"The less I type, the better": How AI Language Models can Enhance or Impede Communication for AAC Users

Stephanie Valencia
Carnegie Mellon University, Google Research
Pittsburgh, PA, United States
svalenci@andrew.cmu.edu

Richard Cave
UCL Department of Language and Cognition
London, United Kingdom

Krystal Kallarackal
Google Research
Cambridge, MA, United States

Katie Seaver
MGH Institute of Health Professions
Boston, MA, United States

Michael Terry
Google Research
Cambridge, MA, United States

Shaun K. Kane
Google Research
Boulder, CO, United States
shaunkane@google.com

ABSTRACT

Users of augmentative and alternative communication (AAC) devices sometimes find it difficult to communicate in real time with others due to the time it takes to compose messages. AI technologies such as large language models (LLMs) provide an opportunity to support AAC users by improving the quality and variety of text suggestions. However, these technologies may fundamentally change how users interact with AAC devices as users transition from typing their own phrases to prompting and selecting AI-generated phrases. We conducted a study in which 12 AAC users tested live suggestions from a language model across three usage scenarios: extending short replies, answering biographical questions, and requesting assistance. Our study participants believed that AI-generated phrases could save time, physical and cognitive effort when communicating, but felt it was important that these phrases reflect their own communication style and preferences. This work identifies opportunities and challenges for future AI-enhanced AAC devices.

KEYWORDS

accessibility, communication, artificial intelligence, large language models

ACM Reference Format:
Stephanie Valencia, Richard Cave, Krystal Kallarackal, Katie Seaver, Michael Terry, and Shaun K. Kane. 2023. "The less I type, the better": How AI Language Models can Enhance or Impede Communication for AAC Users . In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23), April 23–28, 2023, Hamburg, Germany*. ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3544548.3581560

1 INTRODUCTION

Individuals who are unable to speak using their voice, or who have difficulty speaking, often rely on an augmentative and alternative communication (AAC) device to assist with communication. People with a wide variety of abilities and disabilities may use AAC devices to assist with communication. Given the variety of AAC users, AAC devices themselves often vary in their use: users may interact with the device through gaze, touch, or with a physical switch; users may select words through menus, by typing them with a keyboard, or with some combination of the two; AAC output may be read aloud, shared as text, or stored for later use [2].

While AAC users and devices may vary, there are some general challenges that affect many AAC users. For example, AAC users often communicate more slowly than non-AAC users [30]. As a result, they may feel pressure to respond in time or struggle to participate in a conversation. Some AAC users report that using an AAC device requires high physical and cognitive effort, which impacts AAC users' ability to effectively express themselves [16, 20]. Much research around AAC focuses on the goals of reducing the effort of AAC input and increasing the speed of AAC composition.

A primary strategy for improving AAC performance is to predict what the user intends to type and offer it as a suggestion [34]. These predictions can come from many sources, including static language models [34], photographs [12, 13], or contextual information about the user [18, 19]. AAC users themselves may attempt to predict what they will discuss in the future and pre-write messages that they can later retrieve via their AAC device [20].

Recently, advances in large, neural language models (LLMs) such as GPT-3 [6] and BERT [10] have created new opportunities for improving the usability and efficiency of AAC devices. Current LLMs are able to generate text that is indistinguishable from text written by a human [8], potentially enabling AAC users to generate human-level speech with minimal effort. Preliminary research with simulated user data has shown that LLMs can retrieve contextually relevant sentences [31] and expand user abbreviations [7], theoretically reducing an AAC user's keystrokes by up to 75%.

While these potential gains are encouraging (and in fact will likely continue to improve), it is important that AAC users be involved in the process of combining LLMs with AAC devices. This involvement helps ensure that LLM output meets the users' expectations, and that interactions between LLMs and AAC users support the users' communication preferences, all while maintaining privacy, autonomy, and control.

In this paper, we present a study with 12 adult AAC users in which participants generated speech suggestions from an LLM and

CHI '23, April 23–28, 2023, Hamburg, Germany

ACM ISBN 978-1-4503-9421-5/23/04.
https://doi.org/10.1145/3544548.3581560

provided feedback about those suggestions. Our participants had a variety of disabilities that affected their speech production but not their language use or understanding. To support participants' experimentation with the LLM, we introduce the concept of *speech macros*: LLM prompts that transform abbreviated user input into full sentences, with a focus on achieving conversational goals such as requesting help with something or answering a biographical question. Participants tested each of these macros over a remote video call, trying various inputs and commenting on the outputs, and later provided feedback about their experience via an online questionnaire. Our study was guided by the following research questions:

- *RQ1.* What are the benefits for AAC users, if any, of directly interacting with large language models (LLMs)?
- *RQ2.* How do AAC users evaluate communication suggestions made by an existing LLM?
- *RQ3.* What concerns do AAC users have about integrating LLMs into their own AAC devices?

Overall, our participants were excited about the possibility of using AI-generated suggestions in their AAC device, but articulated some requirements for these suggestions to be usable. Specifically they requested that these suggestions are contextually appropriate, match the user's personal conversation style, and provide the ability to customize, edit, and remove suggestions.

This paper makes several contributions toward the goal of enabling AAC users to benefit from the capabilities of LLMs. First, we present a study in which AAC users interacted with an LLM in real time and provided feedback about the suitability of the LLM's output for their own communication needs. Second, we introduce the concept of *speech macros* as a way to leverage the generative capabilities of LLMs to support the specific communication needs of AAC users. Finally, we identify opportunities and challenges to creating AI-powered AAC systems.

2 RELATED WORK

Our work builds on prior research on the use of context to improve AAC systems, the exploration of AI as a design material, and the use of large language models (LLMs) and natural language prompting.

2.1 AAC and Social Interaction

When communicating using an AAC device, people with speech disabilities have to first compose a message by typing out individual characters or selecting predicted words or icons before they share their thoughts. Speaking using an AAC device usually takes longer due to message composition delays that lead to speaking rates of 3 to 20 words per minute (WPM) while non-AAC users communicate at higher rates (100-140 WPM) [30]. This time asymmetry can cause some social interaction challenges between augmented and non-augmented communicators. For example, augmented communicators (ACs) have reported feeling pressure to respond in time and having others not wait long for their answers [35]. AAC users who lost their ability to speak later in life have reported loosing a lot of their expressivity like their ability to joke or express sarcasm once they became AAC users [20]. AAC users often have to decide how much they want to say and consider both the physical effort and the amount of time they will require before they decide if they want to compose their message [20]. It is then often the case that AAC users participate much less than their non-AAC user counterparts in conversation, are left behind in group conversation, or struggle to demonstrate the relevance of their comment that is shared some minutes delayed after their topic they are addressing has passed [29, 30, 35].

2.2 Improving AAC Performance

As noted above, a significant challenge experienced by many AAC users is a much slower rate of communication. Several approaches have been explored for improving the text entry rate for keyboard-based AAC. Most commonly, these systems adopt a linguistic model to predict the next word or words that the user intends to type. In ideal cases, word prediction has been shown to improve typing speed by more than 50 percent [34]. A second approach is to adapt user input methods to support more efficient input, such as by dynamically adjusting the dwell time needed to select keys using eye gaze [27] or supporting dwell-free, gesture-based typing [22]. Both word prediction (including user-specific word and phrases) and dwell-free typing are available in commercial AAC devices, such as the Tobii Dynavox Communicator 5[1].

These methods assume that users will type out their intended message letter by letter, selecting word suggestions when possible. AAC researchers have also explored alternative methods for composing messages, such as typing in the initial letter of each word [7] or by using a phoneme-based keyboard [33]. These systems assume that users will compose their messages and then convert them into an abbreviated form; in contrast, our system explores how users can provide input suggestions that are expanded into phrases by the AI language model.

2.3 Context Awareness in AAC

In addition to improving performance using linguistic predictions, context-aware AAC solutions have been proposed to leverage context sources such as location [18], conversation partner word completions [11], and more recently the use of large-language models that utilize additional context such as a user's persona, keywords, and conversational context to suggest relevant content to an AAC user [31]. Using general dialogue data, some studies have shown substantial improvements in word prediction when language models consider the partner's speech in their prediction [31, 36]. Recent advances in large language models point to a great opportunity to leverage LLMs to support AAC communication. Nonetheless, we do not know how, in practice, AAC users could benefit from these enhanced uses of language models and how they would like to interact with the models (give information, make selections in an interface) and how they would use them. In this study, we showed design concepts that showcased a system using partner speech to have conversations with participants about this idea, its usefulness and concerns around it.

