A-CONECT: DESIGNING AI-BASED CONVERSATIONAL CHATBOT FOR EARLY DEMENTIA INTERVENTION

Junyuan Hong[†1], Wenqing Zheng[†1], Han Meng[2], Siqi Liang[2], Anqing Chen[1], Hiroko H. Dodge[3,4], Jiayu Zhou[2], Zhangyang Wang[1] *
[1]University of Texas, Austin, [2]Michigan State University, [3]Massachusetts General Hospital, [4]Harvard Medical School
{jyhong,w.zheng,benjamin.c0427,atlaswang}@utexas.edu,
{menghan1,liangsi4,jiayuz}@msu.edu, hdodge@mgh.harvard.edu

ABSTRACT

Mild Cognitive Impairment (MCI) is a prodrome stage of Alzheimer's Disease and related dementias. Its detection is essential for early intervention and trial cohort enrichment. A recent clinical trial demonstrated that engaging in frequent cognitively stimulating conversations could be an effective strategy against social isolation and cognitive decline. However, the widespread deployment of such interventions faces challenges, particularly due to the need for trained human interviewers to conduct the conversations. In this paper, we study an innovative solution that uses an AI-based chatbot to replace human interviewers, thus greatly improving the accessibility of this therapeutic approach. We integrate the protocols used in the previous intervention trial into the automatic chatbot for stimulating cognitive functions through cognitively demanding, engaging, and user-friendly voice-based conversations. To evaluate the effectiveness, we create MCI digital twins–virtual replicas of MCI patients–offering a scalable and realistic assessment method. With the digital twins, we provide an end-to-end framework for evaluating and iterating the chatbot. Our experiments show the chatbot's proficiency in fostering natural conversations and its potential as a cost-effective, accessible tool in dementia intervention. A demonstration of our chatbot system is available at https://a-conect.github.io/.

1 INTRODUCTION

Mild Cognitive Impairment (MCI) is a prodrome stage of Alzheimer's Disease and related dementias. It manifests through a decline in memory, executive functions, and emotional expressiveness, predominantly affecting older adults (Petersen, 2004). It is estimated that 12% to 18% of individuals aged 60 or older live with MCI (Association et al., 2022). Within this age group, approximately 10% to 15% of MCI annually progress to dementia, with one-third transitioning to Alzheimer's disease within five years (Association et al., 2022; Petersen et al., 2014; Ward et al., 2013). Despite the high prevalence of MCI, developing effective intervention remains challenging due to its intricate underlying mechanisms (Cooper et al., 2013).

Social isolation and loneliness are modifiable risk factors of dementia (Evans et al., 2019; Donovan et al., 2017; Wilson et al., 2007a). Older adults experiencing limited social interactions or dissatisfaction with the quality and frequency of their social contacts are at increased risk of developing dementia (Dodge et al., 2014; Evans et al., 2019; Donovan et al., 2017; Fang et al., 2023). Remarkably, reducing social isolation could potentially prevent 4% of dementia cases, surpassing the prevention rate for diabetes (2%) and physical inactivity (2%), well-established risk factors (Livingston et al., 2020). This underscores the potential impact of preventing isolation and loneliness on dementia prevalence through enhanced social interactions; even modest postponements of cognitive decline can substantially impact the dementia prevalence (Brookmeyer et al., 1998). Recent studies advocate for

*Correspondence to Zhangyang Wang. † indicates equal contributions.

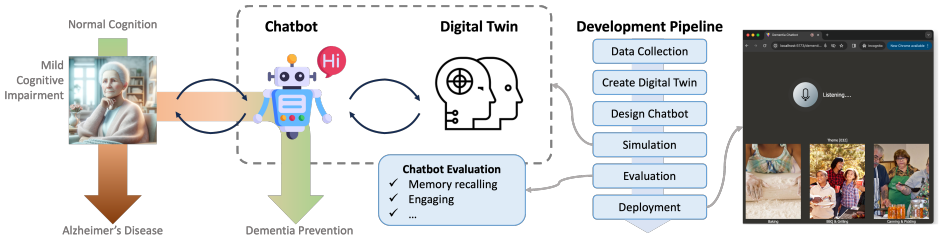


Figure 1: Overview of the paper including (1) a chatbot for early dementia intervention, (2) digital twins for evaluating chatbots, and (3) an end-to-end development pipeline from data to deployment.

enhancing the social connectedness of older adults through structured activities and regular video calls facilitated by trained interviewers (conversational staff) Yu et al. (2021); Dodge et al. (2023). Perry *et al.* highlighted the significance of social bridging and the cognitive stimulation from interactions with diverse social groups in preserving cognitive function (Perry et al., 2022). Recently, an internet-based conversational engagement project (I-CONECT) (NCT02871921), which is a behavioral intervention trial, provided frequent video chats to enhance social interactions among socially isolated older adults. The trial showed that frequent video-chat interactions could significantly boost social connectivity (Yu et al., 2023), improve psychological wellbeing (Yu et al., 2022), and slow down cognitive decline (Dodge et al., 2023; Yu et al., 2021; 2023). They prompted the conversations using daily themes and the associated pictures presented via the internet/webcam (www.i-conect.org).

Though the I-CONECT clinical trial results strongly motivate the development of conversation-based dementia interventions, these interventions are heavily dependent on the availability of trained interviewers. In (Yu et al., 2021; Dodge et al., 2023), interviewers underwent extensive training before engaging in actual video-chats with the trial participants. Therefore, their practical application is hindered by substantial challenges, such as the high costs associated with employing human interviewers and the inflexible scenarios in which communication can occur.

In response to these challenges, we propose an innovative AI-driven solution: employing an LLM-based chatbot as an alternative to human interviewers in dementia intervention programs. LLMs or Large Language Models (Sanh et al., 2022; Raffel et al., 2023; Ouyang et al., 2022) are state-of-the-art Machine Learning models that can conduct open conversations like humans (Bubeck et al., 2023). Compared to human counterparts, LLM-based chatbots are accessible round-the-clock at lower costs. As depicted in Fig. 1, the chatbot is specifically designed to offer cognitively demanding conversations, mimicking the human-interviewer-based previous I-CONECT trial, and providing patient-friendly interfaces to effectively engage with older adults (Dodge et al., 2023) . Furthermore, it ensures natural, effortless conversations through a soothing vocal interface, thereby accommodating older adults with varying cognitive capacities.

For a chatbot to be deployed to a larger population, it is crucial to validate the efficacy and adherence against the previous trial Therefore, we aim to provide an end-to-end development pipeline that evaluates and iterates the designs. A conventional evaluation involves enrolling numerous MCI patients in clinical studies to interact with the chatbot. However, the limited availability of patient data and the broad spectrum of user variations pose challenges to the empirical robustness of such studies (Hoang et al., 2023). Therefore, developing an alternative strategy for **scalable assessment** is essential, allowing for extensive and diverse interactive testing with the chatbot. In this study, we develop a simulation experiment in which the chatbot interacts with *patients' digital twins*. These digital twins are computational analogs that simulate patient characteristics and behaviors (Coorey et al., 2022; Goh et al., 2023). Compared to traditional human trials, digital twins offer the advantage of generating a vast array of conversations on demand, with highly customizable behaviors for varied testing scenarios. Our experiments confirm that the digital twin proficiently mimics the abnormal conversation patterns associated with MCI, providing a credible means to assess the chatbot.

