# Designing Scaffolding Strategies for Conversational Agents in Dialog Tasks of Neurocognitive Disorder Screening

Jiaxiong Hu
hujx@ust.hk
The Hong Kong University of Science and Technology
Hong Kong SAR, China

Junze Li
junze.li@connect.ust.hk
The Hong Kong University of Science and Technology
Hong Kong SAR, China

Yuhang Zeng
Dongjie Yang
Danxuan Liang
yzengay,dyangaj,dliangac@connect.ust.hk
The Hong Kong University of Science and Technology
Hong Kong SAR, China

Helen Meng
hmmeng@se.cuhk.edu.hk
The Chinese University of Hong Kong
Hong Kong SAR, China

Xiaojuan Ma*
mxj@cse.ust.hk
The Hong Kong University of Science and Technology
Hong Kong SAR, China

## ABSTRACT

Regular screening is critical for individuals at risk of neurocognitive disorders (NCDs) to receive early intervention. Conversational agents (CAs) have been adopted to administer dialog-based NCD screening tests for their scalability compared to human-administered tests. However, unique communication skills are required for CAs during NCD screening, e.g., clinicians often apply scaffolding to ensure subjects' understanding of and engagement in screening tests. Based on scaffolding theories and analysis of clinicians' practices from human-administered test recordings, we designed a scaffolding framework for the CA. In an exploratory wizard-of-Oz study, the CA empowered by ChatGPT administered tasks in the Grocery Shopping Dialog Task with 15 participants (10 diagnosed with NCDs). Clinical experts verified the quality of the CA's scaffolding and we explored its effects on task understanding of the participants. Moreover, we proposed implications for the future design of CAs that enable scaffolding for scalable NCD screening.

## CCS CONCEPTS

• **Human-centered computing → Human computer interaction (HCI)**.

## KEYWORDS

Health, Aging, Scaffolding, Conversational Agent, Neurocognitive Disorder Screening

*Corresponding author


*CHI '24, May 11–16, 2024, Honolulu, HI, USA*

ACM ISBN 979-8-4007-0330-0/24/05
https://doi.org/10.1145/3613904.3642960

**ACM Reference Format:**
Jiaxiong Hu, Junze Li, Yuhang Zeng, Dongjie Yang, Danxuan Liang, Helen Meng, and Xiaojuan Ma. 2024. Designing Scaffolding Strategies for Conversational Agents in Dialog Tasks of Neurocognitive Disorder Screening. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24), May 11–16, 2024, Honolulu, HI, USA.* ACM, New York, NY, USA, 21 pages. https://doi.org/10.1145/3613904.3642960

## 1 INTRODUCTION

Neurocognitive disorders (NCDs), such as Alzheimer's disease, are common among older adults [5]. NCDs not only impair the quality of life of patients and their families but also increase the burden on caregivers and healthcare systems [19] and impose a financial strain on society [52]. However, NCD-associated symptoms can be managed through early diagnosis and preventive interventions [1, 17]. It is, therefore, crucial to develop a scalable approach for routine screening of cognitive impairment [53]. Currently, NCD screening and diagnostic tests, such as MoCA [50], are primarily administered in person by clinical professionals [67, 82]. Unfortunately, the scalability of human-administered approaches is limited by reduced mobility of some older adults, clinical shortages, and inter-rater variability [42].

To overcome these limitations of traditional NCD screening methods, researchers are developing alternative solutions. As communication impairments are often indicative of NCDs, machine learning algorithms that detect and analyze anomalies in speech, language processing, and language production offer a new lens for NCD screening [18, 34, 48]. However, a practical challenge of this potential method for early NCD assessment is the need for reliable, longitudinal collection of speech data from users. Previous studies suggested that having an automatic conversational agent (CA) chat with users regularly could be an affordable means to capture data [3]. Specifically, if a CA can routinely conduct dialog tasks for NCD screening, such as the recently developed Grocery Shopping Dialog Tasks [21], the user's changes in cognitive abilities can be observed over time with standard assessment and speech data can be gathered for objective analytics. This could facilitate early detection of cognitive decline.