2.4 Interacting with Large Language Models

Large Language Models (LLMs) are machine learning algorithms that can recognize, generate, and transform human languages by having learned patterns from large text-based data sets. Recent

[1] https://us.tobiidynavox.com/pages/communicator-5-ap

LLMs such as GPT-3 [5] have proven to learn from examples or text-based instructions known as prompts [17] to generate language in context. By carefully crafting the inputs given to the model people can directly influence the output of a LLM by defining a desired task. LLMs have previously been used to support accessibility use cases including generating speech for AAC users [7, 31], and providing writing support for people with dyslexia [14].

This capability of understanding language to produce language makes LLMs a great resource that could enable AAC users to produce responses that are detailed and grammatically correct by only inputting a few words. Given a specific text-based instruction or example also known as prompt, LLMs can return plausible continuation or response to the given prompt. For example, *prompt:* give me some fruits that start with "A", *model:* Apple, Apricots, Ananas. There are many advantages to using prompts to retrieve customized output from a language model. Prompting does not require pre-training or fine-tuning of a model which can be expensive and require access to large amounts of data, which is a limitation in the AAC research field. Utilizing prompts as a prototyping tool can enable quick explorations on the types of input and outputs needed for the model to be most useful [17]. In this work, we explore how prompting could support AAC users in their communication by presenting them with different pre-made prompts that can be configured at different levels: the type of context it uses, the type of input it requires from the user, and the task at hand (e.g., add details to a reply, share background information, turn words into requests).

2.5 Designing with AI

Designing with technologies that add an intelligence layer to products or systems is a non-trivial task [39]. Traditional user-centered methods as paper prototyping may fall short to communicate to users what functionalities and errors are possible when interacting with the system. At the same time it is important to show design concepts to users that are feasible and realizable to elicit feedback and encourage discussion and reflection that can uncover design opportunities as well as ethical risks and concerns. One approach to communicating technology capabilities and limitations to users, designers, and technologists has been the use of boundary objects [40], representations of abstract ideas in the form of interactive prototypes or other artifacts that facilitate communication among different stakeholders [23]. For example, boundary objects in the form of interactive notebooks have been used to discuss how language intelligence could support creative writers [38]. Other approaches have developed open-ended working prototypes to explore what tasks an AI system can and cannot do for specific users [26, 28]. In this work we developed interactive prototypes that served as boundary objects to communicate LLMs functionalities to AAC users and encourage discussion and reflection about integrating LLMs into AAC.

3 INTEGRATING LLMS INTO AAC WITH SPEECH MACROS

In considering how LLMs can provide useful suggestions to AAC users, we landed upon the concept of *speech macros* as a way to explore scenarios in which LLMs generate content for AAC users. Our approach is similar to KwickChat's *bag-of-keywords* model [31], in that an AAC user provides one or few input words that are then converted into a complete sentence. However, in contrast to prior work, our speech macros go beyond sentence expansion to support a variety of connections between input and output.

3.1 Design Process

We began this project with an exploration of how LLMs could support use cases common to AAC users, and how we might explore those benefits in the context of a user study. We conducted several brainstorming and sketching sessions within our research team, which contains HCI/accessibility researchers, researchers with experience related to LLMs, and speech language pathologists.

Through this process, we identified a set of potential benefits that LLMs can provide to AAC users, including some that have been explored in prior work:

(1) ability to create full sentences from abbreviated input (as explored in [7, 31]);
(2) ability to draw from conversational or user context (also explored in [7, 31]);
(3) ability to generate grammatically correct sentences in response to a question;
(4) ability to customize the tone and content of output.

While our prototype includes elements of all of these benefits, we ultimately decided to focus on how LLMs can be instructed to perform a variety of tasks using natural language prompts. For example, prompts provided to a general purpose LLM can be used to quickly prototype French-to-English translation [17]. Prior research about AAC users has often identified challenges with specific forms of communication, such as when talking to a physician or telling a long story [18], and that AAC users often conduct extensive work before a meeting to prepare what they wish to communicate [20]. Thus, we chose to explore how specific conversational tasks, such as requesting help with a particular object, or answering questions about one's background, could be supported by prompting an LLM. This approach is complementary to work that is focused on improving AAC expansions in everyday conversation [7, 31].

3.2 Speech Macros

We created Speech Macros to act as boundary objects and design probes to exemplify LLMs' capabilities and to demonstrate real-time output based on different conversational situations and user inputs.

Speech Macros were designed to be purpose-driven shortcuts that can generate complete sentences from a brief input, such as a single word. Informed by prior work that uncovered challenges in AAC-based social interactions [16, 20, 35], we created multiple prompts using the transformer-based large language model LaMDA developed for dialogue applications [32]. This model's output can be customized through zero and few-shot prompts. In our tests, we found we could provide 1-3 examples and a description of the desired output to produce reasonable results (see the Appendix for the prompts we used). The model produced a variable number of responses, which varied in length from a few words to several sentences. However, since the multiple sentences generated by the model often contained unrelated "hallucinations", we delimited each response to include only the first sentence generated. We restricted

our macros to showing the first four responses generated by the model so that they fit on the screen without requiring the user to scroll, and to provide a manageable number of suggestions to read and evaluate, comparable to the number of suggestions provided by existing AAC systems.

After testing different prompts through word choice and example iterations, we generated three Speech Macros that produced phrase suggestions for users based on different available contexts, underlying task instructions, and user inputs. We selected Speech Macros that performed well under different conversational situations, and with different types of user inputs. We then created a web-based prototype for each macro (Figure 1).

3.2.1 Extend Reply. Phrases produced by LLMs can leverage specific conversational context, like the ongoing dialogue, to provide specific responses that can help reduce misunderstandings among conversation partners while helping the AAC user be more specific about what they want to say. Motivated by the known problem that current AAC input methods may limit how detailed an AAC user's response can be (as more detail means more effort), the first Speech Macro, *Extend Reply*, extends a user's short input with more details that fit an ongoing conversation. To demonstrate this LLM functionality and to support users in sharing more detailed responses with less effort, the Extend Reply prototype has three main features: (1) a place where we represented the model knew what a conversation partner had just said (the current conversational context), (2) a place for user input to respond to the current conversation, and (3) suggested phrases by the model generated based on the instruction to extend the user input into a contextually relevant sentence that could be used in conversation.

3.2.2 Reply with Background Information. In addition to supporting user input during a conversation, we explored the possibility of allowing users to fill out information ahead of time and use that stored information to generate suggestions in a later conversation. During the study, we asked AAC users how an LLM could reduce their effort, and several participants mentioned that they often get asked the same questions repeatedly throughout the day, and prior work has shown that AAC users often write out things that they might want to say before a meeting so that it can be quickly retrieved during the conversation [20]. The *Reply with Background Information* Speech Macro accepts a paragraph of text in which the user includes information that they might wish to retrieve later. When they are asked a question, the AAC user can generate responses based on the previously supplied information. To maintain our interaction model of combining a conversation partner's question with user input to generate a response, the user does not enter additional text in this example, but instead presses the button to automatically generate potential responses, although future versions could certainly combine stored content with live input.

One feature of this Macro is that the system can automatically generate responses that match the phrasing of a specific question, regardless of how they originally wrote the information. For example, an AAC user might include a declarative statement in their bio such as "I have a cat named Kevin." If the conversation partner asks "Do you have any pets?" the system would reply with "Yes, I have a cat named Kevin.", while if the conversation partner asked "What is your cat's name?" the system would respond with "My cat's name is Kevin."

