some initial training to use R-SPEAK, it might be something that an SLT could do or perhaps a training video guide is built into the system.

Textbox 10. Perceived ease of use: Support from family or therapists' aids usage

"A family member or someone else could actually help somebody use the app potentially. It's the sort of thing that somebody could help them with if needed" (SLT3).
"Speech therapist say […] outlines the system. Yeah. And then full system, I do it with a guide".P4

Perceived Usefulness

Improved independence

Both PwA and HCPs, as shown in (Textbox 11), thought that R-SPEAK could help improve independence.

Textbox 11. Perceived Usefulness: Enables independence and reduces reliance on others

"Well, Melanie (wife) has got she has me and her on the phone. She [...] regulates"(P4).
"He wants more freedom. He doesn't want me having to sit here or go up to the reception desk or up to the bar or order a meal, which he does do because you know, he wants to. But sometimes people are getting confused, There's a bit of embarrassment and stress levels are going up and how, you know, if you're out for a meal, you don't want to be. He wants to be able to do that himself. He's 62" (C1)

"If this was a way of helping [PwA] get a bit more independence of going out and trying to communicate with strangers. Maybe that's where it could come in handy" (SLT5)

Useful in many scenarios

Many ideas, as shown in (Textbox 12), were generated about possible scenarios where R-SPEAK could be used. These included: between patients and care teams on wards, in restaurants, at the doctor's surgery, during family conversations with young children, during phone calls or online

meetings and perhaps to support with writing emails or messaging. HCPs also suggested that R-SPEAK might be used therapeutically as a way to practice communicating.

Textbox 12. Perceived Usefulness: Useful when therapists unavailable and for home practice

"I know a lot of nursing staff sometimes struggle to understand what some with aphasia is saying, and it's usually the weekend when the speech therapist isn't there to try and interpret what it is they're trying to say. So it could be useful If it's helpful for them" (SLT5).

"I can imagine certain people sitting at home using it as a therapy tool, practising saying things, getting the feedback and choosing it and repeating after it, which is also another really useful use for it" (SLT3).

Table 3. Themes from interviews and focus group mapped onto TAM constructs

TAM Construct	Themes	Subthemes
Attitude towards using		
	High hopes vs Scepticism or somewhere in between	PwA and carers had high hopes
		Clinicians more sceptical
		Cautious optimism
	False assumptions about using AI	
	Cost could be prohibitive	
Perceived ease of use		
	Dependent on type and severity of aphasia	
	Other neurological impairments may pose challenges	
	Tech easy for some	
	Try before you buy	
	Some training needed	
Perceived usefulness		
	Improved independence	
	Could be useful in many scenarios	

System design features

PwA and HCPs liked that it was designed for use on a mobile phone and they liked the interface due to its simplicity, font size, use of colour, alignment and spacing of features.

In terms of functionality, interviewees wanted to improve the accuracy and speed of the responses. Both groups made suggestions for improved functionality and suggested modifications including spoken responses for those that have difficulty reading, editable responses, option to re-record answers or give two options to choose from. Table 4 summarises the combined feedback on system design features.

System Usability Scale

Ratings on the system usability scale can be found in (Figure 6) and (Figure 7). Individual SUS scores ranged between 62.5 and 85 with a mean group score of 75 which overall demonstrates acceptability.

Table 4. Combined feedback and suggestions for improvements on the R-SPEAK's design features

System design features	Positives	Negatives	Suggestions for modifications or improvements
Interface	Clear Simple Size/type of font Use of colour Alignment of features On a mobile phone	No ability to modify	Include modifiable set up for: Language/accents Font size Bolding key words in text Colour of background/text

Functionality	Accuracy sometimes good	Slow Accuracy sometimes poor	Improve accuracy of output Improve speed Include ability to hear response or read response Editable responses Re-record option Recognise intonation of speaker e.g. statement vs question. Present graphics instead of text output Learn from user to improve output Use information from the internet to improve output Available on all devices

Figure 6. System Usability Scale (SUS) - individual participant scores.

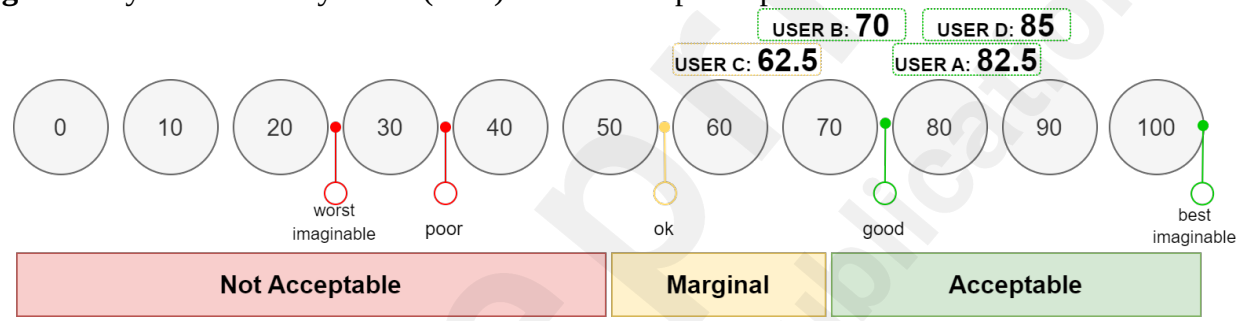
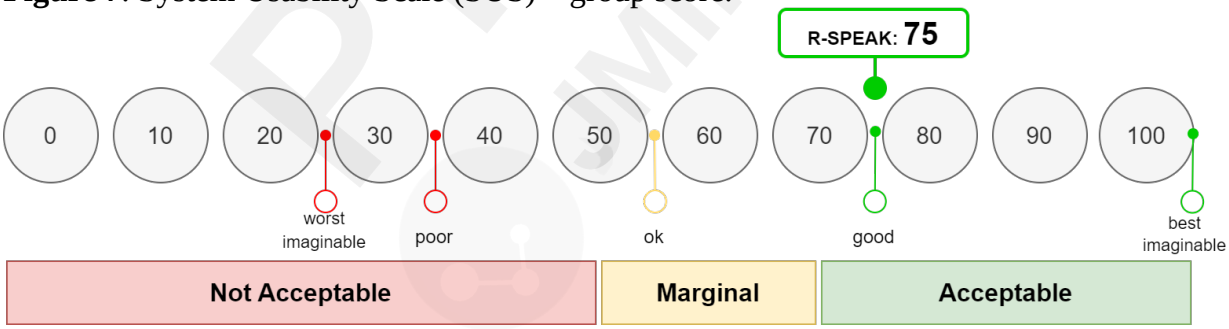