While existing NCD screening administered by CA systems typically begins with a simple dialogue strategy [3], communicating with people who may experience reduced cognitive abilities often requires special conversational skills. For instance, simplifying sentences and rephrasing is important [15]. Prior work has shown that equipping a CA with back-channeling skills could ensure a natural and smooth NCD screening process [14]. However, there are many other more complex conversational skills and techniques that humans adopt in communication with NCD patients [15]. Among them, scaffolding – a process that keeps NCD patients engaged and facilitates activity understanding – is considered critical to support NCD patients' daily activities through dialogues [20, 25, 26, 32].

Scaffolding from conversational partners has been shown to help sustain the involvement of people with NCDs in joint activities [26, 46, 61]. Many NCD screening tests can be considered joint activities between the tester and the subject. However, it is common for subjects to give up early in these tests [14] and scaffolding may play a helpful role in keeping them engaged. Compared to the learning [38, 77] and decision-making [57] domains, designing CA's scaffolding strategies for dialog tasks of NCD screening is still under-explored and bears unique challenges. First, the target users are people with potential NCDs, suggesting that the CA design should be rooted in scaffolding theories and practices for these people particularly. Second, the objective of the CA is to facilitate the user's task understanding during the screening and their exhibition of current cognitive abilities, rather than providing direct answers or teaching them new knowledge. This raises two research questions. **RQ1**: How can a conversational agent for NCD screening be designed to exercise clinical professionals' scaffolding strategies? **RQ2**: What are the effects of the conversational agent's scaffolding strategies on the user's task understanding during the NCD screening?

To address RQ1, we proposed a design process for the CA that administered the dialog task for NCD screening. To identify clinicians' scaffolding strategies, we first analyzed the conversation recordings between clinicians and subjects during the dialog task of NCD screening, i.e., the GSDT [21]. The GSDT assesses multiple cognitive functions by requiring a subject[1] to narrate the process of purchasing ingredients on the floorplan of a grocery store for cooking a dish. Rooted in theories of scaffolding for NCD patients and the Zone of Proximal Development (ZPD), we proposed a framework for designing the CA's scaffolding to imitate human skills. Based on the proposed framework, we used an iterative design approach [24] to implement a prototype of a semi-automatic CA system that provided scaffolding. We automated the core "scaffolding strategy retrieval and reasoning" feature in CA with the support of ChatGPT (gpt-3.5-turbo) [2], a large language model with great potential in public health [8]. As the existing APIs for speech recognition and production are not of satisfactory quality in our context, especially when the local dialect is a comparatively low-resource language, we involved a human wizard to manage these functions.

To answer RQ2, we used this prototype CA as an exploratory research probe to understand the effects of the CA's scaffolding. Following the evaluation method employed in previous HCI literature [24], we conducted a "first-use" study, inviting 15 participants (age=55-76) to complete the GSDT administered by our CA. The goal of this study was to check the appropriateness of the CA's scaffolding and to explore its effects on the user's understanding of the dialog task. Note that we did not conduct a comparative study against a CA without scaffolding. This is because such a baseline system is not capable of handling challenges caused by users' cognitive conditions such as misunderstanding of the task, and it is likely to fail the whole dialog task when those situations occur. We plan to carry out a comparison with human-administered tests when our proposed CA is fully automated in the future.

Clinical experts from a local hospital verified the appropriateness of 89.45% of the CA's scaffolding adoption and 68.25% of ChatGPT's scaffolding recommendation. Participants also subjectively perceived the helpfulness of the CA's scaffolding in task understanding and evaluated the CA with generally good usability. Interestingly, we found that the CA's scaffolding use declined as the task progressed, which was consistent with the human testers' scaffolding. This suggested that the CA's scaffolding changed the ZPD status of participants and facilitated their task understanding. To the best of our knowledge, this research could be among the first to explore a CA's scaffolding strategies in a setting of dialog tasks for cognitive assessment. Our key contributions are threefold. First, we propose a process for designing the CA's scaffolding strategies for NCD screening. Second, we explore the effects of the CA's scaffolding on the task understanding of the user during NCD screening through an exploratory study. Third, we propose implications for the future design of CAs that enable scaffolding for scalable NCD screening.