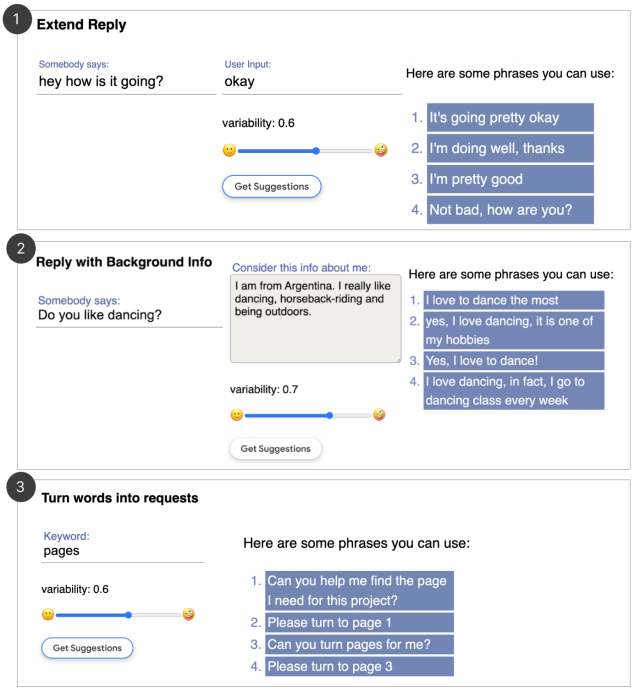

Figure 1: Prototype of *speech macros* used in the study. Each screen includes the name of the current macro, the conversation partner's question context, space for the user's input, and a 'variability' slider to adjust the temperature of the model, and thus the diversity of the responses generated. Output from the LLM is presented on the right side of the screen.

3.2.3 Turn Words into Requests. Another important key functionality of LLMs is that they can be prompted to complete specific tasks such as turning a word into a help request. We wanted to communicate this functionality to AAC users so we developed the *Turn Words into Requests* Speech Macro. We imagine AAC users could create their own instructions or prompts in the future to retrieve outputs from a model that fit their needs. The Turn Words into Requests Macro prototype consisted of only two components: (1) a place where the user could input a word they wanted to ask help with and (2) a space to see generated help request suggestions.

3.3 Prototype User Interface

Even if an LLM can be trained to provide high quality suggestions, there remains the challenge of integrating LLM feedback into the AAC user interface. At the same time, conducting early stage design studies requires communicating a lot of information to the user under significant time constraints.

For this study, we chose to sidestep any detailed questions about the user interface, and instead focus on a prototype that enabled AAC users to test the model and evaluate its output. Our prototype features a minimal user interface that highlights three main

components: a question from a conversation partner, user input, and suggested phrases from the LLM (Figure 1). The user is able to change either of the inputs and regenerate the suggestions; thus they can explore how different inputs lead to different suggestions, or how a particular input would function in response to different questions. For our study, the input fields were pre-populated with example text so that participants could immediately test the system and see live output from the model. Additionally, our prototypes included a *variability* slider that helped modify the model's output during the study in cases where the model produced the same text suggestions repeatedly. A higher variability value creates more random output. We explained the variability value to users and set it to an approximate mid-point of 0.6 and only changed it when the model did not suggest sufficient phrases or suggested repeated phrases. Our choice of 0.6 was based on multiple testing of our prompts and variability combinations that produced varied phrase suggestions.

4 EVALUATION STUDY

Knowing that LLMs can generate diverse outputs from one short set of inputs, we wanted to understand what type of prompts would be most useful to AAC users, what types of inputs they could provide, and what types of outputs were the most useful to support their communication. We designed a user study that would first introduce participants to LLM capabilities, focusing on three main abilities we thought could be the most relevant for AAC users: how a model can (1) suggest words based on conversational context, (2) draw from general world knowledge and (3) learn from examples and specific instructions (prompts). During the user study, participants also experienced the three speech macro design concepts and provided feedback during the study and in a post-study survey.

4.1 Participants

We recruited 12 adult expert augmented communicators, who use a variety of AAC devices, to test all three Speech Macro concepts. Our 12 participants (Table 1) included two eye gaze AAC users, four switch users and six AAC users who used direct selection to interact with their communication devices. Our participants used AAC solutions for multiple reasons including degenerative chronic illnesses, apraxia, cerebral palsy, autism, and also a combination of all these factors. None of our participants had aphasia or any disabilities affecting language use, only verbal speech production. Participants resided in the United States, the United Kingdom, or Canada.

In preparation for the study session we asked participants if they wanted to join the call with a person that could support their participation. We also asked about preferred communication styles and broadly about how we could make the study accessible specifically to them. Some participants joined the study session with a support person that helped them connect to the video call or communicate. Support persons were often family members or speech and language therapists. We did not consider support persons to be active study participants, but in some cases they did provide comments during the study, and we include those in the paper when appropriate.

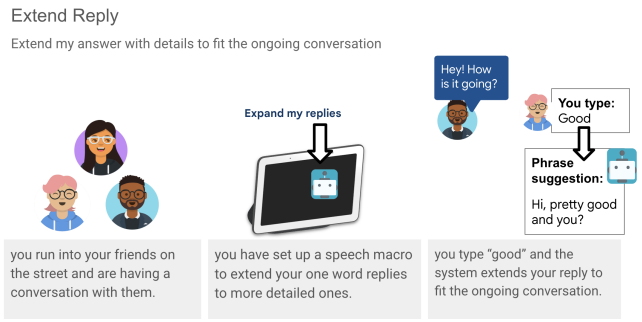

Figure 2: Storyboard exemplifying a sample interaction with the Extend Reply Speech Macro.

4.2 Procedure

We gathered participants' feedback through a 90 minute remote video call and a post-study survey. We divided the remote study session in three main parts: (1) introducing language technologies, what they are and how they appear in some products (10 minutes), (2) testing the three speech macros to evaluate their usefulness and their outputs (60 minutes), and (3) reflecting on other uses for speech macros (10 minutes). We also offered participants an optional 5 minute break (or more time if needed) that could be taken at any point during the study.

4.2.1 Introduction to LLMs. To introduce participants to LLMs, we presented different examples of AI-based language technologies that use LLMs, such as word prediction and word completion keyboards, auto-complete, and translation software. We explained that the LLM they would interact with during the study had learned patterns about vast amounts of text-based data from the internet. We noted that this made it a useful tool to support conversations since it could "understand" dialogue and suggest possible responses given a specific task or instruction. We introduced participants to Speech Macros as a way to define tasks the model could perform to assist in communication.

4.2.2 Trying out the Speech Macros. We then introduced participants to each Speech Macro by first illustrating a scenario using a Storyboard to illustrate specific use cases (Figure 2) and then shared the interactive prototypes via screen-sharing. All participants were introduced to the three Speech Macros in the same order: *Extend Reply*, *Reply with Background Information*, and *Turn Words into Requests*. We first explained each prototype and how these were just examples to test the model's functionality, rather than finalized communication device concepts.

For the *Extend Reply Speech Macro*, participants were presented with two example conversation scenarios. For the first, we said to imagine a conversation partner asked: "What did you do today?" and they, the AAC user, used a short word (*i.e.*, "work") to get extended replies from the system. For the second example, we explained that the system could adapt to different situations, such as the conversation partner asking, "Do you want to get pizza?" and then providing the same user input of "work". After each example, we asked participants what they thought about the suggestions, and

Table 1: Our 12 study participants (3 females, 2 non-binary (NB)) used a diverse set of AAC devices and techniques and had different levels of speech use.

ID	Age group	Gender	AAC Device	Years using AAC	Access Method	Non-verbal?
P1	45-54	M	Tobii Dynavox	1	Eye gaze	Fully
P2	45-54	M	Tobii Dynavox	4	Eye gaze	Often
P3	65-74	F	Proloquo4Text App on iPad	9+	Direct: mouse/joystick	Fully
P4	65-74	M	Grid 3	3	Direct: keyboard	Fully
P5	45-54	M	iPad with Predictable App	16	Direct: touch	Often
P6	25-34	F	Type on phone and Google docs	8	Head movement and switch	Sometimes
P7	25-34	NB	Dynavox maestro	23	Switch Scanning	Fully
P8	35-44	F	iPhone and Android phone	5+	Direct: touch	Sometimes
P9	25-34	M	Tobii Dynavox I-15 series	32	Switch Scanning	Fully
P10	55-64	M	Speech Assistant app on Galaxy S20+	9	Switch Scanning	Fully
P11	25-34	NB	Android tablet with Predictable, Coughdrop, and Speech Assistant Apps	8+	Direct: touch	Often
P12	45-54	M	iPad Pro with Proloquo4Text App	40+	Direct: toes	Often

whether they would accept any of the suggested phrases[2]. After reviewing the examples, we asked participants if they would like to try a user input to reply to either conversational scenario. If time allowed, participants could try more than one user input to reply to a conversation situation or to suggest a question they often get asked. At the end of the macro, participants were asked to rate how useful the functionality of having the system extend their reply would be to them and to comment on their rating using a scale ranging from "Not at all useful" to "Extremely useful".