Figure 7. System Usability Scale (SUS) – group score.



Refine and redesign

To improve the speed of processing in converting input to output we conducted further tests with open-weight lightweight LLMs. We aimed to reduce response time whilst preserving response relevancy, faithfulness and completeness.

We evaluated 12 open-weight lightweight LLMs, ranging from 0.5b to 3.8b parameters, across 30 the clinician–patient dialogue samples from individuals with post-stroke aphasia. Each model generated interpretations under 10 stochastic decoding iterations (temperature = 0.7), yielding 3,600 total responses. Performance was assessed along three clinically grounded dimensions: answer relevancy

(alignment with the original question), faithfulness (absence of hallucinations relative to the aphasic utterance), and completeness (coverage of key intent elements present in expert-curated reference interpretations). An external LLM-as-a-judge system (Qwen3:8b) scored each output on a [0,1] scale, with a pass threshold of 0.5 per metric. Inference latency (in seconds) was recorded per generation to capture practical deployability.

Overall Performance and Model Rankings

The Qwen2.5:3b model emerged as the top performer across all metrics, achieving an overall mean score of 0.814, as shown in (Table 5). It attained 0.794 in relevancy, 0.866 in faithfulness, and 0.781 in completeness, with 100% pass rates across all samples and iterations. Hermes3:3b ranked second overall (0.796), excelling in answer relevancy (0.802, the highest among all models), though slightly trailing in faithfulness (0.826). Gemma3:4b placed third (0.792), demonstrating strong faithfulness (0.837) but showing higher variance (σ = 0.184), likely due to sensitivity to iterative stochasticity.

Table 5. Performance comparison of tiny and small LLMs across key metrics

Model	Answer Relevancy	Faithfulness	Completeness	Mean Score	Response Time (s)
Qwen2.5:3b	0.794	**0.866**	**0.781**	**0.814**	0.600
Hermes3:3b	**0.802**	0.826	0.761	0.796	0.895
Gemma3:4b	0.787	0.837	0.753	0.792	0.738
Phi4-Mini:3.8b	0.779	0.796	0.741	0.772	0.815
Llama3.2:3b	0.777	0.777	0.743	0.766	0.704
Phi3:3.8b	0.776	0.761	0.735	0.757	0.800
Qwen2.5:1.5b	0.741	0.775	0.745	0.753	0.436
Gemma3:1b	0.679	0.806	0.753	0.746	0.451
Llama3.2:1b	0.670	0.772	0.707	0.716	0.487
SmolLM2:1.7b	0.700	0.759	0.678	0.712	**0.425**
DeepScaler:1.5b	0.607	0.712	0.668	0.662	5.162
Qwen2.5:0.5b	0.573	0.566	0.503	0.547	0.720

At the lower end, Qwen2.5:0.5b struggled across all dimensions (mean = 0.547), with particularly poor faithfulness (0.566) and completeness (0.503), suggesting that sub-1B architectures lack sufficient capacity to reliably disambiguate fragmented aphasic input. DeepScaler:1.5B, while moderate in quality (0.662), exhibited prohibitively high latency (mean = 5.16s), rendering it unsuitable for real-time assistive applications.

To contextualise the performance of lightweight models, we compared them against the top three performing larger LLMs (viz. mixtral:8x7b, gemma2:9b, and Qwen2:7b). While these models achieved marginally higher overall scores (0.838, 0.833, and 0.807, respectively), this quality improvement comes at a disproportionate computational cost that could be prohibitive for interactive clinical deployment. Crucially, the performance gain over the best lightweight model (Qwen2.5:3b, 0.814) is minimal (+0.024 for mixtral:8x7b), while inference latency increases dramatically from 0.60 seconds to 2.42 seconds, representing a 4.0x slowdown that would disrupt the conversation flow.

This speed-performance profile is particularly significant for real-time clinical interactions. In therapeutic settings, response latency directly impacts conversational flow and patient engagement. Typical conversational gaps last only a few hundred milliseconds; silences of a second or more feel

long and disrupt the flow. Individuals with communication disorders experience these prolonged pauses more often, increasing the effort needed to plan, produce, and process speech [31]. The Qwen2.5:3b model achieves near-optimal quality (within 3% of larger models) while maintaining sub-second response times essential for interactive applications. Even the fastest lightweight contender, SmolLM2:1.7b (0.43s), offers clinically adequate performance (0.712 on average) at 5.6x the speed of the best larger model.