As outlined in Fig. 1, our contributions are in four aspects. • We introduce a novel chatbot tailored for older adults (especially those with mild cognitive impairment), implementing protocols from the previous I-CONECT clinical trial. • We establish a development pipeline for the chatbot's design, scalable validation, and iterative improvement, incorporating innovative metrics for automated evaluation by LLMs or through human assessments. • We create digital twins as proxies for MCI patients, ensuring realistic and valid simulations. • Our experimental results affirm that the chatbot can adhere to the protocol, providing fluent conversational experiences and helping digital patients to

recall past experiences, express their thoughts, and be engaged in conversations. Plus, the chatbot is expected to illuminate and expedite preliminary experiments or proof-of-concept studies in MCI clinical trials.

2 Related Works

Using virtual humans in medical practices has attracted attention due to its low costs, efficiency, and convenience. This includes robotics and chatbots (Cruz-Sandoval & Favela, 2019; Huang et al., 2012; Saito et al., 2015; Nayak et al., 2023). For example, (Cruz-Sandoval & Favela, 2019) uses Socially Assistive Robotics (SAR) for interaction with old people with dementia. They designed a semi-autonomous robot, Eva, that can send personalized utterances, display emotions on Eva's face, present predefined activities (greetings, jokes, farewell), and search and play songs. Eva needs remote control by humans. In contrast, our chatbot is totally autonomous and can be accessed through web browsers. Closely related to our work, chatbots have been explored for old facing potential dementia risks (Huang et al., 2012). (Huang et al., 2012) uses a chatbot as a memory assistant companion that aims at increasing the social activities of socially isolated adults. (Saito et al., 2015) uses an embodied chatbot to analyze the attitudes of adults with dementia. In (Leo et al., 2019), a chatbot is used as a caregiver who will use organized interesting stories to help reconstruct cognitive functions. Comprehensive reviews have been provided in (Hocking et al., 2023; Ruggiano et al., 2021; Rampioni et al., 2021) about using chatbots in rehabilitation for adults with brain-related neurological conditions. The emergence of generative AI or LLMs revolutionizes chatbots with more natural language capabilities. The potential and limitations of AI in advancing psychological measurement, experimentation, and practice were recently discussed in (Demszky et al., 2023).

Another use case of virtual humans in health is a digital twin that simulates the behaviors of humans from patients' data. Previously, the digital twin has been used in healthcare to guide the diagnosis and treatment of cardiovascular disease (Coorey et al., 2022). (Elayan et al., 2021) applies the digital twins into an IoT healthcare system for heart problem detection. Digital twin technology also has the potential to enable personalized healthcare services (Sahal et al., 2022). As a data augmentation strategy, simulating conversations by virtual patients was also shown to be effective in enhancing predictive dementia modeling (Tang et al., 2020).

Our work is inspired by the advances both in dementia healthcare and the large language models (LLMs), with unique contributions in designing chatbots for early dementia intervention. To our knowledge, there is still an absence of an end-to-end framework for designing a chatbot that is applicable to old adults with mild cognitive impairment. Our work is a pioneer exploration toward filling the gap and yielding a viable chatbot for dementia intervention and beyond. In addition, we gain motivation from the effectiveness of conversational chatbots in MCI simulation (Tang et al., 2020) to utilize digital twins for validating our chatbots. With digital twins, we generate natural conversations for realistic simulations of cognitive impairment, facilitating the practical evaluation.

3 A-CONECT: AI-based CONversational Engagement for Clinical Trials

In this section, we present the essential design principles for developing a chatbot dedicated to dementia intervention. First, an AI-powered chatbot framework should follow the conversation strategies that have been practiced in the MCI clinical trial (I-CONECT (Dodge et al., 2023)). Second, recognizing the cognitive challenges faced by users, it is imperative to incorporate user-friendly designs for effectively engaging users. Based on these principles, we articulate a series of design choices tailored for old adults with MCI, that is, AI-based conversational engagement for the clinical trial project (A-CONECT), modified from the original human-based trial (I-CONECT).

Principles. Evidence from previous research shows that more participation in cognitive, leisure, and social activities correlates with a decreased chance of being diagnosed with dementia (Wilson et al., 2002; 2007b; Park et al., 2014). Similarly, (Park et al., 2014) found that sustained engagement in cognitively demanding, novel activities enhances old adults' memory function. These findings drive the conversational intervention for cognitive impairments. For instance, (Dodge et al., 2023) designed a behavioral intervention where a human interviewer regularly makes video chats with trial participants. The authors conduct a trial to explore if such regular conversation can mitigate the MCI symptoms. Our proposed method uses a chatbot to resemble the human interviewer in dementia intervention and, therefore, should follow the same protocol used in the clinical trial (Dodge et al.,

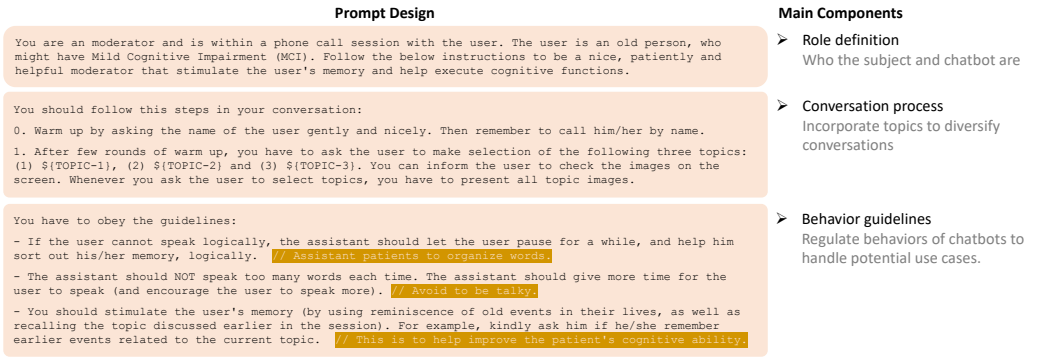


Figure 2: The prompt design for chatbot. Yellow-highlighted texts are comments explaining the functions of the specific instructions, which are not included in actual prompts.

2023).

P1: Cognitively-demanding Conversation Strategies. The primary hypothesis that has been verified in the existing clinical trial (Yu et al., 2021; Dodge et al., 2023) is "Increasing novel social stimuli may lead to cognitive reserve against neurodegeneration". That means the conversations should drive the users to stimulate and then strengthen their brain functions despite pathological insults. To be more specific, we break down the principle into two parts with concrete actionable goals. **P1.1: Novel chat experiences.** Previous research Park et al. (2014) pointed out that engagement in novel experiences can enhance cognitive function. Conversational interactions with diverse chat topics can be one of the possible novel experiences especially those living in social isolation (Dodge et al., 2023). **P1.2: Stimulating executive functions.** MCI patients suffer from impairment in executive functions, and the chats could help patients enhance their critical thinking and well-organized oral expression (Huang et al., 2012). In particular, we tried to stimulate memory(episodic and semantic memory) and executive functions which include conveying their thoughts effectively to others.

P2: Patient-Friendly Interaction. Other than the first principle, a user-friendly design is essential to create incentives for old adults to use the software, as outlined in (Dodge et al., 2023). A chatbot that is barely used will have negligible effects on the disease. The challenge lies in the fact that old adults have difficulty using PC/mouse/Internet and may not feel comfortable trusting virtual chatbots. We define two principles for promoting old adults' motivation for using the chatbot. **P2.1: User-friendly interface** that does not have any barriers for the oldest adults to use the software/hardware. Elder adults may have difficulty operating modern devices and may not like modern electronic characters instead of humans in conversation. Hence, it is imperative to streamline the software operations and enhance user interaction to achieve a more natural experience. **P2.2: Engaging conversations** that should encourage active patient participation. We aim to create an atmosphere where a subject is willing to talk to the chatbot actively. Traditional AI is often defined as an assistant that only responds on demand. In contrast, a chatbot in our scenario should actively ask questions that motivate users to think, talk, and then share thoughts.