For the *Reply with Background Information Speech Macro*, participants were first presented with a fictional sample biography: "I am from Argentina. I really like dancing, horseback-riding and being outdoors. I do not like insects. I love to eat ceviche, arepas, and tacos. I have a cat named Stella". The test conversational scenario was the question: "Do you like animals?" After discussing how the macro worked and rating the output for the example scenario, participants were invited to add their own biographical details to the existing text so that the model could use their own background information. Participants shared information about their favorite animals, favorite sports teams, hobbies, or country of origin.

For *Turn words into Requests*, participants were directly asked to think about any items or actions they would like to ask help with and suggest them to try as user inputs. After trying each suggested input, participants were asked to rate the generated phrase suggestions. After trying various inputs participants were asked to rate how useful was the Turn words into Requests functionality by using a scale ranging from "Not at all useful" to "Extremely useful".

[2]We originally asked participants to rate each output on a scale. However, participants were frequently unable to choose a rating for a single phrase set, so we omit these individual ratings from analysis and instead focus on the comments they provided after each phrase set, as well as the ratings and comments about each speech macro.

4.2.3 Post-study survey. A post-study survey was sent to participants to capture any additional open-ended feedback they did not get a chance to share during the study session. The post-study survey contained both multiple choice and open questions and was organized in sections: (1) feedback about specific Speech Macros and ideas for additional speech macro functionalities not covered, (2) feedback about the concept of Speech Macros (where they would be useful and where they would not be, what were some benefits, what were some concerns), (3) Priorities for future versions of speech Macros (important and less important use cases); (4) Using personal data in AAC (concerns about personal data use and information they would feel comfortable sharing), and (5) General feedback (Any other things you wish an AI-based communication system could do for you? Any additional feedback you would like to share?).

4.3 Data Collection and Analysis

Study sessions were video and audio recorded; audio recordings were transcribed using automatic speech recognition and corrected manually by the research team. Transcripts of the session were combined with a log of the session's text chat and researcher notes into a single document. Participant responses to the post-study survey were stored in a separate document.

Two members of the research team analyzed the data; both have several years experience in human-computer interaction and accessibility research. Both researchers had prior experience in conducting participatory design research with AAC users. One researcher had experience using accessible technology in their everyday life.

We performed qualitative data analysis to organize the findings and identify common themes [4]. First, the two researchers independently read through the 12 transcripts and post-study surveys, selecting quotes and observations and copying them onto separate notes, which were organized through several rounds of affinity diagramming [25]. We identified four categories of data: feature suggestions, potential use cases, comments about the quality of suggested phrases, and observations about using AI-enabled AAC in daily life. Feature suggestions and use cases were organized by which macro they related to, and are mostly presented in Sections 5.1 and 5.2. The remaining data were analyzed through several iterations of discussion, note-taking, and affinity diagramming, beginning with identifying the most common high-level themes in the data (characteristics of good/bad sentence suggestions, AAC use as self-expression, how AAC influences perceptions of its users, and concerns about AI) and grouping data into subcategories within them. These themes are largely discussed in Sections 5.3 and 5.4.

5 FINDINGS

We first report on participants' experience using Speech Macros, the inputs they tried for each macro, and the suggestions participants provided. We then report on participants' feedback about the model's output, and lastly on key user concerns to consider when integrating AI-based language technologies into communication devices.

As our participants had different speaking rates, we tried to spend an equal amount of time discussing each Speech Macro with each participant (around 20 minutes per Speech Macro). Some participants with faster speaking rates provided more user inputs of their own, while others provided at least one user input or conversational scenario suggestion. The statistical Median of interactions with the *Extend Reply* macro was three, and two for the *Reply with Background Information* and *Turn Words into Requests* macro.

All participants completed the post-study survey and provided extended written replies and thoughts (Median: 376.5 words, Max: 1023 words, Min: 43 words). Only three responses were less than 100 words while all the others were above 250 words.

5.1 Benefits and Uses of Speech Macros

Participants found the conversational tasks that the speech macros supported to be in general very useful to them (Figure 3). From all three speech macros, the "Turn words into Requests" was rated more often to be either extremely or very useful followed by "Reply with background information" and then "Extend Reply."

5.1.1 Extend Reply. Participants liked only having to input a few words to get phrases extended by a macro since it could help reduce typing effort and fatigue. *"[the extend reply macro] would enable me socialize and network more because I would be able to type faster and would require less time effort and energy and also lessen frustration, sometimes I just don't initiate conversations because I don't have the energy to type a lot and I can't answer quickly enough"* (P6). P11 also shared that phrase suggestions could help them alleviate cognitive effort in trying to remember what to say that could appropriately fit the social situation. *"That sounds super helpful for knowing what words to use that are socially acceptable rather than getting stuck and trying to remember and at the same time trying to go through the physical actions of using AAC."*

The Extend Reply macro generated alternative responses to the same question, for example "I'm hungry" or "I'm not hungry", to reply to "do you want to get pizza?" (Table 2). P6 really liked having options as it would allow her to choose different suggestions depending on her mood, the circumstance, and the person she is addressing.

Given that macros were created using few-shot prompts that were straightforward expansions of an input (see Appendix for prompts), some phrase suggestions seemed too cold-cut for social conversation. When asked to rate how appropriate the model's suggestions were, participants brought up needing more information about the scenario: would they be using the extend macro on a mobile phone or a computer? (P5) who are they addressing?(P2,P3,P6,P10) what is their mood?(P6,P11). For scenarios in which they imagined talking to their friends, participants preferred phrase suggestions that were casual and friendly instead of more straightforward options (Table 2).

5.1.2 Reply with Background Information. One strength of LLMs is that they can draw general information about the world into their generated responses. This ability seemed to be very useful when the participants tested how the model suggested phrases for them in the "Reply with Background" macro. The ability to guess and imply information from a short biography provided by participants seemed helpful when it related to factual knowledge about the world. For example, when telling the model that she was from Sri Lanka originally, the model made suggestions like "I love to go to the beach" when replying to the question "what do you like to do for fun?"

> *"those activities are things that I would do back home. It knew somehow that because I am from Sri Lanka, that is something important to me, so I love that it took that piece of information and took it as my background. I think that is important."* - P8

In another case, P9 mentioned his favorite soccer team, the *Chivas*, and was pleased when the system was able to identify the location of the team, as it suggested the phrase "my favorite team is Chivas de Guadalajara". While P9 was impressed by the system's knowledge, he ultimately chose a different response that better conveyed his lifelong enthusiasm for the team: *"My favorite soccer team is Chivas, I have supported them since I was ten".*

On the other hand, the language model often guessed and made up information to fill in the details that seemed random and unrealistic or "wrong funny" as expressed by P11. The model guessed that P11's dog was "a cute little dog" based on its name "Pippin," which P11 provided in their background information. In reality, Pippin is a 100 pound German Shepherd. *"The false assumptions would make me want to go add bio information later but not right then."* - P11.

Participants shared examples of meaningful context they would want the system to draw from to make more useful suggestions. Participants wanted the model to know medical information about them so that they could have informed conversations with their medical doctors (P10, P11). Participants also wanted the model to automatically draw information about the time of day and their location (P7, P10, P12). Some AAC users wanted the device to know

Table 2: Extend Reply Speech Macro inputs given by our study participants, the suggestions returned by the speech macro, their preferred option (shown in bold), and their rating of how appropriate that suggestion would be for the question asked and their personal communication style.