Discussion

Principle Results

We co-designed a protype app using the Mixtral LLM to enhance the clarity and fluency of verbal expression of PwA. We tested the app with PwA and demonstrated it to HCPs. Both HCPs and PwA groups felt that R-SPEAK had the potential to be less stigmatised than other forms of AAC. As a mobile application, the software is more modern and can be integrated into the users existing devices compared to older AAC devices, which may lead to it being more socially accepted. Additionally, HCPs considered certain use-cases in which R-SPEAK would be useful, such as improving patient-doctor communication in clinical settings or even as a therapy tool.

The concept of practicing real life communication as an intervention isn't novel, speech therapy often involves role playing in different scenarios, and EVA Park [32] is an online virtual world that gives people with aphasia opportunity to practice communicating but both examples rely on therapists to facilitate. In another study PwA reported that the use of an app for communication practice would be highly beneficial for them [33]. R-SPEAK has the potential to be used in such a way. Both clinicians and PwA agreed that the potential use-cases for R-SPEAK could improve patients' independence, lessening the need for SLT or carer intervention. This would remove some of the burden for the constant presence of carers and allow for PwA to gain more independence and confidence when it comes to communicating with others.

An average SUS score given by the PwA of 75 indicates a good level of usability which is superior to other AAC apps for PwA and the app's interface was reported as easy to use. This implies R-SPEAK is acceptable to PwA. However, both groups expressed concern at the long response time of the prototype app. In the refine and redesign phase we found that Qwen2.5-3B LLM achieved the strongest overall performance with high faithfulness and sub-second latency when compared to 11 other models.

There was a degree of disparity between the attitudes of PwA and HCPs; PwA were positive and expressed optimism about its potential to improve communication, whereas clinicians were more cautious about R-SPEAK's capabilities to dramatically change the lives of PwA. HCPs doubted how well the tool would perform in actual conversation, where meaning can be highly contextual and implicative. Cost was also mentioned as a potential limitation to the software; charging PwA for a tool may deter use or even exacerbate the inequalities faced by PwA, particularly if they have idealistic expectations about R-SPEAK's capabilities. A 'try before you buy' feature was suggested to address this issue, to ensure that users are fully aware of the software's capabilities before committing to payment.

Furthermore, concerns were expressed about the AI-powered nature of the app, which may deter potential users from engaging with the software, however, this sentiment was not shared by the PwA. PwA having higher expectations about R-SPEAK's capabilities highlights a disparity between those with a specific lived experience of aphasia and those with purely clinical exposure. This conveys a potential struggle to align the expectations of the two groups. This observation highlights the crucial

nature of co-design in the current study for developing and evaluating clinical tools such as R-SPEAK. It is also essential to understand important aims and outcomes of R-SPEAK for both the patient and the clinician for robust and meaningful testing.

While the participants of the current study were able to use the software with ease, both groups had concerns about potential R-SPEAK users who were less technologically literate, had different types and severities of aphasia or other stroke related, speech, physical, cognitive or visual impairments. This might limit the usability of the software and prompt more sophisticated (and therefore costly) training or personalisation features. Therefore, incorporating a wider group of people with aphasia in further development and testing will be essential to ensure it is accessible to as many people as possible and so the target population could be refined. It was also suggested that HCP-provided training could be beneficial, although an in-app tutorial may be just as effective without the cost of having to formally train clinical staff.

Limitations

R-SPEAK was developed specifically for patients with mild-to-moderate expressive aphasia, reflected in the sample of five patients who were recruited for the co-design process. Given that the five patients were able to use the technology with ease, it might be that R-SPEAK could be adapted to suit patients with a wider range of needs, such as more severe cases of aphasia. However, four of the five patients had moderate aphasia and a decent level of technological literacy. This means that the positive attitudes and feedback given by the PwA may not be fully generalisable to those who have a milder or more severe diagnosis of aphasia, or to those who are less technologically literate. In addition, during the testing, PwA were supported to use a web-based version of the app which likely reduced the cognitive and linguistic demands on the participants and might mean feedback was overly optimistic [33]. This was a proof-of-concept study; thus participant numbers were intentionally small, but this is a limited representation of the target population and further testing on a wider group is imperative.

Another limitation stems from the initial testing of the LLMs, which involved evaluating their interpretation of utterances collected from AphasiaBank. The ‘gold standard’ interpretations were decided upon by members of the team; however, it is difficult to confirm that these are in fact the true interpretations of the utterances, because they were collected from unknown speakers out of the context of conversation. Moreover, the LLM used in the development of R-SPEAK was still prone to occasional underperformance, despite being rated the best LLM. Mixtral sometimes produced inaccurate interpretations of the patient’s utterance and overly wordy responses, which could lead to unproductive or confusing conversation turns. Further work in improving accuracy in outputs as misalignment between intended message and output reduces “trust” in AI.

Comparison to prior work

R-SPEAK is a leap away from existing AAC tools, which require pointing to pictures or writing key words to express basic needs. A US group [33] are co-developing AI tools to support PwA in real-time communication and to prepare for future conversations. They used OpenAI’s DALL·E 3, ChatGPT 3.5 Turbo, and Whisper to build the tools. Whilst it is not a complete product, the systems they are developing might be useful to consider as features in R-SPEAK such as the double checking of important words, using key words to generate sentences and use of a diary to capture meaningful experiences. They are targeting a greater range of types and severities of aphasia including those with comprehension difficulties and they have included a range of participants in their prototype development. Participants reported the tools reduced ambiguity and ability to recount personal experiences with less effort, but trust in the AI tools reduced due to technical limitations of AI.