## 2 RELATED WORKS

### 2.1 Conversational Agents for Populations with Neurocognitive Disorders and Their Caregivers

Conversational agents (CAs) have been adapted to serve users with neurocognitive disorders (NCDs) and their caregivers due to CAs' accessibility and interactivity [9, 14, 60, 73, 74]. Robin, for example, provided voice prompts and task guidance for individuals with NCDs to improve their independence in daily tasks [10]. Anne was designed with a female avatar to engage users with NCDs naturally in the dialogues [63, 64]. Louise, another CA, could monitor the attention of older adults with cognitive impairment [74] and was evaluated with good usability and acceptance when supporting them in simple tasks such as taking medicine and measuring blood pressure [73]. Additionally, AlzBot enhanced the socialization of patients with NCDs by providing them with a 24-hour online chatting feature [80].

Caregivers of people with NCDs often report higher levels of burden, stress, depression, and poor health conditions due to a lack of caregiving support, including emotional support and knowledge support, such as strategies to cope with the complexity of dementia care, especially under the COVID-19 pandemic [5]. Prior work utilized the CA to support the dyadic interactions in daily living between older adults with NCDs and their caregivers [85]. Additionally, CAs on public platforms, such as websites and application stores, also provided support for people with NCDs and their

[1] In this paper, we use the term "subject" generally for those who take part in any NCD screening tests administered by human testers, and "participant" for those who are recruited in our study.
[2] https://openai.com/blog/chatgpt

caregivers, mainly focusing on memory, reminiscence, and education about NCDs [60]. Many CAs have been designed to support the daily activities of NCD patients and their caregivers, however, CAs for NCD screening, which may aid early diagnoses and timely preventive interventions, are under exploration.

## 2.2 Automating Tests for Neurocognitive Disorder Screening

There are many valid neurocognitive tests for NCD screening [4, 11, 41]. Comprehensive tests that cover multiple key cognitive functions have been widely used in clinical practice, such as the Montreal Cognitive Assessment (MoCA) [50], the Addenbrooke's Cognitive Examination Revised (ACE-R) [47], the Short Test of Mental Status (STMS) [33], and The Grocery Shopping Dialog Task (GSDT) [21]. Among them, GSDT further covers non-verbal episodic memory, such as spatial memory, and it is developed for Hong Kong older residents [21].

Currently, these tests for NCD screening are generally conducted in person by professional clinicians [67, 82]. Though digital tools, such as tablets, touchscreens, and smartphone applications, for NCD screening tests, have been developed to release the clinicians' burden [28, 43, 45, 51, 62, 72], many of them still need to be applied in a supervised manner. Self-administered tests, such as SATURN [7], showed the feasibility of application in a fully independent manner [66]. However, language functions such as verbal fluency were not sufficiently covered in these tests [7]. As language impairment is an essential marker of NCDs, many machine-learning approaches have been developed to identify or detect NCDs in the user's speech [18, 34, 48]. The user's speech could be either elicited during NCD screening tests such as the corpus DementiaBank [6] or spontaneous utterances such as the dataset in the ADReSS challenge [44].

Data collection from users is necessary for automated NCD screening. Conversational agents (CAs) such as intelligent virtual agents and automatic interviewers were adapted to collect speech data from older adults [3, 48]. There are challenges to building CAs to collect speech from people with NCDs. Elderly speech displays significant differences from that of non-aged adults due to increased voice perturbations, articulatory imprecision, reduced speaking rates, increasing dysfluencies, and decreasing intensities, which requires effective automatic speech recognition [82]. Further, people with NCDs need special support from the CA, such as back-channeling [14] and dialog act recognition [16]. Scaffolding, on the other hand, is also considered crucial support for people with NCDs, especially during everyday collaborative activities with their caregivers [20, 25, 26, 32]. However, little is known about how to design a CA that can scaffold people with NCDs, and automating NCD screening tests with CAs is still under exploration [49].