Question	ID	Input	Suggestions	Rating
Do you want to get pizza?	P6	Later	1. I'm feeling pizza later, are you? **2. What time do you want to get pizza?** 3. Sure, let's get pizza 4. Ok, where do you want to go?	Extremely appropriate
	P1	Eat	**1. I'm always up for pizza** 2. I'm sorry, I'm not hungry right now 3. I'm not hungry 4. I'm hungry	Very appropriate
	P3	Work	**1. I'm sorry, I can't go out this afternoon** 2. I'm already at work 3. I'm busy with work today 4. I'm working today	Not at all appropriate

about their favorite topic, like their favorite soccer teams (P5 and P9), and even be aware about common topics shared with specific conversation partners (e.g., co-workers, friends):

> *"I think background information could encompass a great deal beyond demographics. For example, being in IT includes a kind of sub-vocabulary relevant to speaking with colleagues, as a chess player there exists a sub-vocabulary, and the roles of husband, father, grandfather, church brother, and more could each have identifiable "background information" that could influence phrase generation."* - P10

5.1.3 Turn Words into Requests. The "Turn Words into Requests" macro worked well when requesting help with inputs about common tasks (i.e., tea and biscuits, bathroom, sleep), and was described positively by participants indicating that it could be useful to them (Figure 3). Nonetheless, when receiving input like "smoke", in a case where the user wanted to try to request his caregiver to take him out for a smoke, the model steered the user from this specific activity. During the study, we tried different input including smoke and cigarette and for both tries the model suggested phrases such as: "no smoking in the house", "I do not want to smoke", "please don't smoke a cigarette". We speculate this subject was caught by a safety or policy layer intended to avoid promotion of this specific activity.

We found the system was not able to suggest relevant requests related to access or health needs. For example, when P3 tried the word "transfer" to see if the model could suggest requests related to needing help transferring from her chair to the bed using her home-installed lift, the model only suggested phrases related to transferring money (i.e.,"transfer money to my bank account"). Other user inputs related to medical and health requests also did not work well. Both P1 and P10, who have a tracheostomy and often need suction to clear their breathing pathway, tried using "suction" to generate a request but the model suggested something unrelated like, "I have a suction cup that needs to go on the wall". Many help requests that users wanted to generate were high stakes and needed to be specific. So when users tried inputs as "meds" or "itch", the model suggested very generic help requests that participants tended to evaluate as less useful and less appropriate.

5.1.4 Additional Conversational Tasks. Participants also shared different ideas about how they would use speech macros. Several participants said they would program them to make specific help requests to their voice assistants (P1, P6). Others would like to have speech macros help with routine tasks like asking for help with self-care tasks (P7) and ordering at a restaurant (P4, P12) or at a coffee shop (P10), as the model could draw information about what type of food the restaurant sells or what type of coffee order the user always asks for. Some participants also shared that they would not mind using the "Expand Reply Macro" to get suggestions on how to answer common questions like "How are you doing?" which can be frustrating to answer again and again (P1). Speech macros that could draw from background information could be used to plan conversations with doctors (P11) if the medical data given to the system was guaranteed to be secured and private (P10, P11).

5.2 Learning Input Mappings

While the macros were able to suggest reasonable phrases that matched the questions asked, participants expressed uncertainty on how the system used the input it was given. In other instances, AAC users proposed use of short-hand as input in order to have more control over the generated output.

Participants suggested using specific notation like adding symbols to their inputs to overcome not knowing what implicit tone the model would decide to go with when suggesting responses. P10 suggested using symbols that could hint the model towards having a more positive phrasing: *"typing 'word +' would lead to positive responses."*. Overall, participants expressed a preference for reliable and short inputs. P4 explained he had a macro programmed in his device that expanded "1y" to yes and automatically played it out loud saving him time. P8 used "WON1" to say wonderful when others asked how she was doing. P2 used "SYS" as a shorthand that expanded into "see you when I see you".

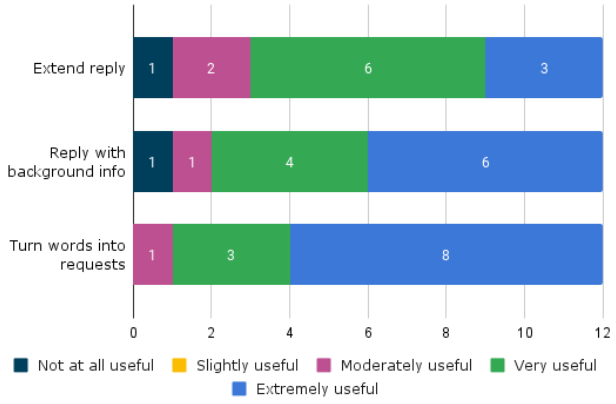

Figure 3: Most participants found speech macros to be either very or extremely useful. Turning words into requests was considered very and extremely useful by most participants. Participants identified they could use the turn words into requests macro to help them daily to create requests for routine, self-care and accessibility related tasks.

An unexpected LLM behavior resulted from having phrase suggestions insert non-factual information, a certain tone, or preference that was not specified by the user. The "Extend Reply" macro would sometimes suggest arbitrary responses that seemed too specific like "I worked from 8am to 6pm" as a response to the question "what did you do today?". P11 was interested in knowing where the suggestion came from: *"it goes into assumptions of how long you worked. Is that something it learns? That you have patterns?"* Participants were also curious about how to control the variability of the options as sometime the macro repeated phrase suggestions at the testing variability value we tried (0.6).

P10 also noted that some of the phrases suggestions carried an implicit "negative tone" that he would not necessarily intend if he typed the input "work" as a reply to the situation "Do you want to get pizza?" In another case, P3 tried to retrieve neutral answers to the question "how is it going?" by using the phrase "it's going" as input. The Extend macro did not catch that the user wanted to express a neutral response and instead suggested phrases that complemented the users input: *"it's going pretty good"*. P3 understood that the underlying speech macro task did not catch her meaning and added: *"if things weren't going well, I would learn to not confuse the program. and use a different input. [I would use] "not great" for instance."*

5.3 Evaluating Suggested Phrases

While our speech macros performed in an expected way and suggested phrases that were mostly relevant within the conversation scenarios tested, LLM output was insufficient in supporting AAC users in adding their personal tone and style and representing their personality and identity. Participants also shared concerns about how the LLM-generated output could affect their social relationships by being too abrupt or just by the fact that others could know they were automated responses.

Many phrases were either considered as very appropriate or were phrases that were close to something the participant would say but that needed a little more editing to get it right. Participants suggested ways in which they would use the model's generated output as a starting point to build on top of it during a real-time conversation. *""the first phrases will be like i almost want to answer this it gives me an idea of what i want to say, sometimes but it doesn't necessarily fit exactly with what i want to say... i might end up using a predicted phrase and then delete the last word and write my own"*(P11).

5.3.1 Tone and Style. When discussing the different outputs generated through the three speech macros, participants' most common critique was that the phrases did not reflect their personal style and the image of themselves they wanted to portray. Nonetheless, the way each participant described what qualities of their background mattered to them and what they wished the model knew about them varied. Some participants talked about their personality and how they aim to convey it through their words. Other participants shared the importance of their culture, including their country of origin and their faith and how that impacts the way they talk. Lastly, participants talked about the impressions they wanted to leave on others and how they were concerned the generated phrases could affect their relationships.

5.3.2 Communicating personality, style, and identity. While AAC users found value in being able to get phrase suggestions by typing less, they indicated they would not necessarily use these phrases as they lacked their personal style and did not reflect personality. In many cases, users found phrase suggestions to be "too bland" and impersonal while their personality was "witty" (P2) or "positive" (P8). When sentences lacked a way to convey an important part of the participant's personality and their values, participants said they would not use them. P8 and P10 talked about their faith and personal belief and how that influenced the way they talked to others. *"For me I would add something positive...because I think that positivity in the world is lacking these days so I add positivity into my answers"* (P8). Participants also expressed that they did not wish to sound too scripted. P9 was concerned about sounding "robotic" and P12 raised an important future risk: *"If the system is being used in the future, are all AAC users going to talk the same way? That's something we need to think about."*

5.3.3 Maintaining social standing. Participants expressed concerns related to how the generated phrases could impact their personal relationships. Some participants commonly found generated phrases to be too abrupt (P6, P8, P9, P10, P12) when suggesting possible replies to social questions like "do you want to get pizza?" in the Extend Reply macro scenario, and said that the phrase suggestions were not appropriate for this reason. Participants more commonly preferred phrases that were more socially correct like, "I am sorry, I can't go out this afternoon" rather than the more abrupt ones: "I'm already at work" (Table 2). P12 explained why abrupt responses, even though they matched the question, were not appropriate: *"it's not very appropriate because if someone asked me that question and I typed in work, I want phrases like, what time do you want to go? Or I'm working, could we reschedule or give me a minute and I will get back to you about lunch?"*

Table 3: Participants suggested different items to be turned into help requests. Items spanned three main categories: medical, accessibility related, and daily tasks.