Design 1: LLM-Driven Conversation. Recent advances show that Large Language Models can be trained to follow instructions embedded in language prompts (Ouyang et al., 2022). The ability leads to a new method, **prompt engineering**, that customizes the behaviors of LLMs simply in instructions. To incorporate the principles into the conversation, we use ChatGPT as the backbone to follow text instructions precisely. At a high level, we design the prompt to include *role definition*, *conversation process* and *behavior guidelines*, as demonstrated in Fig. 2. **(1) Role definition.** First, we define the role of the chatbot and the subject it will interact with during the conversation. We assume the subject will be either with normal cognition or MCI. **(2) Conversation process.** The process definition aims to specify the actions that should be taken during the conversation. The process exactly follows the I-CONECT trial (Dodge et al., 2023) to warm up the conversation and then discuss three topics. **(3) Behavior guidelines.** In these guidelines, we provide specific instructions to define the mindsets and conversation strategies that our chatbot should follow.

Based on these techniques, we integrate the principles (**P1**) into the prompt engineering. **(1) Diverse chat topics for novel chat experiences.** We aim to make the conversation to encourage users to do in-depth discussions that help preserve cognitive abilities. Inspired by (Dodge et al., 2023), we use diverse chat topics from a pre-defined pool that can be updated. We apply the idea into the

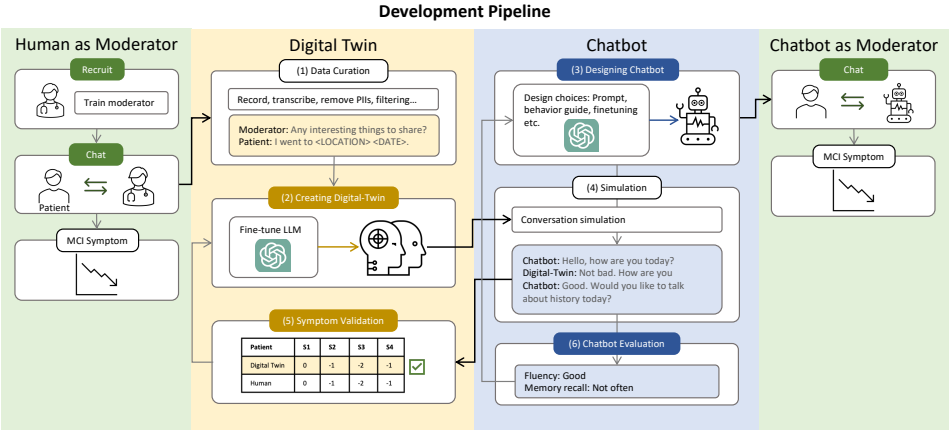


Figure 3: The development process of building a chatbot prioritized for MCI patients. We collect MCI patients' conversation data from the human clinical trial (I-CONECT) and build digital twins to simulate conversations with the designed chatbot.

conversation process as shown in Fig. 2. Each time, we randomly select a theme out of 150 ones developed in the previous trial (Dodge et al., 2023). Given a theme including three topics, used in the I-CONECT clinical trial (Dodge et al., 2023), the chatbot will ask the user to select one interesting topic. The 150 candidate themes and topics are collected by the I-CONECT trial, where topics are selected based on the interests of multiple stakeholders and potential participants. **(2) Stimulating cognitive functions by reminiscent and speech organization assistance.** Cognitive demanding activities could improve cognitive reserve (brain resilience against pathological insults). Therefore, we encourage the user to talk more than the chatbot. Reminiscences make the user think about their past experiences and motivate them to describe experiences in detail. However, MCI users may have difficulty organizing speech well when trying to describe their experiences in detail. As a result, the expression from MCI users tends to be restrained and non-informative, which can easily bring the conversation to an end. Taking these two factors into account, we instruct ChatGPT to ask some questions actively such that the users will be prompted to invoke past experiences and organize their expression properly. Moreover, our instruction, "*If the user cannot speak logically, the assistant should let the user pause for a while, and help him/she sort out his/her memory logically.*", provides the strategy when the users are not talking logically. **(3) Visual stimulation for informative context.** We also present images to guide the users for topic selection. The images will be presented automatically during the conversation to enhance the contextual richness of the dialogue. Each topic is associated with such a picture collected and copy-righted by the I-CONECT.

Design 2: Patient-Friendly Interaction The principle **P2** requires a joint design in the user interface, language generation, and software system. Especially, we need to consider the users as MCI patients. **(1) User-friendly interface.** Cognition-impaired patients could face challenges in using modern devices, especially those with complex operations. To streamline the software operations, we minimize the operations required to start the conversation by only one button click. Also, recognizing that typing might be challenging for old adults with motion diseases, like Parkinson's disease, we introduce voice conversation into our system. When activated by a button, the chatbot will be served automatically by listening and responding in voice, providing a conversation experience akin to interacting with a human. Last but not least, the entire service will be accessible on a website and is compatible with the widely-used Chrome browser on many terminal hardware, including laptops, pads, and smartphones.
(2) Engaging conversation. We aim to build an immersive chatbot system that is capable of listening to understand subjects' speech, responding with natural and proper language, and delivering a touching voice. *i) Avoid being talky.* As noted in previous research (Wei et al., 2023), LLMs tend to be talky, which is against our intention for stimulating users' speech and memory recalling. Therefore, we include the instruction, "*The assistant should NOT speak too many words each time. The assistant should give more time for the user to speak (and encourage the user to speak more).*". *ii) Avoid robotic-like conversation.* We instruct the chatbot to avoid robotic-like words that make users feel non-realistic and, therefore, less connected. In the role definition, we define the role that the chatbot will play and the subject it will face. Therefore, it will better understand the scenario and have proper mindsets. *iii) Friendly LLMs.* Last, we follow the common practice to make LLMs nice to humans by the instruction"*Follow the below instruction to be a nice, patient, and helpful interviewer*".

4 DEVELOPMENT PIPELINE

The language functionality lies at the core of the proposed conversational chatbot and therefore should be validated before deployment. In this section, we provide an end-to-end pipeline that evaluates and iterates the LLM-based design. The crux of the pipeline is an interactive test environment in which we can probe the capabilities of the chatbot. As outlined in Fig. 3, the pipeline includes 6 steps. (1) We curate data of conversations between patients and interviewers. (2) Then, we utilize the data to create digital twins for each patient. (3) We design the chatbot based on pre-trained LLMs. (4) We simulate conversations between the digital twin and the chatbot. (5) We evaluate the digital twins by examining the symptom similarity between the conversations by digital twins and those by human patients. (6) With the simulated conversations, we evaluate if the chatbot can realize the design principles. Based on the evaluation results, we can iterate or rank design choices in step (3). Finally, we deploy our chatbot in conversations with humans.

Data Curation. We use conversation data from the clinical trial, I-CONECT (NCT02871921). The objective of this clinical trial was to explore how frequent video chat conversations might affect cognitive abilities and the psychological health of individuals with MCI. The demographics of all participants are included in the appendix. The duration of each conversation session is approximately 30 minutes. To be specific, we call the MCI participants patients onwards. The I-CONECT data are transcribed from verbal conversations without punctuation and include privacy-sensitive information like dates and names. To protect participants' private information and improve data quality, we add punctuations using the Nemo BERT models (Harper et al.), and identify and remove private identifiable information (PII) using the Flair NLP toolbox (Akbik et al., 2019). conduct two preprocessing operations.

Creating Digital Twins for Patients. To refine the chatbot's heuristic design, the chatbot has to be run in vast conversations. However, recruiting human volunteers could be challenging, especially, for extensively interacting with the chatbot. With a limited number of conversations, the experimental evaluation could lack statistical significance. Toward a solid evaluation of the proposed chatbot, we propose using interactive virtual participants, namely digital twins (Venkatesh et al., 2022), to replace human participants. A digital twin is a model that distills the characteristics of a human participant and is able to generate infinitely many conversations. An effective digital twin should reproduce a subject's symptoms that are related to MCI. For example, some subjects may struggle to find proper words for converting ideas. In our study, the construction of digital twins includes two steps: *creating a digital twin chatbot* and *validating the MCI-related symptoms*.