Suggestions for future work

The next step with R-SPEAK, is to complete Phase 3 of the design process, the 'Refine and Redesign' phase, in which design improvements are made based on feedback. Future testing may involve a larger sample of patients with a broader range of aphasia, confidence in technology and other stroke related impairments. This will ensure that the technology is usable by the entire target population, but also to observe whether patients outside of the initial target population might also benefit from R-SPEAK. Additionally, future testing may want to examine the software's performance in real-world situations. This would involve testing the software in a specific context with an interlocutor, such as communicating patient needs in a hospital to a nurse.

Further work should also be done to improve the performance of the LLM technology for its application to R-SPEAK. Mixtral was found to underperform occasionally due to inaccuracy and its tendency to produce verbose outputs. Verbosity is a tendency of LLMs that have been trained on written text, since spoken language is nearly always more concise than written language. Future prototypes of R-SPEAK could leverage LLM technology by fine-tuning the model on speech data or using a model that has been pre-trained on speech. This would prime the technology to produce more speech-like responses and therefore improve its performance.

The adaptable nature of LLM technology could be leveraged further to improve the user's experience with R-SPEAK by offering features to tailor the software to the individual. For instance, the LLM model could be fine-tuned on the individual patient's communication preferences, which would allow the model to learn their linguistic behaviour and produce more accurate and personal responses. As mentioned by the PwA and HCPs in the focus groups, different patients may have specific difficulties; for example, they may struggle with reading comprehension or have motor problems. Implementing R-SPEAK as an experience tailored to the individual's needs could both improve the general usability of the software and broaden R-SPEAK's target population. For example, for those with motor impairments, a fully voice-controlled system might be more usable and therefore more beneficial. To develop and implement these features properly, further iterations of co-design and testing would be crucial to ensure that the software remained helpful and usable.

Conclusions

Overall, the current study proposes R-SPEAK, a promising piece of software that has the potential to be a more beneficial and less stigmatised alternative to classic AAC tools. Initial testing revealed clinicians' cautious optimism towards R-SPEAK's impact and its generalisability, but PwA were enthusiastic and had high hopes for improving their lives. Cost and usability were main concerns; however, opportunities for improvement such as model enhancements, personalisation features, and training materials pose opportunities to make R-SPEAK better-performing and more accessible to a wider range of patients.

Acknowledgments

Qwen2.5 and Grammarly was used to check grammar and sentence structure. For the purpose of open access, the authors have applied a Creative Commons Attribution (CC BY) licence to any Authors Accepted Manuscript version of this paper arising from this submission.

Funding

This work was supported by the Engineering and Physical Sciences Research Council (EPSRC) under Project Reference EP/W000679/1, URL: (https://www.rehabtechnologies.net/fundedprojects).

Conflicts of Interest

none declared.

Data Availability

The datasets generated or analysed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

AA: Conceptualisation; Methodology; Software; Validation; Data Curation; Writing – Original Draft; Writing – Review & Editing; Visualization; Project Administration; Funding Acquisition.
JA: Conceptualization; Methodology; Investigation.
JB: Methodology; Validation; Data Curation; Formal Analysis; Writing – Original Draft; Writing – Review & Editing; Supervision; Project Administration; Funding Acquisition.
CS: Methodology; Investigation; Writing – Review & Editing.
CG: Writing – Original Draft; Funding Acquisition; Validation.
RL: Methodology; Investigation; Resources.
DT: Formal Analysis; Writing – Original Draft.
MD: Data Curation; Software; Validation.
DW: Writing – Original Draft; Writing – Review & Editing.
KR: Methodology; Formal Analysis; Resources; Writing – Original Draft; Writing – Review & Editing; Supervision; Project Administration; Funding Acquisition.

Abbreviations

BCW: Behaviour Change Wheel
COREQ: Consolidated Criteria for Reporting Qualitative Research
HCP: Health Care Professional
JMIR: Journal of Medical Internet Research
PT: Physiotherapist
PwA: People with aphasia
RCT: Randomized Controlled Trial
R-SPEAK: Revolutionising Speech Enhancement in Aphasia Using Knowledgeable-AI
SLT: Speech and Language Therapist
SUS: System Usability Scale
TAM: Technology Acceptance Model

References

1. Mitchell C, Gittins M, Tyson S, Vail A, Conroy P, Paley L, et al. Prevalence of aphasia and dysarthria among inpatient stroke survivors: describing the population, therapy provision and outcomes on discharge. Aphasiology. 2021;35(7):950-60. doi: 10.1080/02687038.2020.1759772.
2. Whitworth A, Webster J, Howard D. A cognitive neuropsychological approach to assessment and intervention in aphasia : a clinician's guide. Hove: Psychology Press; 2005. ISBN: 9786610175048.
3. Wisenburn B, Mahoney K. A meta-analysis of word-finding treatments for aphasia. Aphasiology. 2009;23(11):1338-52. doi: 10.1080/02687030902732745.
4. Fama ME, Lemonds E, Levinson G. The Subjective Experience of Word-Finding Difficulties in People With Aphasia: A Thematic Analysis of Interview Data. American journal of speech-language pathology. 2022;31(1):3-11. doi: 10.1044/2021_AJSLP-20-