	User input	Participants
Daily living	TV	P2
	Outside	P2
	Sleep	P2
	Bed	P2
	Bathroom	P2, P6, P12
	Window	P3
	Service	P4
	Cigarette / Smoke / I want to smoke	P5
	Tea and Biscuits	P5
	Drink	P6
	Book / Book pages	P8
	Starbucks	P10
	Water	P12
Accessibility	Access	P7
	Transfer	P3
Medical	Suction	P1, P10
	Itch	P1
	Meds	P11
	Pain	P3

Participants also had concerns about the system suggesting the wrong thing and making the participants look bad. P4 worried the system would suggest phrases that would reveal information about how he truly felt about someone, or if he had talked badly about someone with somebody and the potential for the system to reveal that fact, embarrassing him. P2 and his wife also had concerns about inputting sensitive words to the system and then having it show unwanted output. P2 wanted to use "bed" to have the model make requests related to helping him prepare for bed. P2's caregiver was concerned about what the system would say and so did not want him to use that word so they decided to try the word "sleep" as input instead: *"He usually says I'd like to get ready for bed and I didn't want to use bed because I didn't want to see what would come up..."*

P2 and his caregiver also shared that for more personal requests like going to the bathroom, they might not want to use the word "bathroom" as a user input but instead a euphemism. *"He would probably use other words at home, and say other words in public to keep it private."* P2 shared he would say: *"I have to powder my nose."*

5.4 Concerns About Using AI Suggestions

5.4.1 Privacy and data concerns.

> *" I would love a local version of this [reply with background information macro] which i don't worry about privacy that i could put medical history information into and it could help me make sure i actually give doctors details when i am in a high stress situation and everything is going so fast i can't keep up"* - P11

Sharing background information with a system was considered extremely useful by most participants. Participants said they wanted to use this feature but would be concerned about other people using their data without their consent, identity theft and data breaches, having other humans reading their personal or medical background information, and on-cloud data processing. Participants asked whether there was a way for them to confirm their data was encrypted and how they could turn off the system from hearing the conversations all the time, as it would need to do this to catch the current dialogue. Participants understood their data and conversation data would improve the model and that the speech macros would indeed be very useful but if there were no privacy in place or clear transparency about how it would work, they would absolutely not use it.

5.4.2 AI suggestions could undermine autonomy. Selecting a generated phrase from the system could have social consequences beyond what the content of that phrase is and how it is interpreted. Participants reflected on how selecting an automated phrase, even a pre-stored phrase they had created beforehand, made others believe the system did all the work for them. P8 shared a story where using her AAC device for a job interview allowed her to get the job because people understood that she had prepared answers to the interview questions beforehand using her device and attributed her with being responsible and well prepared. Unfortunately in another occasion, people thought the opposite: *"I pressed a button for a customized answer and someone said 'oh, the device did that for you'. That was insulting to get that answer. They thought I was an idiot because I had these customized phrases pre-installed. I was saddened. I had to stand up for myself, so I said I had made the preparation for the meeting."*

P10 shared that he only allows very few intimate people to watch him type as he communicated and he feels that if they see him select an automatically generated phrase instead of typing his own that could have a negative consequence or change the meaning of that intimate communication.

> *"... AI generated responses may become the answer for the types of communication that are centered around content and timing. But there are more intimate forms of communication where AI might get in the way of personal expression. I have found with my wife, daughter, and a few close friends that sitting beside me and watching me type the reply to their comment (clearing my response without speaking it) is a secure form of communication for them with me. I would go so far as to say, if the exact expression desired popped up on the list, choosing it would mean something different to an observing intimate friend than if I were to type it."* - P10

Similarly, P8 shared she would not use speech macros with people close to her: *"[I would not use this] with my family and friends who know me personally. I think sometimes the human feelings cannot be translated by devices."*

6 DISCUSSION

In this work, we presented AAC users with an interactive prototype in which they could evaluate text suggestions produced by a large language model. Based on this work, we reflect on how

language models may be integrated into AAC devices and about our experiences integrating AAC users into the research process.

6.1 Can Speech Macros Improve AAC Use?

Overall, our participants were excited about the possibility of using AI to improve their AAC systems. Participants were clear about the amount of effort they expended when communicating, and the value in reducing some of that effort. At the same time, participants already had ideas about how to improve output from the system, to more clearly reflect their own preferences and communication styles. They also presented some concerns about sharing data, and about the potential loss of *control.*

One question that we considered throughout the work is whether it is useful to focus on specific use cases, as we have done here, as opposed to a more general system that predicts phrases across all conversation contexts, as in some related projects [7, 31]. As a research tool, we found that this approach was successful in introducing participants to the relevant concepts, and providing specific contexts in which to test and evaluate LLM output. Participants also gravitated toward specific macros (especially "Turn Words into Requests"), which may help to prioritize future work. However, whether speech macros should be introduced into AAC devices remains an open question. Existing AAC devices often have multiple modes, such as a mode to replay stored phrases, and it is possible that specific speech macros could similarly be incorporated into current AAC user interfaces. Alternatively, it may make sense to provide AAC users with the ability to create and customize their own prompts to the LLM, enabling them to customize output through prompt programming.

6.2 Design Challenges and Trade-offs

While our prototype provided an intentionally simplified interface, creating and testing the prototype revealed tensions between the potential benefits of integrating AI and potential negative effects which we summarize here.

6.2.1 Reducing effort vs. maintaining control. As noted by the majority of our participants, communicating using their current AAC devices can sometimes be both frustrating and ineffective. AI generated suggestions offer the possibility to reduce the amount of effort. However, participants were often unsatisfied with the output from the system, finding it had the wrong tone or was simply incorrect. Participants noted that in some cases they could edit the response to get the result they wanted, but in others they would need to rewrite the entire response. In either case, editing or rewriting an input is counterproductive to the goal of reducing keystrokes. Nonetheless, while our participants wanted to type less to save physical and cognitive effort when responding to routine questions, participants were concerned about how automatic phrase generation could impact their relationships. Participants shared how putting effort into their communication by preparing long messages demonstrated to others that they cared, about a job interview (P8) or about a close family member (P10). These findings align with prior work highlighting how effort invested in computer-mediated communication can be a symbol for caring [21]. While keystroke savings can reduce time and effort, future explorations with AI should consider how views on authorship and effort may impact relationships among AAC and non-AAC users.

When encountering low quality suggestions, participants sometimes tried to enter longer queries, or invented new input conventions, such as entering a plus after their input to retrieve positive responses. Providing more robust input options could provide users with more control. Similarly, showing users potential predictions as they type could help users make decisions about when to try predictions *vs.* typing out messages themselves.

6.2.2 Composing in real time vs. using stored content. AAC users optimize and plan for their communication and social interactions. A lot of preparation happens, they create and store phrases that anticipate potential questions or any high pressure situations they may encounter. While this study does not engage with situations in which AAC users could use the technology to plan content for future encounters, participants gave clear feedback of how the output could be better and even how they wanted to combine phrases created by them with the ones generated by the system.

It is important not to ignore all the work and setup AAC users have already established that work for them, like the shortcuts they already have in place (i.e, 1y for yes, WON1 for wonderful, etc). If AI will be used within an AAC device, it should be flexible so that it does not impose a scheme but learns from and is customized by the user who has already developed a system for their communication.