(1) Fine-tuning LLMs to imitate MCI language patterns. For each patient, we use his/her conversations with interviewers to train a language model as *digital twin* such that the language model can generate distributional similar conversations. The fine-tuning is done along with a simple system prompt: "*You are an old person with Mild Cognitive Impairment. You will talk to an interviewer.*". Then, the LLM can be prompted to simulate the language patterns of the source patient. Every time, the LLM will only be tuned by a small set of conversations (10 in our experiments) such that the LLM can be periodically updated to synchronize the progression of cognitive functions in the future.

(2) Validating digital twin's symptoms. As a digital twin is designed to reproduce the MCI-related symptoms, we draw inspiration from existing literature to quantify the symptoms and, therefore, validate the reproducibility. Previously, Yeung *et al.* found that MCI/AD patients present recognizable speech characteristics in coherence and difficulty in finding words (Yeung et al., 2021) in structured conversations (Becker et al., 1994). They further showed that lexical and acoustic metrics can serve as the automatic language biomarkers for MCI/AD. The finding inspires us to seek language biomarkers (symptoms) in the semi-structured conversations in the I-CONECT dataset. As demonstrated by (Yeung et al., 2021), using multiple lexical scores was demonstrated to be correlated with speech capability and MCI/AD.

Based on the findings, we create **lexical profiles** for each patient including 7 lexical scores. We select 7 lexical scores that are most important for distinguishing MCI and NC participants by logistic regression. The full list and details of the selected lexical scores are enclosed in the appendix (see Table 4). A patient's profile, as shown in the left figure of Fig. 4, is constructed using selected lexical scores. We denote such lexical biases relative to the normal range of normal participants as a *lexical symptom*. The biases

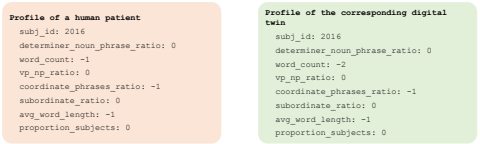


Figure 4: Comparison of patient's and the corresponding similar digital twin's profile.

are discretized as -2, -1, 0, 1, and 2 based on the variance bins. For example, -1 means that the lexical score is in the $[-2\sigma, -\sigma)$ range given the standard deviation σ of normal participants. A well-designed digital twin should also present a similar bias in simulation. With the lexical profiles, we justify a digital twin to be an effective replica of its source patient if its lexical profile is similar to the the patient's. For example, Fig. 4 presents a digital twin (right) that can reproduce similar lexical symptoms as the patient (left).

Testing the Chatbot in A Salable and Interactive Environment. The simulation of conversations with digital twins can be used to evaluate the chatbot and then refine it. In this section, we will provide a brief view of potential testing methods, gaining insights into the functionality of the chatbot in executing the designs in Section 3. As the interface- or system-based functions (diverse topics, visual stimulation, user-friendly interfaces, and data privacy) are guaranteed by design and program implementations, our evaluation will focus on the AI-based language functions.

The execution effectiveness of chatbot is verifiable through linguistic metrics or event detections. The goal, stimulating cognitive functions, includes two actionable designs: reminiscence and speech organization. In other words, we expect that the (virtual) patient will be stimulated to talk more about his/her past experience and will be more fluent in conversations. Therefore, we define the degree of reminiscence and speech organization as the frequency of experience recalling and fluent turns. The turns can be recognized by prompting LLMs, which has excellent zero-shot ability in semantic recognition (Wei et al., 2022). The other goal of engagement can be evaluated by how many words the patient has said in the whole conversation (similar to the strategy in (Yu et al., 2021)) and how frequently the chatbot asks questions to engage the patient. The detailed metric formulations will be elaborated in the experiment section.

5 EXPERIMENTS

In this section, we present empirical evaluations of how well our chatbot implements design principles for dementia intervention. Through our experiments, we use GPT3.5-turbo (ChatGPT) as the base LLMs both for the chatbot and the digital twins. For the digital twins, we use 10 conversations from each patient to fine-tune the ChatGPT on the OpenAI platform. Two LLMs are used in the conversation simulation. One is the digital twin fine-tuned on a patient. The other is the chatbot. We let the conversation last until one of the two parties says 'goodbye' or other similar sentences. We limit the conversation to 4096 tokens to reduce the costs.

Automatic Evaluation of Digital Twins. Digital twins are replicas of patients and are trained from patients' conversations with interviewers. Our experiments evaluate if a digital twin of a specific patient can reproduce similar symptoms in 20 simulated conversations per patient.

Results. In Fig. 4, we show that the profile measured on simulated data can match the real profile. We select a patient (No. 1007) that has significant biases in multiple lexical metrics. The patient tends to be conservative in conversations and therefore has fewer counts of words. He/she presents difficulty in using complex words and prefers shorter words. The preference is echoed by the negative one score of `avg_word_length` and the low ratio of coordinate phrases (`coordinate_phrases_ratio`). When the patient is simulated by our digital twin, we also observe similar symptoms. `coordinate_phrases_ratio` and `avg_word_length` exactly matches the scores of the real patient.

To demonstrate the generality, we extend the experiment to more patients and demonstrate the similarity between digital twins and their source patients in conversations. We use a distance metric to measure the similarity. Formally, we define the distance between two lexical profile vectors v_1 and v_2 as $\text{Dist}(v_1, v_2) = \|\text{sign}(v_1) - \text{sign}(v_2)\|_1$. We use ChatGPT without fine-tuning as a baseline for virtual participants. Two kinds of interviewers are evaluated in the conversation: ChatGPT and our chatbot. In Table 1, we show that our fine-tuned digital twins can effectively reproduce the lexical characteristics of the source patients. When facing different interviewers, the digital twin will present a slight difference in the distance metrics. It turns out that our chatbot can induce virtual patients to present better-aligned lexical symptoms.

Automatic Evaluation of Chatbots. With the built digital twins, we examine how chatbots perform in simulated conversations. Though our chatbot can also be used by adults without MCI, we focus on the simulation with digital twins of MCI patients, the more challenging scenario than that with NC users. To show the advantage of our design, we compare the proposed design with different

Table 1: Validating how well digital twins can resemble the patients' lexical behaviors in conversation with ChatGPT or our customized chatbot. The ChatGPT is also used as a participant baseline. Experiments are conducted with data from 9 human patients (with ID included per row) with MCI. Lexical distance represents the dissimilarity between the simulated participant's language and the referred participant's. On average, digital twins present the highest similarity with the patients in conversation with our chatbot.

Participant Interviewer	ChatGPT ChatGPT	ChatGPT Chatbot	Digital Twin ChatGPT	Digital Twin Chatbot
Average Lexical Distance (↓)	0.905	0.683	0.492	**0.397**

design choices. For models, our default model is *ChatGPT* which is the base model of ChatGPT for conversations. As a comparison, we fine-tuned GPT-3.5 (FT) on 10 conversation samples per participant, which is a popular practice to customize LLMs for specific tasks. Different from digital twins, fine-tuned chatbots learn the patterns of human interviewers instead of participants in the conversation samples. In experiments, we evaluate three design choices: (1) *ChatGPT*: We use the ChatGPT without any prompt or instructions; (2) *Ours*: The chatbot using full prompt proposed in this paper; (3) *Ours w/o Behavior Guidelines (BG)*: The proposed prompt is used but without behavior guidelines; (4) *Ours+FT*: In addition to the prompt, we further fine-tune the ChatGPT on the conversation data.