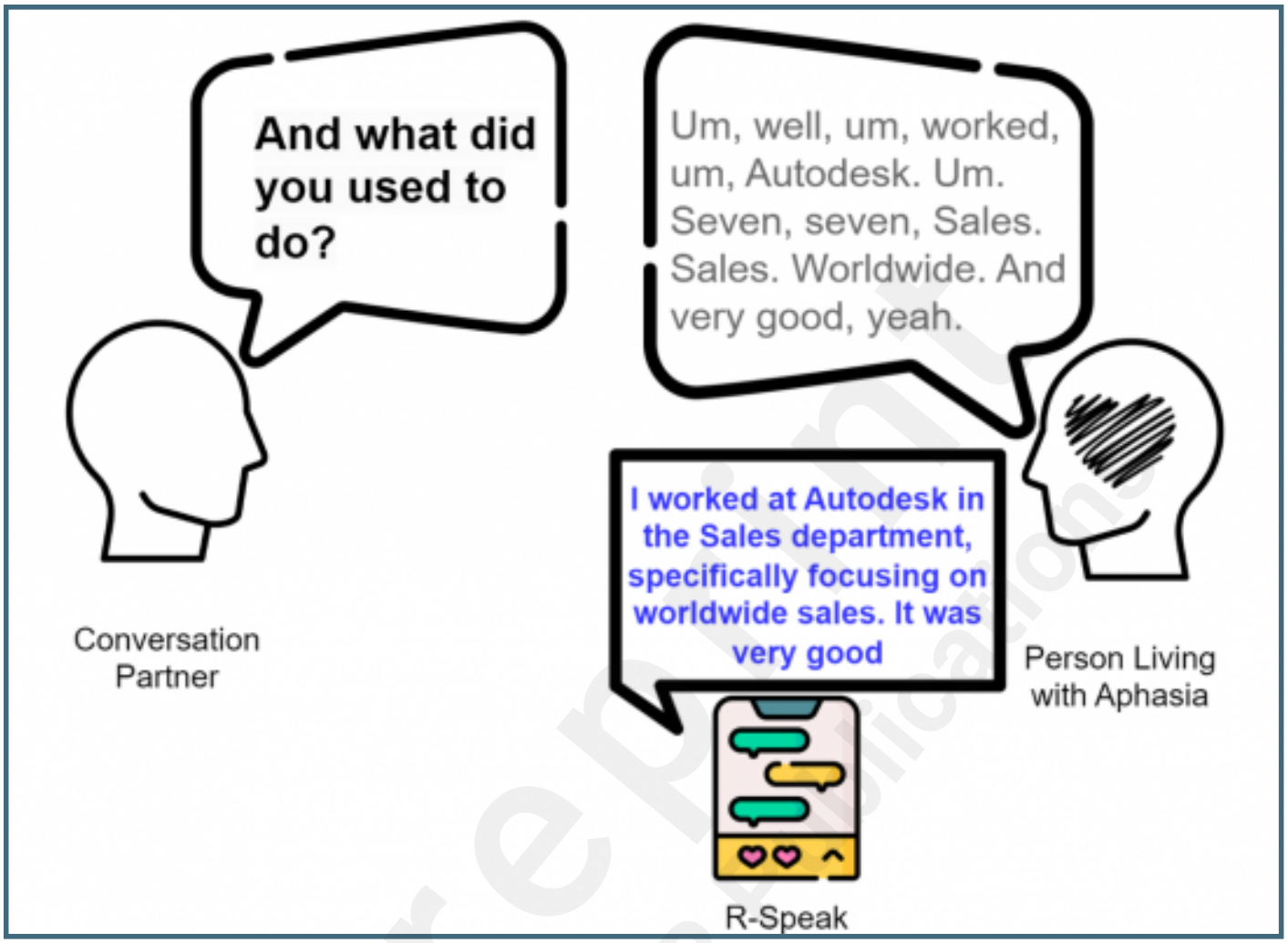
00265.
5. Paolucci S, Matano A, Bragoni M, Coiro P, De Angelis D, Fusco FR, et al. Rehabilitation of left brain-damaged ischemic stroke patients: the role of comprehension language deficits. A matched comparison. Cerebrovascular diseases (Basel, Switzerland). 2005;20(5):400. doi: 10.1159/000088671.
6. Hilari K, Northcott S. "Struggling to stay connected": comparing the social relationships of healthy older people and people with stroke and aphasia. Aphasiology. 2017;31(6):674-87. doi: 10.1080/02687038.2016.1218436.
7. Wray F, Clarke D. Longer-term needs of stroke survivors with communication difficulties living in the community: a systematic review and thematic synthesis of qualitative studies. BMJ open. 2017;7(10):e017944. doi: 10.1136/bmjopen-2017-017944.
8. Dietz A, Wallace SE, Weissling K. Revisiting the Role of Augmentative and Alternative Communication in Aphasia Rehabilitation. American journal of speech-language pathology. 2020;29(2):909-13. doi: 10.1044/2019_AJSLP-19-00041.
9. Bloch S, Beeke S. Co-constructed talk in the conversations of people with dysarthria and aphasia. Clinical linguistics & phonetics. 2008;22(12):974-90. doi: 10.1080/02699200802394831.
10. Brown E, Cairns P. A grounded investigation of game immersion. Extended Abstracts on Human Factors in Computing Systems New York: NY: Association for Computing Machinery.; 2004. p. pp. 1297–300.
11. Yu D, Deng L. Automatic speech recognition: A deep learning approach: Springer; 2015.
12. Mhasakar M, Sharma S, Mehra A, Venaik U, Singhal U, Kumar D, et al. Comuniqa : Exploring Large Language Models for improving speaking skills. Ithaca: Cornell University Library, arXiv.org; 2024.
13. van den Oord A, Dieleman S, Zen H, Simonyan K, Vinyals O, Graves A, et al. WaveNet: A Generative Model for Raw Audio. Ithaca: Cornell University Library, arXiv.org; 2016.
14. Arik SO, Chen J, Peng K, Wei P, Zhou Y. Neural Voice Cloning with a Few Samples. Ithaca: Cornell University Library, arXiv.org; 2018.
15. Yang Z, Ishay A, Lee J, editors. Coupling large language models with logic programming for robust and general reasoning from text . Findings of the Association for Computational Linguistics; 2023.
16. Vatsal S, Dubey H. A survey of prompt engineering methods in large language models for different NLP tasks. . arXiv Preprint [Internet]. 2024.
17. Thiel L, Conroy P. 'I think writing is everything': An exploration of the writing experiences of people with aphasia. International journal of language & communication disorders. 2022;57(6):1381-98. doi: 10.1111/1460-6984.12762.
18. Tobin J, Nelson P, MacDonald B, Heywood R, Cave R, Seaver K, et al. Automatic Speech Recognition of Conversational Speech in Individuals With Disordered Speech. Journal of speech, language, and hearing research. 2024;67(11):4176-85. doi: 10.1044/2024_JSLHR-24-00045.
19. Vinotha R, Hepsiba D, Vijay Anand LD, Andrew J, Jennifer Eunice R. Enhancing dysarthric speech recognition through SepFormer and hierarchical attention network models with multistage transfer learning. Scientific reports. 2024;14(1):29455-23. doi: 10.1038/s41598-024-80764-w.
20. Hrivnáková D, Václavová E, Laco M, editors. Modified double diamond design methodology for innovative interfaces in the medical domain. . Proceedings of the Future Technologies Conference (FTC) 2025; 2026: Springer.
21. DSDM Consortium. DSDM: Business Focused Development. DSDM Consortium, editor. https://www.agilebusiness.org/resource/ebook-hb-agileframework.html2001.
22. Aphasia Bank. The Talk Bank System. [database on the Internet]. 2017.
23. Davis FD. Perceived Usefulness, Perceived Ease of Use, and User Acceptance of