## 2.3 Scaffolding in NCD Patients' Daily Life

Scaffolding is originally a psychological process between the "expert" and the "novice" [78]. The process allows a novice to accomplish a task or reach a goal that would be too difficult for them to do on their own. The scaffolding process involves an expert guiding and supporting the novice by taking control of the more challenging aspects of the task so that the novice can focus on and complete the parts that they are capable of doing. The collaborators in the process were described as an expert and a novice, which could be an adult and a child [78], or a caregiver and a person with NCDs as well [32]. Many joint or collaborative activities with NCD patients involve scaffolding, such as cooking and having a meal [20, 27]. For example, the caregiver may provide scaffolding to help the patient maintain the ceremonial order and become active during the meal or other social activities [20]. The conversational partner also scaffolds the person with NCDs to support their well-being and social identity [32, 61]. Scaffolding with narrowing questions e.g., yes/no questions, helps include and sustain the involvement of the people with NCD [46]. In practice, scaffolding may take place at three levels [26], as shown in Table 1, which motivates this study as well.

**Table 1: Scaffolding for NCD patients takes place at three levels [26].**

| Scaffolding Level | Description |
|---|---|
| Activity | Help to frame, reframe, and remind about the ongoing or planned activity. |
| Action | Help to construct and perform joint actions. |
| Repair | Help to identify and solve problems in joint actions. |

Another scaffolding-related theory is the Zone of Proximal Development (ZPD) [71]. The ZPD is the zone that lies between what a learner can do on their own and what they are unable to do even with assistance as shown in Fig. 1. This is the range where the learner can perform with the guidance of a teacher or a peer who has greater knowledge or expertise, also known as a "more knowledgeable other." Also, the ZPD is considered a counterbalance to the zone of actual development (ZAD) [35]. ZPD-based approaches were constructed to improve prose memory and social interaction of older adults with NCDs [65]. Assistant technology for NCD patients was considered working in the ZPD to provide "just enough" support to allow them to carry out activities without feeling like they had lost control or repeatedly failed [75].

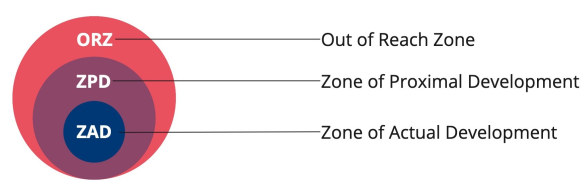


**Figure 1: The Zone of Proximal Development theory [71].**

In the scenario of NCD screening, many tests involve the tester's scaffolding for the subject. For example, the tester was found to use follow-up questions or encouragement to help the subject talk more in the language fluency test of MoCA [14]. In a more comprehensive and complex screening test such as GSDT [21], the subject may need scaffolding for locating and navigating to complete the spatial cognition test. Inspired by the previous work that looked into the tester's behaviors during NCD screening [14], this study will investigate the tester's scaffolding strategies in GSDT.