Here, we quantify and evaluate the below abilities of the chatbot. Our 6 automatic metrics, denoted as M1-6, provide fine-grained quantitative analysis of the instantiation of our principles. Essentially, the first three metrics (M1-3) evaluate if the chatbot can *stimulate cognitive functions* (P1.2). M4 evaluates if the chatbot can provide a fluent chatting experience and therefore contribute to a user-friendly interface (P2.1). M5 evaluates how the chatbot can engage the patient by asking questions (P2.2) rather than being a passive AI robot.

M1: Ability to help recall memory. The chatbot is instructed with "*If the user cannot speak logically, the assistant should let the user pause for a while, and help him sort out his/her memory, logically.*" We test whether the chatbot can actually follow the instructions. We use GPT-4 to justify whether a turn by the patient is talking about some experience. If it is an experience, we count it as a (memory) recalling turn. The metric is denoted as *Memory Recall Ratio* in a conversation.

$$\text{Recall Ratio} = \frac{\#\text{recalling turns}}{\#\text{turns in the conversation}}.$$

A higher ratio indicates more frequent memory recall.

M2: Ability to help organize words. The goal of the chatbot is to help patients rehabilitate their communication abilities, such as conveying a story clearly. We measure the fluency of patients' expressions in the conversation as a measurement of the goal conduction. Each conversation turn will be assessed by GPT-4 to determine its fluency. The frequency of fluent turns in a conversation is denoted as its *Fluency Ratio* (higher is better). The general fluency of a party (patient or interviewer) in the conversation is defined as

$$\text{Fluency Ratio} = \frac{\#\text{fluent turns}}{\#\text{turns in the conversation}}. \tag{1}$$

When calculating that for a patient, only responses from patients will be used to count the fluent turns and turns in conversation.

M3: Patient's Verbosity. According to the clinical protocol Yu et al. (2021), it is crucial to maximize patient engagement while minimizing the interviewer's verbosity. Therefore, we measure the ratio of the interviewer's tokens versus the patients' tokens in our design as a measurement of the patient's verbosity. An engaging patient will tend to talk as much as possible, implying higher patient's verbosity. We denote the criteria as *Patient/Interviewer Word Ratio*. For a given conversation, we compute the ratio as

$$\text{Pat/Int Word Ratio} = \frac{\#\text{words by the patient}}{\#\text{words by the interviewer}}.$$

M4: Chatbot Fluency. A qualified chatbot should be fluent in conversation. Similar to evaluating patients' fluency by Eq. (1), we use GPT-4 to judge if the chatbot is fluent in the language. The metric is computed based on the chatbot's turns.

M5: Chatbot Question Frequency. The question frequency reflects how actively the chatbot tends to involve the participants in conversations. Participants are often passive or conservative if not actively queried. Thus, higher question frequency should be preferred. Here, we use the ratio of questions in the conversation as a measurement:

$$\text{Question Ratio} = \frac{\#\text{turns of asking questions}}{\#\text{words by the interviewer}}.$$

A turn is defined as a question if a question mark appears in the turn.

The effectiveness of the Recall Ratio and Fluency Ratio relies on the accuracy of GPT-4 in detecting the recalling and fluent turns. To validate the effectiveness, we create two datasets for the two detection tasks, each of which includes 100 turns sampled from our generated data. The samples are manually labeled. The GPT-4 presents 63% accuracy in detecting the recalling events and 82% in detecting the fluent events.

Table 2: Averaged results and t-test on the difference between our chatbot and base ChatGPT as interviewers. Higher values are favored for all metrics and we highlight significant differences if p-value < 0.05. When talking with the proposed chatbot, digital twins (patients) are more likely to recall past experiences and talk more.

	ChatGPT	Ours	p-value
Patient Behaviors			
M1: Recall Ratio	0.271	**0.423**	**0.001**
M2: Fluency Ratio	0.462	0.421	0.212
M3: Pat/Int Word Ratio	0.425	**0.701**	**0.014**
Chatbot Behaviors			
M4: Fluency Ratio	0.999	0.999	0.859
M5: Question Ratio	0.213	**0.616**	**0.000**

Main Results. In Table 2, we extend the evaluation to 9 digital twins of patients and report the t-test results on the difference between ChatGPT and our chatbot as interviewers. Each digital twin is evaluated with 20 conversations. Across all patients, we observe significant changes in Recall Ratio, Patient/interviewer Word Ratio, and Question Ratio. The changes imply that the chatbot is actively involving the patients in conversation probably by asking more questions. The engagement leads to more cognitively demanding recalling of past experiences and more speeches by the patients. In terms of the fluency of patients or the interviewer, there are no significant differences. In other words, the chatbot is as fluent as the ChatGPT.

Cost comparison. LLM-based provides a cost-effective solution for dementia intervention and we provide estimated costs based on the market up to February 2024. According to OpenAI pricing, the cost for one thousand tokens is 0.03 U.S. dollars (USD) for input and 0.06 dollars for output by GPT-4. For a conversation including 4096 tokens (the limit of standard GPT-4), we assume there will be 20 turns of back and forth in a 30-minute conversation. Then, the estimated cost will be $0.5 \times 20 \times 4096 \times 0.06/1000 = 2.45$ USD assuming linear token accumulation per turn. According to ElevenLabs pricing, the cost will be 5 USD per month for 30,000 times of voice generation. If a patient completes 16 30-minute conversations (8 hours in total), the total cost will be $16 \times 2.45 + 5 = 44.2$ USD. In comparison, hiring a human for 8 hours demands $24 \times 8 = 192$ USD according to the virtual voice assistant wage in the US. Our chatbot can effectively reduce the cost to the 23.0% of hiring a human interviewer.

6 Conclusion

In the paper, we introduce an innovative AI-based chatbot as an early intervention for cognitive declines based on the accumulating evidence that social interactions could enhance cognitive reserve and resilience, thereby delaying the onset of dementia. It utilizes state-of-the-art Large Language Models. This solution addresses the limitations of human interviewer-based interventions by providing a scalable and cost-effective alternative. Our chatbot is designed with previously conducted efficacy-proven clinical trial protocols that could enhance social engagement and improve cognitive functions. We provide an end-to-end development pipeline together with empirical results demonstrating that the proposed LLM-based chatbot aligns with design principles. We envision our work can provide the foundation for future clinical trials, which compare the efficacy of chatbots with human interviewers in enhancing cognitive functions through social interactions.

Acknowledgments

This material is based in part upon work supported by the National Science Foundation under IIS-2212174 and IIS-1749940 and the National Institute of Aging: 1RF1AG072449, R01AG051628, R01AG056102.

REFERENCES

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pp. 54–59, 2019.

Alzheimer's Association et al. More than normal aging: Understanding mild cognitive impairment. *Alzheimer's Disease Facts and Figures*, 2022.

James T Becker, François Boiler, Oscar L Lopez, Judith Saxton, and Karen L McGonigle. The natural history of alzheimer's disease: description of study cohort and accuracy of diagnosis. *Archives of neurology*, 51(6):585–594, 1994.

Ron Brookmeyer, Sarah Gray, and Claudia Kawas. Projections of alzheimer's disease in the united states and the public health impact of delaying disease onset. *American journal of public health*, 88(9):1337–1342, 1998.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.

Claudia Cooper, Ryan Li, Constantine Lyketsos, and Gill Livingston. Treatment for mild cognitive impairment: systematic review. *The British Journal of Psychiatry*, 203(4):255–264, 2013.