Information Technology. MIS quarterly. 1989;13(3):319-40. doi: 10.2307/249008.
24. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. International journal for quality in health care. 2007;19(6):349-57. doi: 10.1093/intqhc/mzm042.
25. Marikyan D, Papagiannidis S. Technology Acceptance Model: A review. In: Papagiannidis S, editor. TheoryHub Book. https://open.ncl.ac.uk2024.
26. Green J. Qualitative methods for health research / Judith Green & Nicki Thorogood. 4th ed. ed. Thorogood N, editor. London: London : SAGE; 2018.
27. Lewis JR. The System Usability Scale: Past, Present, and Future. International journal of human-computer interaction. 2018;34(7):577-90. doi: 10.1080/10447318.2018.1455307.
28. Simmons-Mackie N, Kagan A, Shumway E. Aphasia Severity Rating. 2018.
29. Braun V, Clarke V. Using thematic analysis in psychology. Qualitative Research in Psychology. 2006;3(2):77-101. doi: 10.1191/1478088706qp063oa.
30. Braun V, Clarke V. One size fits all? What counts as quality practice in (reflexive) thematic analysis? Qualitative research in psychology. 2021;18(3):328-52. doi: 10.1080/14780887.2020.1769238.
31. Thomas BK. Quantifying speech pause durations in speakers with nonfluent and fluent aphasia. https://scholarsarchive.byu.edu/etd/8939 Brigham Young University); 2021.
32. Marshall J, Devane N, Talbot R, Caute A, Cruice M, Hilari K, et al. A randomised trial of social support group intervention for people with aphasia: A Novel application of virtual reality. PloS one. 2020;15(9):e0239715. doi: 10.1371/journal.pone.0239715.
33. Mao L, Lee JH, Faroqi-Shah Y, Valencia S, Oakley I, Nisi V, et al., editors. Design Probes for AI-Driven AAC: Addressing Complex Communication Needs in Aphasia. Proceedings of the 2025 ACM Designing Interactive Systems Conference; 2025 2025; New York, NY, USA: ACM.