6.2.3 Achieving functional communication goals vs. expressing oneself. Prior work has reported that the way people communicate changes once they start using AAC systems [20, 24]. Individuals who acquire speech disabilities later in life often lose the ability to express sarcasm, humor, and nuance. The social timing pressures that exist for AAC device users make it hard to add nuance and talk about other things beyond requests or short utterances [18]. This is why sometimes there are large efforts into enabling basic communication with AAC devices, and even though there has started to be a shift into how AAC devices could also play a role to maintain social relationships [9], we did not know what qualities of generative output could be important to know about for future systems. Our study revealed that AAC users want more from LLMs in addition to keystroke savings and achieving that model-generated golden reply that is reasonable within context. Users want to be able to customize output to their needs (and this is different for each person); this is the key to unlocking the potential of LLMs. Moving away from scriptedness and transactional conversation support towards customized use of these systems.

While prior work established that computer-mediated communication (CMC) grants greater control over the impressions people convey to others as people can edit and plan their messages before sending them [15, 37], AAC communication is a unique type of CMC where responses are expected sooner, and the time window to achieve self-representation is shorter. Perhaps by enabling post-processing of LLM suggestions or co-authoring of a response we might support AAC users in personalizing their responses on the go.

Additionally, our study revealed the importance of customized user information to better tailor LLM-generated output to. Prior work created simulated personas mostly comprised a one to two sentence description of a person's hobbies or personal preferences

to generate conversational phrases [31]. Through our study we gathered other important information about what AAC users would like a model to know about them: medical details, details about their relationship with conversation partners (co-worker, family), details about their work and context information about their location or time of day.

6.3 Conducting Human-Centered AI Research with AAC Users

Conducting participatory and open-ended research with AAC users can be challenging as new methods are always needed to elicit real-time feedback in a way that aims to maximize participation and reduce user burden [1, 3]. These challenges are amplified by the need to conduct studies remotely during the COVID-19 pandemic. Despite these challenges, participation from AAC users is necessary to ensure that technology accurately meets their needs. A primary goal of this research was to enable AAC users to interact with, and provide feedback about, a real AI model. All AAC users were able to successfully suggest inputs and provide feedback about generated outputs. Additionally, the use of a post-study survey was successful in allowing participants to share more extended thoughts, anecdotes, and feedback.

We conducted each of our 12 interviews remotely, and encountered some anticipated and unanticipated challenges while doing so. First, we knew that the pace of interaction during the study might be slow, and that the technologies we wished to discuss might be unfamiliar to our users. Second, we expected that some participants would be limited in their ability to give feedback during the session. Third, because most participants would be using their devices for communication, we could not install software on their devices.

Generally, we were able to adapt our study protocol to these circumstances. First, we designed our three speech macros to be simple and easy to understand. We designed our prototype to quickly show the language model by allowing the user to change the contextual information or the prompt and quickly see the results. Through building in a variability slider we were able to run the model at other randomness values if needed. By using screen sharing, we were able to accept spoken or typed user input through chat and input it to the prototype. Finally, as in previous studies [20], we combined our in person interview with a follow-up questionnaire so that participants could compose longer responses. Our study allowed us to explore potential uses of AI for AAC by enabling users to directly interact with a language model via digital prototypes. While our speech macro concepts served us to communicate LLM's capabilities and usage scenarios, co-designing future speech macros or prompts with AAC users may enable us to understand how AAC users create mental models and expectations for LLMs.

Overall, the design of our prototype and study protocol enabled our participants to see the LLM in action, test it with several queries, and provide feedback during and after the study session. While this configuration worked well for this study, conducting longer deployments of this technology would require an alternative setup.

6.4 Limitations and Future Work

While we designed our prototype and study to maximize the amount of interaction participants would have with the language model, they were still limited to completing 5-10 inputs in the study session. As a result, participants gave feedback based on this limited experience only. Future studies could feature a longer deployment, so that participants could input more prompts and gain a better sense of the strengths and weaknesses of this approach.

The language model used in the study was designed for generic dialog-based applications; each speech macro was a prompt written for that specified activity. We did not customize speech macros with any data about specific users. While the speech macros proved effective for collecting user feedback, this study does not provide a clear picture of how well current language models could perform for AAC suggestions, and it is likely that tuning the speech macro prompts or adding user-specific data to the prompts would improve suggestion quality. Furthermore, participants provided suggestions for new macros; providing end users with the ability to edit and customize their own macros could further improve results.

We designed our *speech macros* prototype with the goal of quickly introducing users to the concept and enabling them to try out several macros during the study session. By design, the prototype does not include details of user interaction, such as whether suggestions should be added to an on-screen keyboard or placed in a menu, or what actions are needed to generate suggestions. During our study, participants emphasized the importance of having control over predictions and the usefulness of composing messages ahead of time. Participants were also aware that the system might not produce the output they intended, and that they might need to fall back to typing a message themselves. Future work could explore the design of appropriate user interfaces that combine contextual information, saved settings, and live input, and allow users to correct or override the system if it fails to produce usable output in a particular situation. There may also be ways to provide users with more control without requiring more typing, such as by allowing users to specifically request a positive or negative response, or to request short or long responses.

For this study, we recruited participants who used AAC due to motor difficulties; our participants had typical language skills and did not have conditions such as aphasia that would impair their understanding of spoken or written messages. We were thus able to assume that participants were able to understand the purpose of the macro and its expected input, correctly format their input messages, and choose from the suggested phrases without difficulty. We expect that this approach could be made useful to those with aphasia and other language disorders by designing appropriate interfaces to detect and highlight potential issues related to comprehension, or by personalizing user models that accept input and generate output in a format more appropriate for that individual user.

7 CONCLUSION

We conducted an early stage evaluation of large language models as a tool to support AAC users in generating phrases, introducing *speech macros* as a method for AAC users to benefit from the generative capabilities of these models. Our study found that AAC users were enthusiastic about the potential of language models to support their communication, provided that they maintain control of their personal expression. The potential of LLMs for different

types of AAC use should be explored through future design work and experiments.

ACKNOWLEDGMENTS

We thank our participants who shared their insights and time. We also thank Google's People and AI Research team members, especially Mahima Pushkarna, whose feedback and support made this work possible.

REFERENCES

[1] Erin Beneteau. 2020. Who are you asking?: Qualitative methods for involving AAC users as primary research participants. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.

[2] David R Beukelman and Pat Mirenda. 2013. *Augmentative and alternative communication: Supporting children and adults with complex communication needs*. Paul H. Brookes Pub.

[3] Sarah W Blackstone, Michael B Williams, and David P Wilkins. 2007. Key principles underlying research and practice in AAC. *Augmentative and alternative communication* 23, 3 (2007), 191–203.

[4] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[6] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 1877–1901. https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

[7] Shanqing Cai, Subhashini Venugopalan, Katrin Tomanek, Ajit Narayanan, Meredith R Morris, and Michael P Brenner. 2022. Context-Aware Abbreviation Expansion Using Large Language Models. *arXiv preprint arXiv:2205.03767* (2022).

[8] Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A Smith. 2021. All That's 'Human' Is Not Gold: Evaluating Human Evaluation of Generated Text. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Online, 7282–7296.

[9] Jiamin Dai, Karyn Moffatt, Jinglan Lin, and Khai Truong. 2022. Designing for Relational Maintenance: New Directions for AAC Research. In *CHI Conference on Human Factors in Computing Systems*. 1–15.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[11] Alexander Fiannaca, Ann Paradiso, Mira Shah, and Meredith Ringel Morris. 2017. AACrobat: Using mobile devices to lower communication barriers and provide autonomy with gaze-based AAC. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. ACM, 683–695.

[12] Mauricio Fontana de Vargas, Jiamin Dai, and Karyn Moffatt. 2022. AAC with Automated Vocabulary from Photographs: Insights from School and Speech-Language Therapy Settings. In *Proceedings of the 24th International ACM SIGACCESS Conference on Computers and Accessibility* (Athens, Greece) *(ASSETS '22)*. Association for Computing Machinery, New York, NY, USA, Article 23, 18 pages. https://doi.org/10.1145/3517428.3544805

[13] Mauricio Fontana de Vargas and Karyn Moffatt. 2021. Automated Generation of Storytelling Vocabulary from Photographs for use in AAC. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Online, 1353–1364. https://doi.org/10.18653/v1/2021.acl-long.108

[14] Steven M Goodman, Erin Buehler, Patrick Clary, Andy Coenen, Aaron Donsbach, Tiffanie N Horne, Michal Lahav, Robert Macdonald, Rain Breaw Michaels, Ajit Narayanan, et al. 2022. LaMPost: Design and Evaluation of an AI-assisted Email Writing Prototype for Adults with Dyslexia. *arXiv preprint arXiv:2207.02308* (2022).