Genevieve Coorey, Gemma A Figtree, David F Fletcher, Victoria J Snelson, Stephen Thomas Vernon, David Winlaw, Stuart M Grieve, Alistair McEwan, Jean Yee Hwa Yang, Pierre Qian, et al. The health digital twin to tackle cardiovascular disease—a review of an emerging interdisciplinary field. *NPJ digital medicine*, 5(1):126, 2022.

Dagoberto Cruz-Sandoval and Jesus Favela. Incorporating conversational strategies in a social robot to interact with people with dementia. *Dementia and geriatric cognitive disorders*, 47(3):140–148, 2019.

Dorottya Demszky, Diyi Yang, David S Yeager, Christopher J Bryan, Margarett Clapper, Susannah Chandhok, Johannes C Eichstaedt, Cameron Hecht, Jeremy Jamieson, Meghann Johnson, et al. Using large language models in psychology. *Nature Reviews Psychology*, pp. 1–14, 2023.

Hiroko H Dodge, Oscar Ybarra, and Jeffrey A Kaye. Tools for advancing research into social networks and cognitive function in older adults. *International psychogeriatrics*, 26(4):533–539, 2014.

Hiroko H Dodge, Kexin Yu, Chao-Yi Wu, Patrick J Pruitt, Meysam Asgari, Jeffrey A Kaye, Benjamin M Hampstead, Laura Struble, Kathleen Potempa, Peter Lichtenberg, et al. Internet-based conversational engagement randomized controlled clinical trial (i-conect) among socially isolated adults 75+ years old with normal cognition or mild cognitive impairment: Topline results. *The Gerontologist*, pp. gnad147, 2023.

Nancy J Donovan, Qiong Wu, Dorene M Rentz, Reisa A Sperling, Gad A Marshall, and M Maria Glymour. Loneliness, depression and cognitive function in older us adults. *International journal of geriatric psychiatry*, 32(5):564–573, 2017.

Haya Elayan, Moayad Aloqaily, and Mohsen Guizani. Digital twin for intelligent context-aware iot healthcare systems. *IEEE Internet of Things Journal*, 8(23):16749–16757, 2021.

Isobel EM Evans, Anthony Martyr, Rachel Collins, Carol Brayne, and Linda Clare. Social isolation and cognitive function in later life: a systematic review and meta-analysis. *Journal of Alzheimer's disease*, 70(s1):S119–S144, 2019.

Fang Fang, Tiffany F Hughes, Andrea Weinstein, Hiroko H Dodge, Erin P Jacobsen, Chung-Chou H Chang, Beth E Snitz, and Mary Ganguli. Social isolation and loneliness in a population study of cognitive impairment: The myhat study. *Journal of Applied Gerontology*, 42(12):2313–2324, 2023.

Ethan Goh, Bryan Bunning, Elaine Khoong, Robert Gallo, Arnold Milstein, Damon Centola, and Jonathan H. Chen. Chatgpt influence on medical decision-making, bias, and equity: A randomized study of clinicians evaluating clinical vignettes. *medRxiv*, 2023.

Eric Harper, Somshubra Majumdar, Oleksii Kuchaiev, Li Jason, Yang Zhang, Evelina Bakhturina, Vahid Noroozi, Sandeep Subramanian, Koluguri Nithin, Huang Jocelyn, Fei Jia, Jagadeesh Balam, Xuesong Yang, Micha Livne, Yi Dong, Sean Naren, and Boris Ginsburg. NeMo: a toolkit for Conversational AI and Large Language Models.

Bao Hoang, Yijiang Pang, Hiroko H Dodge, and Jiayu Zhou. Subject harmonization of digital biomarkers: Improved detection of mild cognitive impairment from language markers. In *PACIFIC SYMPOSIUM ON BIOCOMPUTING 2024*, pp. 187–200. World Scientific, 2023.

Judith Hocking, Candice Oster, Anthony Maeder, and Belinda Lange. Design, development, and use of conversational agents in rehabilitation for adults with brain-related neurological conditions: a scoping review. *JBI evidence synthesis*, 21(2):326–372, 2023.

Hung-Hsuan Huang, Hiroki Matsushita, Kyoji Kawagoe, Yoichi Sakai, Yuuko Nonaka, Yukiko Nakano, and Kiyoshi Yasuda. Toward a memory assistant companion for the individuals with mild memory impairment. In *2012 IEEE 11th International Conference on Cognitive Informatics and Cognitive Computing*, pp. 295–299. IEEE, 2012.

Pietro Leo, Grazia D'Onofrio, Daniele Sancarlo, Francesco Ricciardi, Michele De Petris, Francesco Giuliani, Giuseppe Ciarambino, Silvia Peschiera, Angelo Failla, Fabrizio Renzi, et al. Vita: Virtual trainer for aging. In *Ambient Assisted Living: Italian Forum 2017 8*, pp. 199–208. Springer, 2019.

Gill Livingston, Jonathan Huntley, Andrew Sommerlad, David Ames, Clive Ballard, Sube Banerjee, Carol Brayne, Alistair Burns, Jiska Cohen-Mansfield, Claudia Cooper, et al. Dementia prevention, intervention, and care: 2020 report of the lancet commission. *The Lancet*, 396(10248):413–446, 2020.

Ashwin Nayak, Sharif Vakili, Kristen Nayak, Margaret Nikolov, Michelle Chiu, Philip Sosseinheimer, Sarah Talamantes, Stefano Testa, Srikanth Palanisamy, Vinay Giri, et al. Use of voice-based conversational artificial intelligence for basal insulin prescription management among patients with type 2 diabetes: A randomized clinical trial. *JAMA Network Open*, 6(12):e2340232–e2340232, 2023.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35: 27730–27744, 2022.

Denise C Park, Jennifer Lodi-Smith, Linda Drew, Sara Haber, Andrew Hebrank, Gérard N Bischof, and Whitley Aamodt. The impact of sustained engagement on cognitive function in older adults: The synapse project. *Psychological science*, 25(1):103–112, 2014.

Brea L Perry, William R McConnell, Siyun Peng, Adam R Roth, Max Coleman, Mohit Manchella, Meghann Roessler, Heather Francis, Hope Sheean, and Liana A Apostolova. Social networks and cognitive function: An evaluation of social bridging and bonding mechanisms. *The Gerontologist*, 62(6):865–875, 2022.

Ronald C Petersen. Mild cognitive impairment as a diagnostic entity. *Journal of internal medicine*, 256(3):183–194, 2004.

Ronald C Petersen, Barbara Caracciolo, Carol Brayne, Serge Gauthier, Vesna Jelic, and Laura Fratiglioni. Mild cognitive impairment: a concept in evolution. *Journal of internal medicine*, 275 (3):214–228, 2014.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer, 2023.

Margherita Rampioni, Vera Stara, Elisa Felici, Lorena Rossi, and Susy Paolini. Embodied conversational agents for patients with dementia: thematic literature analysis. *JMIR mHealth and uHealth*, 9(7):e25381, 2021.

Nicole Ruggiano, Ellen L Brown, Lisa Roberts, C Victoria Framil Suarez, Yan Luo, Zhichao Hao, and Vagelis Hristidis. Chatbots to support people with dementia and their caregivers: systematic review of functions and quality. *Journal of medical Internet research*, 23(6):e25006, 2021.

Radhya Sahal, Saeed H Alsamhi, and Kenneth N Brown. Personal digital twin: a close look into the present and a step towards the future of personalised healthcare industry. *Sensors*, 22(15):5918, 2022.

Naoko Saito, Shogo Okada, Katsumi Nitta, Yukiko Nakano, and Yuki Hayashi. Estimating user's attitude in multimodal conversational system for elderly people with dementia. In *2015 AAAI spring symposium series*, 2015.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Tali Bers, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M. Rush. Multitask prompted training enables zero-shot task generalization, 2022.