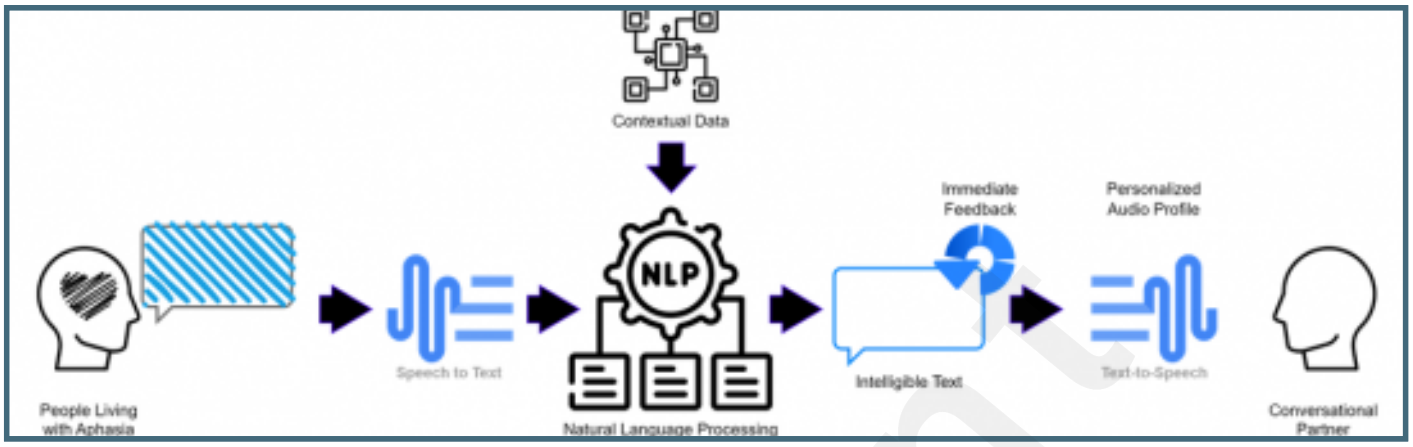
Supplementary Files

Figures

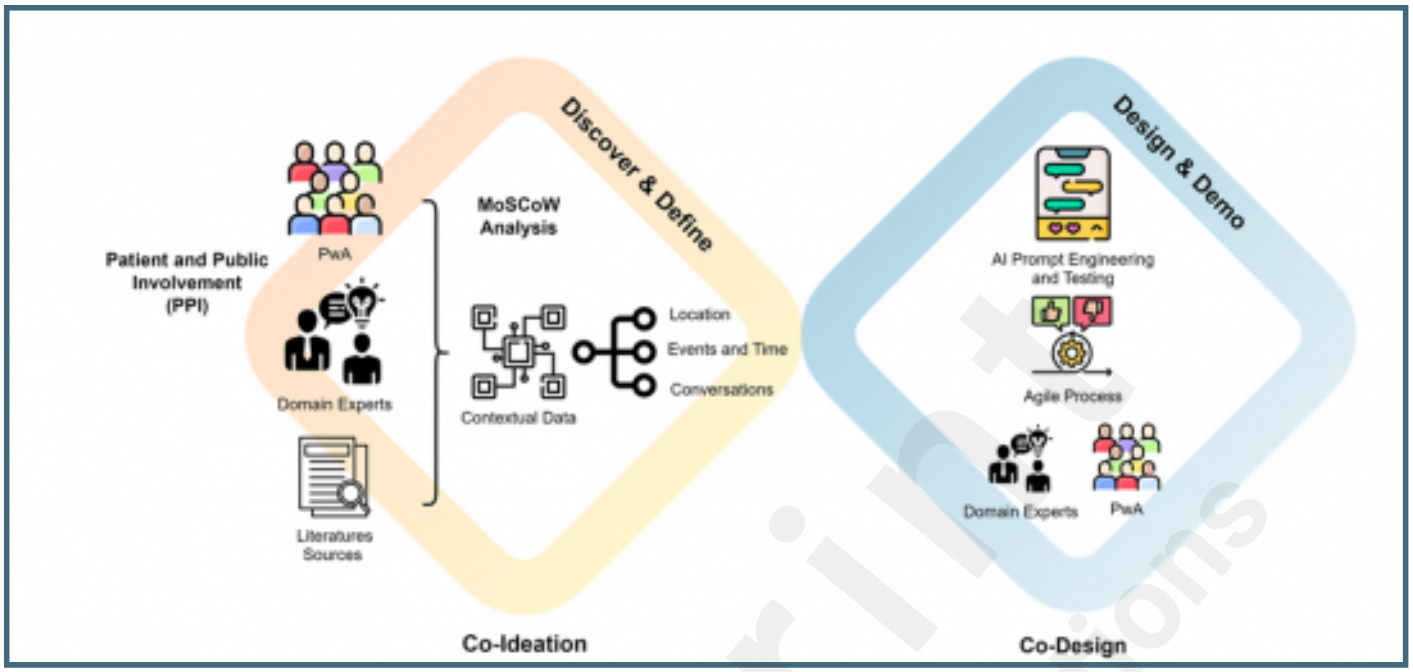
Example of how R-SPEAK converts real-world aphasic speech into comprehensible speech.



Overview of the R-SPEAK system workflow for enhancing communication in people with aphasia (PwA).