## 2.4 Conversational Agents' Scaffolding Techniques

Empowering CAs with scaffolding techniques has been studied in various scenarios. Scaffolding is one of the principles to guide the design of educational CAs [36]. For instance, Sara, a scaffolding-based CA designed for online programming tasks, utilized scaffolding questions to enhance students' comprehension and promote meaningful learning [77]. Aside from scaffolding questions, topic recommendations from the CA were also considered a form of scaffolding to expand the user's knowledge and interests [81]. In the classroom, CAs were deployed to scaffold and promote students' open-ended discussions [68, 69]. Compared to paper-based static scaffolds, the CA's dynamic scaffolds proved to be more valuable for long-term skill development [76]. Thus, CAs have applied natural language processing technologies to generate adaptive scaffolding [2, 23, 29]. A recent study introduced Promptiverse, an approach that leverages traversal paths over knowledge graphs to generate diverse and scalable multi-turn scaffolding prompts [38]. In the scenario of decision-making, the CA was designed with dialogues to encourage brief exchanges of conversation, which would assist the user in making precise estimations and assessments [57]. However, as an important support for people with NCDs, scaffolding has yet to be explored for designing CAs for NCD screening. Inspired by previous works including the idea of adaptive scaffolding [23, 29, 38] and scaffolding forms such as questions and hints [77], this work will design a framework of adaptive scaffolding in different forms for the NCD screening CA.

# 3 DESIGN: SCAFFOLDING STRATEGIES OF A CA FOR NCD SCREENING

In this section, we present our design process, which includes content analysis and iterative prototyping, as illustrated in Fig. 2. To extract design suggestions for the CA's scaffolding, we analyzed the professional testers' scaffolding strategies in the human-human conversation data from video recordings of NCD screening. By combining deductive and inductive coding, the content analysis produced a set of design guidelines and a taxonomy of scaffolding strategies. Based on these, we proposed a scaffolding framework for our CA design. Then, following prior works [24, 30], we performed an iterative process of *design*, *test*, and *analyze* to create a research prototype accordingly.

## 3.1 Content Analysis of the Human-Administered NCD Screening Test Recordings

*3.1.1 The Grocery Shopping Dialog Task (GSDT).* Many NCD screening tests involve dialog between the subject and the tester. Tests such as MoCA [50] and ACE-R [47] involve the tester providing instructions and hints to guide the subject in completing various tasks, such as recalling words or naming objects in pictures. Tasks such as the Cookie Theft picture description task in BDAE [22] involve more complex dialog due to the tester's open-ended, e.g., "What's happening in that picture?" When the engagement is low, the tester may prompt the subject [12], e.g., "What else is going on?" To provide essential guidance, it is crucial to have a clinical professional administer these tests or tasks in person [67, 82]. The Grocery Shopping Dialog Task (GSDT) [21] selected for this study is representative of tests that encompass multiple cognitive functions, involve dialog, and are administered in person by a clinical professional. Furthermore, it is tailored to the local context using materials familiar to the residents. Specifically, subjects are tasked with narrating the process of purchasing ingredients for a commonly known local dish on a grocery store map. Subjects are presented with an image of a dish and are assigned the task of identifying its ingredients. Subsequently, the subject verbally articulates the step-by-step process of purchasing the identified ingredients, simulating an experience in a physical grocery store using a map or floor plan. The tester manipulates a token representing the subject's position on the map, while the subject provides instructions on its movement. The map may feature obstacles like walls or shelves, prompting the subject to devise feasible paths for accessing all the necessary ingredients. The dish and map are rendered on cardboard, with a small humanoid doll serving as the token, akin to a board game piece. The subject's performance is evaluated based on the accurate identification of ingredients in the dish and their corresponding positions on the map (refer to [21] for scoring details). Healthy or typically aged subjects frequently achieve scores exceeding 9.6, similar to the scores (approximately 22) observed in MoCA (see supplementary materials for details). After consulting with clinicians, it was recommended that older adults regularly undergo the GSDT with an interval of 1-6 months, either as part of a comprehensive cognitive test battery or as a standalone assessment.

*3.1.2 The Data.* Our content analysis covered video recordings of 12 subjects (age=63-86) performing the GSDT with human testers in real clinical settings. The 12 subjects were evenly distributed in terms of gender and cognitive level, with six female subjects, and four subjects in each of the following categories: major NCDs (i.e., dementia), mild NCDs (i.e., mild cognitive impairment), and healthy control, all based on clinicians' formal diagnoses. Each cognitive level had an equal number of female and male subjects. The official task description [21] provided a general overview, such as instructions on purchasing the required ingredients, but lacked specific guidance. Consequently, the tester may employ scaffolding techniques to facilitate task comprehension for the subjects. For instance, to initiate the process, the tester might inquire, "Which ingredient would you like to purchase first?"