Fengyi Tang, Ikechukwu Uchendu, Fei Wang, Hiroko H Dodge, and Jiayu Zhou. Scalable diagnostic screening of mild cognitive impairment using ai dialogue agent. *Scientific reports*, 10(1):5732, 2020.

Kaushik P Venkatesh, Marium M Raza, and Joseph C Kvedar. Health digital twins as tools for precision medicine: Considerations for computation, implementation, and regulation. *NPJ digital medicine*, 5(1):150, 2022.

Alex Ward, Sarah Tardiff, Catherine Dye, and H Michael Arrighi. Rate of conversion from prodromal alzheimer's disease to alzheimer's dementia: a systematic review of the literature. *Dementia and geriatric cognitive disorders extra*, 3(1):320–332, 2013.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022.

Jing Wei, Sungdong Kim, Hyunhoon Jung, and Young-Ho Kim. Leveraging large language models to power chatbots for collecting user self-reported data. *arXiv preprint arXiv:2301.05843*, 2023.

Robert S Wilson, Carlos F Mendes De Leon, Lisa L Barnes, Julie A Schneider, Julia L Bienias, Denis A Evans, and David A Bennett. Participation in cognitively stimulating activities and risk of incident alzheimer disease. *Jama*, 287(6):742–748, 2002.

Robert S Wilson, Kristin R Krueger, Steven E Arnold, Julie A Schneider, Jeremiah F Kelly, Lisa L Barnes, Yuxiao Tang, and David A Bennett. Loneliness and risk of alzheimer disease. *Archives of general psychiatry*, 64(2):234–240, 2007a.

Robert S Wilson, Paul A Scherr, Julie A Schneider, Yuxiao Tang, and David A Bennett. Relation of cognitive activity to risk of developing alzheimer disease. *Neurology*, 69(20):1911–1920, 2007b.

Anthony Yeung, Andrea Iaboni, Elizabeth Rochon, Monica Lavoie, Calvin Santiago, Maria Yancheva, Jekaterina Novikova, Mengdan Xu, Jessica Robin, Liam D Kaufman, et al. Correlating natural language processing and automated speech analysis with clinician assessment to quantify speech-language changes in mild cognitive impairment and alzheimer's dementia. *Alzheimer's research & therapy*, 13(1):109, 2021.

Kexin Yu, Katherine Wild, Kathleen Potempa, Benjamin M. Hampstead, Peter A. Lichtenberg, Laura M. Struble, Patrick Pruitt, Elena L. Alfaro, Jacob Lindsley, Mattie MacDonald, Jeffrey A. Kaye, Lisa C. Silbert, and Hiroko H. Dodge. The internet-based conversational engagement clinical trial (i-conect) in socially isolated adults 75+ years old: randomized controlled trial protocol and covid-19 related study modifications. *Frontiers in digital health*, 3:714813, 2021.

Kexin Yu, Benjamin M Hampstead, Katherine Wild, Lisa C Silbert, and Hiroko H Dodge. Examining i-conect intervention effect on psychosocial wellbeing. *Alzheimer's & Dementia*, 18:e059654, 2022.

Kexin Yu, Chao-Yi Wu, Lisa C Silbert, Jeffrey A Kaye, and Hiroko H Dodge. I-conect intervention effects on weekly time spent outside of home and social contacts among socially isolated older adults with normal cognition and mild cognitive impairment. *Alzheimer's & Dementia*, 19:e077984, 2023.

Table 3: Demographics and conversation statistics of participants in the I-CONECT project.

Variable	All (n=74)	NC (n=36)	MCI (n=38)
Age	80.7 ± 4.6	79.7 ± 3.9	81.7 ± 5.0
Gender (% women)	71.6	77.8	65.8
Years of education	15.2 ± 2.5	15.4 ± 2.5	15.1 ± 2.5
Number of conversations	91.5 ± 37.2	92.4 ± 35.8	90.7 ± 38.4

Table 4: Selected lexical scores.

Score Name	Definition
determiner_noun_phrase_ratio	Use of determiner noun phrases
word_count	Count of words
vp_np_ratio	Use of verb phrases with noun phrases
coordinate_phrases_ratio	Use of coordinate phrases
subordinate_ratio	Ratio of the count of subordinate versus coordinate
avg_word_length	Average length (number of letters) of words
proportion_subjects	Ratio of subjects w.r.t. all words

A Experiment Details

Dataset. The objective of this clinical trial was to explore how frequent video chat conversations might affect cognitive abilities and the psychological health of individuals who are 75 years old or older. The trial includes 74 participants who were randomly assigned to the experimental group (as opposed to the control group who received only weekly phone check-in calls in the trial). The experimental group has collectively participated in 6771 conversation sessions. Of these participants, 36 are cognitively normal (NC), and 38 have mild cognitive impairment (MCI). Demographics and conversation statistics of all participants are included in Table 3. The I-CONECT project recruits a diverse group of participants with different ages, genders, and years of education.

Selection of lexical scores. In Table 4, we elaborate on the definitions of selected lexical scores. We select the 7 scores from the full list of lexical scores provided in Yeung et al. (2021). To select the lexical scores, we first discretize the lexical scores based on the σ partitions. Then, we train a logistic regression model on all participants' data yielding an AUC of 82.5%. We select lexical scores with the largest absolute weights, which imply high importance on the prediction.

B Additional Experiments

Ablation Study. In Table 5, we compare the chatbot in conversations with a digital twin of the No. 1007 participant who has the most significant lexical biases. All scores are reported as an average of 20 conversations. Our main findings are as follows. (1) Our chatbot can effectively increase participants' recall ratio, speech fluency but lower aprticipants' chance to speak. This can be caused by the chatbot trying to speak more and ask more under prompts for better conversation context. (2) Though fine-tuning strengthens the spontaneous speech ability that encourages participants to speak more, it significantly lowers the speech fluency of both chatbots and patients. Thus, fine-tuning can lead to inferior conversation quality. (3) Behavior guidance helps with memory recall significantly by three times and increases the speech fluency of both participants and chatbots. (4) A higher question ratio may reduce the participants' speech ratio. This is because participants are more passive listeners rather than active speakers. An active chatbot tends to engage participantsthrough frequent questions.

Simulated conversations. In Fig. 6, we compare the simulated dialog against the real ones. The simulated interviewer has achieved important goals of psychological guidance and healing, such as positive emotional regulation ("*we really are blessed to be able to enjoy it*"), positive cognitive activation and recall ("*Actually, that's what? Do you want your family? Did your family and you used to do something more festive?*"), and active listening skills ("*Yeah, it gets loud. I can imagine*").

Table 5: Ablation study of the proposed chatbot in conversations with a representative digital twin. BG represents the Behavior Guide in the prompt. Ours+FT uses a ChatGPT fine-tuned with conversation data as the base model with our prompt. Higher values are favored for all metrics.

Metric	ChatGPT	Ours w/o BG	Ours	Ours+FT
Patient Behaviors				
Recall Ratio	0.130	0.107	**0.479**	0.215
Fluency Ratio	0.605	0.538	**0.656**	0.146
Pat/Mod Word Ratio	0.368	0.307	0.580	**0.762**
Chatbot Behaviors				
Fluency Ratio	0.998	0.993	**0.999**	0.138
Question Ratio	0.293	0.643	**0.728**	0.500

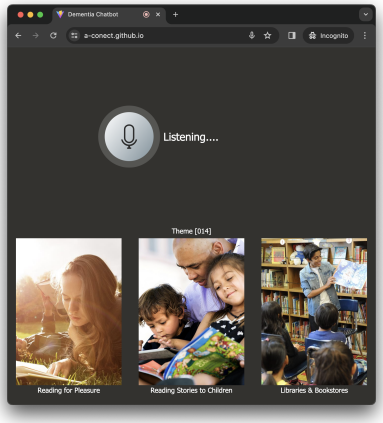


Figure 5: Our chatbot is based on the Chrome browser extension that can be run on any Chrome-compatible platform. The conversation will be automatically detected, and voice response will be generated in real time. The demo system is available at https://a-conect.github.io/.