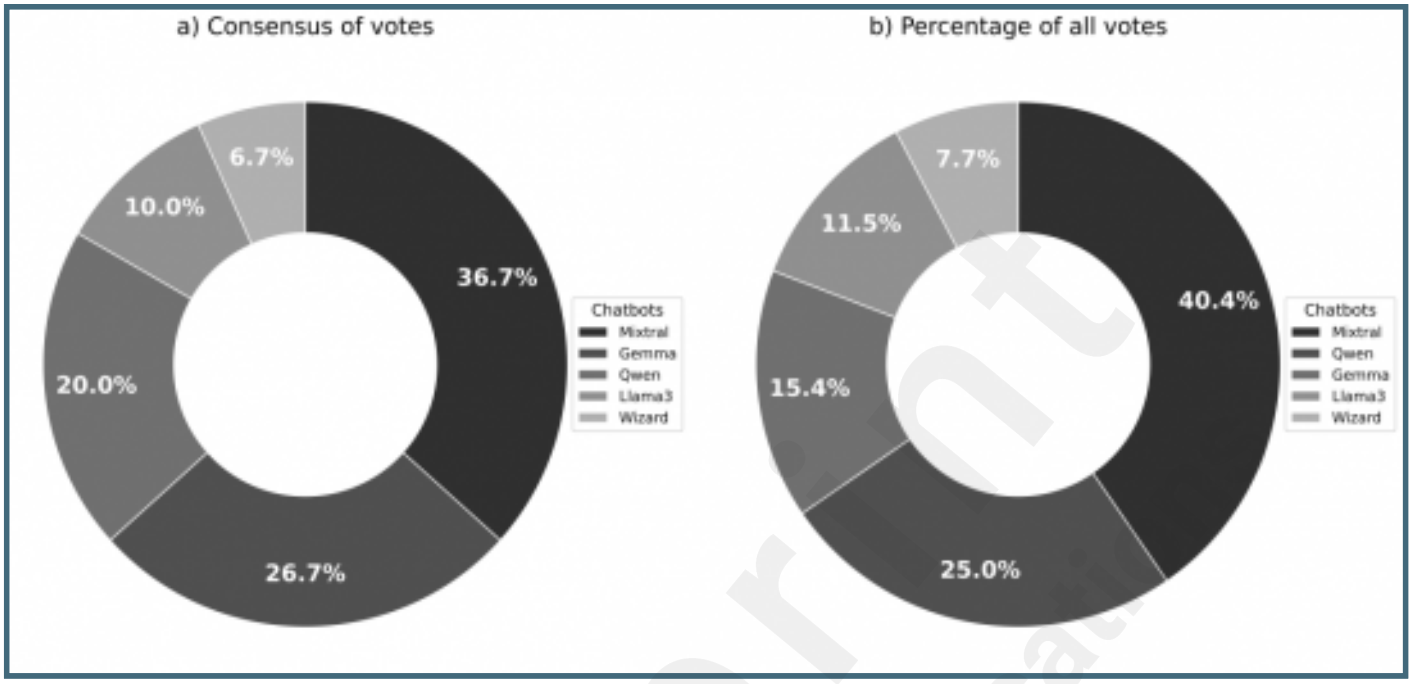
Double Diamond approach to prototype development.



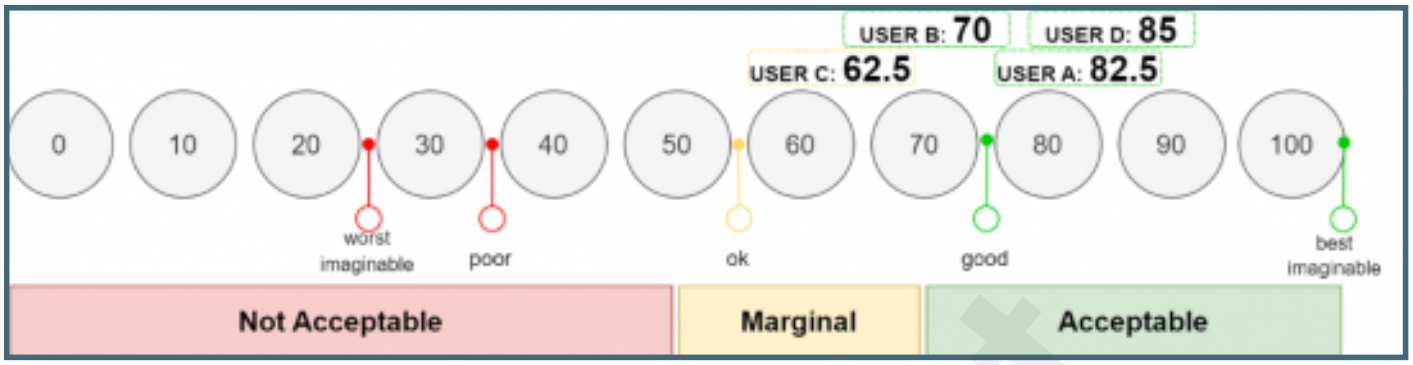
Aphasia Severity Rating (ASR) Scale. This scale provides a structured index of aphasia severity, ranging from complete language impairment (0) to minimal or undetectable difficulties (4), based on speech, writing, and comprehension abilities. Adapted from [28].

Score	Aphasia Severity Rating (ASR)
0	Speech, writing and/or auditory comprehension are not functional. Any attempts to speak or to use different utterances are not understandable to the listener OR the individual is not attempting to speak at all.
1	The individual may occasionally produce words or phrases that are meaningful in context, but communication is fragmentary and not possible without significant support from the listener or augmentative communication tools. The effort to communicate is often described as an enormous effort with very little information conveyed and a sense of a burden. An extremely limited amount of message may be attempted to be exchanged. Misunderstandings or failed communications are very frequent.
2	Basic conversation about familiar and everyday topics is possible but significant breakdown does occur with more complex or difficult conversations. The listener can often understand the intent of the message and assist with repair. The speaker is able to communicate some of the time, but many misunderstandings by the listener require frequent need for repair.
3	Despite some observable issues related to speech fluency or comprehension, there is no significant limitation. The individual may hesitate or look for words or need some effort to find utterances. The listener has a problem with language, which is not severe enough to interfere with the successful exchange of ideas or the use of communication with the listener.
4	Although the individual feels that he/she has a problem with language, this is rarely apparent to the listener who may not detect any problem with speaking or understanding.

Four stroke survivors with aphasia were involved in prototype testing and interviews. Three were assessed as having moderate aphasia with one having milder impairment on the Aphasia Severity Rating scale. Characteristics of participants can be found in (Table 1).



System Usability Scale (SUS) - individual participant scores.



System Usability Scale (SUS) – group score.