[15] Carla F Griggio, Joanna Mcgrenere, and Wendy E Mackay. 2019. Customizations and expression breakdowns in ecosystems of communication apps. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–26.

[16] D Jeffery Higginbotham and DP Wilkins. 1999. Slipping through the timestream: Social issues of time and timing in augmented interactions. *Constructing (in) competence: Disabling evaluations in clinical and social interaction* 2 (1999), 49–82.

[17] Ellen Jiang, Kristen Olson, Edwin Toh, Alejandra Molina, Aaron Donsbach, Michael Terry, and Carrie J Cai. 2022. PromptMaker: Prompt-Based Prototyping with Large Language Models. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI EA '22)*. Association for Computing Machinery, New York, NY, USA, Article 35, 8 pages. https://doi.org/10.1145/3491101.3503564

[18] S.K. Kane and M.R. Morris. 2017. Let's Talk about X: Combining image recognition and eye gaze to support conversation for people with ALS. *DIS 2017 - Proceedings of the 2017 ACM Conference on Designing Interactive Systems* (2017), 129–134. https://doi.org/10.1145/3064663.3064762

[19] Shaun K Kane, Barbara Linam-Church, Kyle Althoff, and Denise McCall. 2012. What we talk about: designing a context-aware communication tool for people with aphasia. In *Proceedings of the 14th international ACM SIGACCESS conference on Computers and accessibility*. 49–56.

[20] Shaun K. Kane, Meredith Ringel Morris, Ann Paradiso, and Jon Campbell. 2017. At Times Avuncular and Cantankerous, with the Reflexes of a Mongoose: Understanding Self-Expression Through Augmentative and Alternative Communication Devices. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (Portland, Oregon, USA) *(CSCW '17)*. ACM, New York, NY, USA, 1166–1179. https://doi.org/10.1145/2998181.2998284

[21] Ryan Kelly, Daniel Gooch, Bhagyashree Patil, and Leon Watts. 2017. Demanding by design: Supporting effortful communication practices in close personal relationships. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. 70–83.

[22] Per Ola Kristensson and Keith Vertanen. 2012. The potential of dwell-free eye-typing for fast assistive gaze communication. In *Proceedings of the symposium on eye tracking research and applications*. 241–244.

[23] Charlotte P Lee. 2007. Boundary negotiating artifacts: Unbinding the routine of boundary objects and embracing chaos in collaborative work. *Computer Supported Cooperative Work (CSCW)* 16, 3 (2007), 307–339.

[24] Miechelle Mckelvey, David L Evans, Norimune Kawai, and David Beukelman. 2012. Communication styles of persons with ALS as recounted by surviving partners. *Augmentative and Alternative Communication* 28, 4 (2012), 232–242.

[25] Matthew B Miles and A Michael Huberman. 1994. *Qualitative data analysis: An expanded sourcebook*. sage.

[26] Cecily Morrison, Edward Cutrell, Martin Grayson, Anja Thieme, Alex Taylor, Geert Roumen, Camilla Longden, Sebastian Tschiatschek, Rita Faia Marques, and Abigail Sellen. 2021. Social Sensemaking with AI: Designing an Open-ended AI experience with a Blind Child. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–14.

[27] Martez E Mott, Shane Williams, Jacob O Wobbrock, and Meredith Ringel Morris. 2017. Improving dwell-based gaze typing with dynamic, cascading dwell times. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 2558–2570.

[28] Changhoon Oh, Jungwoo Song, Jinhan Choi, Seonghyeon Kim, Sungwoo Lee, and Bongwon Suh. 2018. I Lead, You Help but Only with Enough Details: Understanding User Experience of Co-Creation with Artificial Intelligence. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3173574.3174223

[29] Albert B Robillard. 1994. Communication problems in the intensive care unit. *Qualitative Sociology* 17, 4 (1994), 383–395.

[30] Jennifer M Seale, Ann M Bisantz, and David J Higginbotham. 2020. Interaction symmetry: Assessing augmented speaker and oral speaker performances across four tasks. *Augmentative and Alternative Communication* (2020), 1–13.

[31] Junxiao Shen, Boyin Yang, John J Dudley, and Per Ola Kristensson. 2022. KWickChat: A Multi-Turn Dialogue System for AAC Using Context-Aware Sentence Generation by Bag-of-Keywords. In *27th International Conference on Intelligent User Interfaces* (Helsinki, Finland) *(IUI '22)*. Association for Computing Machinery, New York, NY, USA, 853–867. https://doi.org/10.1145/3490099.3511145

[32] Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Vincent Zhao, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Pranesh Srinivasan, Laichee Man, Kathleen Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Aguera-Arcas, Claire Cui, Marian Croak,

Ed Chi, and Quoc Le. 2022. LaMDA: Language Models for Dialog Applications. https://doi.org/10.48550/ARXIV.2201.08239

[33] Ha Trinh, Annalu Waller, Keith Vertanen, Per Ola Kristensson, and Vicki L Hanson. 2012. Applying prediction techniques to phoneme-based AAC systems. In *Proceedings of the Third Workshop on Speech and Language Processing for Assistive Technologies*. 19–27.

[34] Keith Trnka, John McCaw, Debra Yarrington, Kathleen F McCoy, and Christopher Pennington. 2009. User interaction with word prediction: The effects of prediction quality. *ACM Transactions on Accessible Computing (TACCESS)* 1, 3 (2009), 1–34.

[35] Stephanie Valencia, Amy Pavel, Jared Santa Maria, Seunga (Gloria) Yu, Jeffrey P. Bigham, and Henny Admoni. 2020. Conversational Agency in Augmentative and Alternative Communication. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3313831.3376376

[36] Keith Vertanen. 2017. Towards Improving Predictive AAC using Crowdsourced Dialogues and Partner Context. In *Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility*. ACM, 347–348.

[37] Joseph B Walther. 1996. Computer-mediated communication: Impersonal, interpersonal, and hyperpersonal interaction. *Communication research* 23, 1 (1996), 3–43.

[38] Qian Yang, Justin Cranshaw, Saleema Amershi, Shamsi T. Iqbal, and Jaime Teevan. 2019. Sketching NLP: A Case Study of Exploring the Right Things To Design with Language Intelligence. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300415

[39] Qian Yang, Aaron Steinfeld, Carolyn Rosé, and John Zimmerman. 2020. Re-Examining Whether, Why, and How Human-AI Interaction Is Uniquely Difficult to Design. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376301

[40] Nur Yildirim, Alex Kass, Teresa Tung, Connor Upton, Donnacha Costello, Robert Giusti, Sinem Lacin, Sara Lovic, James M O'Neill, Rudi O'Reilly Meehan, Eoin Ó Loideáin, Azzurra Pini, Medb Corcoran, Jeremiah Hayes, Diarmuid J Cahalane, Gaurav Shivhare, Luigi Castoro, Giovanni Caruso, Changhoon Oh, James McCann, Jodi Forlizzi, and John Zimmerman. 2022. How Experienced Designers of Enterprise Applications Engage AI as a Design Material. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 483, 13 pages. https://doi.org/10.1145/3491102.3517491

Appendix A LARGE LANGUAGE MODEL PROMPTS

This section contains the prompts used in the study, which were passed to the large language model using an interface similar to [17]. Variables are marked with [square brackets].

A.1 Extend Reply

```
Q: Do you want to go to the movies?
Input: no
A: No thanks, I'm busy this afternoon.
Q: How are you?
Input: good
A: I'm pretty good. How are you?
Q: [question]
Input: [user input]
A:
```

A.2 Reply with Background Information

```
Consider this background information about myself: [user input]
Q: Where are you from?
A: I am from Argentina, it is the southernest country of south america.
Q: Do you have any hobbies?
A: I love going on hikes and going horseback riding.
Q: [question]
A:
```

A.3 Turn Words into Requests

```
Help: fruit
Phrase: I'd like to have some fruit please
Help: bed
Phrase Can you help me get to bed?
Help: Shoes
Phrase: can you help me put on these shoes?
Help: [user input]
Phrase:
```