How long can the conversation last? The chatbot is expected to talk with a user for as long time as possible. Technically, the challenge comes from the context limit of ChatGPT and the actual tokens to be expected. In clinical experiments, the conversation is expected to be about 30 min. Considering a human can speak 110-150 words (or tokens approximately) per minute, this falls within the range of 3000-4500 tokens, which is within the limit of ChatGPT. But the actual execution length depends on the speech generation and the tone.

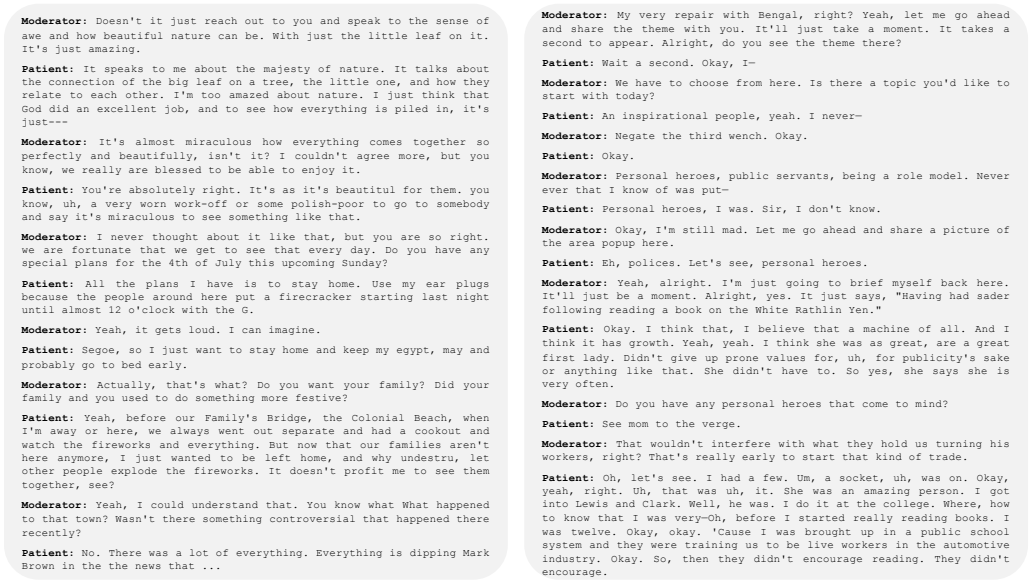
B.1 Designing Human Evaluation of Chatbots

In addition to the automatic evaluation, we designed an evaluation framework to get human feedback, which can extend our experiments in the future. At the high level, we evaluate the three aspects of the chatbot. (1) Fluency: Can the chatbot speak fluently?. (2) Memory recalling: How likely participants will recall memory in conversation? (3) Satisfactory: Is the conversation satisfactory with the chatbot?

The potential participants of the evaluation are general anonymous volunteer users of our chatbot on the Internet. Note that the users are not necessarily the subjects with MCI or patients. We designed a questionnaire that explores different dimensions of our principle.

The evaluation is based on a questionnaire. The questionnaire includes questions falling in the below categories.

- *Interests on Topics*: How long does your every conversation with the chatbot last? How many conversations have you completed? How much are you interested in the topics in general?
- *Experience Sharing Willingness*: Are you willing to share your personal experience in the conversation?



Moderator: Doesn't it just reach out to you and speak to the sense of awe and how beautiful nature can be. With just the little leaf on it. It's just amazing.

Patient: It speaks to me about the majesty of nature. It talks about the connection of the big leaf on a tree, the little one, and how they relate to each other. I'm too amazed about nature. I just think that God did an excellent job, and to see how everything is piled in, it's just---

Moderator: It's almost miraculous how everything comes together so perfectly and beautifully, isn't it? I couldn't agree more, but you know, we really are blessed to be able to enjoy it.

Patient: You're absolutely right. It's as it's beautitul for them. you know, uh, a very worn work-off or some polish-poor to go to somebody and say it's miraculous to see something like that.

Moderator: I never thought about it like that, but you are so right. we are fortunate that we get to see that every day. Do you have any special plans for the 4th of July this upcoming Sunday?

Patient: All the plans I have is to stay home. Use my ear plugs because the people around here put a firecracker starting last night until almost 12 o'clock with the G.

Moderator: Yeah, it gets loud. I can imagine.

Patient: Segoe, so I just want to stay home and keep my egypt, may and probably go to bed early.

Moderator: Actually, that's what? Do you want your family? Did your family and you used to do something more festive?

Patient: Yeah, before our Family's Bridge, the Colonial Beach, when I'm away or here, we always went out separate and had a cookout and watch the fireworks and everything. But now that our families aren't here anymore, I just wanted to be left home, and why undestru, let other people explode the fireworks. It doesn't profit me to see them together, see?

Moderator: Yeah, I could understand that. You know what What happened to that town? Wasn't there something controversial that happened there recently?

Patient: No. There was a lot of everything. Everything is dipping Mark Brown in the the news that ...

Moderator: My very repair with Bengal, right? Yeah, let me go ahead and share the theme with you. It'll just take a moment. It takes a second to appear. Alright, do you see the theme there?

Patient: Wait a second. Okay, I—

Moderator: We have to choose from here. Is there a topic you'd like to start with today?

Patient: An inspirational people, yeah. I never—

Moderator: Negate the third wench. Okay.

Patient: Okay.

Moderator: Personal heroes, public servants, being a role model. Never ever that I know of was put—

Patient: Personal heroes, I was. Sir, I don't know.

Moderator: Okay, I'm still mad. Let me go ahead and share a picture of the area popup here.

Patient: Eh, polices. Let's see, personal heroes.

Moderator: Yeah, alright. I'm just going to brief myself back here. It'll just be a moment. Alright, yes. It just says, "Having had sader following reading a book on the White Rathlin Yen."

Patient: Okay. I think that, I believe that a machine of all. And I think it has growth. Yeah, yeah. I think she was as great, are a great first lady. Didn't give up prone values for, uh, for publicity's sake or anything like that. She didn't have to. So yes, she says she is very often.

Moderator: Do you have any personal heroes that come to mind?

Patient: See mom to the verge.

Moderator: That wouldn't interfere with what they hold us turning his workers, right? That's really early to start that kind of trade.

Patient: Oh, let's see. I had a few. Um, a socket, uh, was on. Okay, yeah, right. Uh, that was uh, it. She was an amazing person. I got into Lewis and Clark. Well, he was. I do it at the college. Where, how to know that I was very—Oh, before I started really reading books. I was twelve. Okay, okay. 'Cause I was brought up in a public school system and they were training us to be live workers in the automotive industry. Okay. So, then they didn't encourage reading. They didn't encourage.

Figure 6: Comparison of the simulated conversation (left) and a real conversation (right).

- *Easiness to exchange thoughts*: Do you think it is easy to exchange your thoughts with the chatbot in the conversation?
- *How well/long engaged?*: How much are you engaged in the conversation? Here are some examples. If you talk a lot and focus on the conversation, then you are well-engaged. If you barely talk and tend to leave the conversation early, then you are not engaged.How can the chatbot improve to better engage you in conversation?
- *Easiness of Interaction*: Do you find any difficulty in using the software?
- *Chatbot Fluency*: Do you think the chatbot is fluent in speech and language?
- *Chatbot Friendliness*: Do you think the chatbot is friendly in language and tone?
- *Improper Expressions*: Do you see any improper expressions by the chatbot that make you uncomfortable?