*3.1.3 Coding.* To code the video recordings, we used ELAN [37]. We developed the initial deductive coding framework based on the ZPD theory [71] and the three-level scaffolding theory [26]. The former postulates when and how much scaffolding should be applied. The latter introduces three analytical levels of scaffolding during joint activities between NCD patients and their caregivers, which is similar to our NCD screening scenario. We followed this initial coding framework to label the type and the context of each scaffolding occurrence identified in the recordings. We observed that most of the testers provided scaffolding during the ingredient purchasing phase of the GSDT. Consequently, we focused on this specific phase for the subsequent inductive coding of detailed scaffolding strategies.

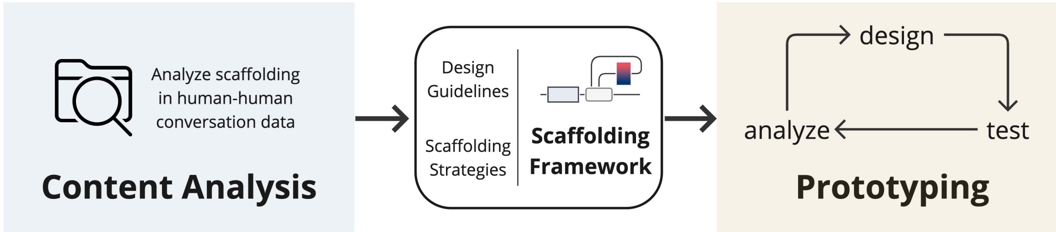


**Figure 2: The process of designing the CA's scaffolding.**

The actual content analysis was conducted as follows. First, five coders transcribed and familiarized themselves with the data. Following the analysis method proposed by W. James Potter and Deborah Levine [55], we assigned each of the 12 recordings to two coders to interpret the subject's responses according to the ZPD (deductive) and label the tester's consequent scaffolding level (deductive). Meanwhile, specific strategies were identified inductively. The coders collaboratively implemented a dynamic codebook to log and align their intersubjective norms. In this process, the five coders met frequently, discussing and consolidating the initial codes by addressing overlapping labels, streamlining the process, and eliminating redundancy.

After completing the coding process, the coding team generated a codebook (see supplementary material) that provides descriptions of the subject's responses, the tester's scaffolding strategies, and their relationship. We elaborate on the coding results in Section 3.3. To ensure codebook consistency and coverage, the coding team labeled video recordings of another five subjects (age=70-78, including two females, and three subjects with mild NCDs) accordingly. For assessing inter-rater reliability between the two coders for each video, Cohen's kappa values were calculated, resulting in .789 for annotations of the subjects' response categories, .866 for annotations of the testers' scaffolding levels, and .787 for scaffolding strategies.

No new codes emerged, and inconsistency was resolved by a third coder. To evaluate whether one could use our codebook to reliably predict the professional human tester's scaffolding strategy exhibited in the videos based on the subjects' utterances only, another co-author (not involved in coding) watched the subjects' parts of the five recordings, chose code labels, and then predicted a scaffolding strategy from our codebook. Using labels from the coding team as the ground truth, the precision for interpreting the subject's responses was 86.0% and for predicting the tester's scaffolding strategies was 74.6%. The lower precision for predicting scaffolding strategies can be attributed to the fact that, besides the ground truth, multiple other strategies were also applicable in that context, as confirmed by the coding team.

From the content analysis results, we derived two guidelines for designing an NCD screening CA capable of scaffolding: adjust scaffolding frequency to align with the user's ZPD status, and follow the three-level scaffolding framework to handle various user responses.

## 3.2 Guideline 1: Adjust Scaffolding Frequency According to the Zone of Proximal Development

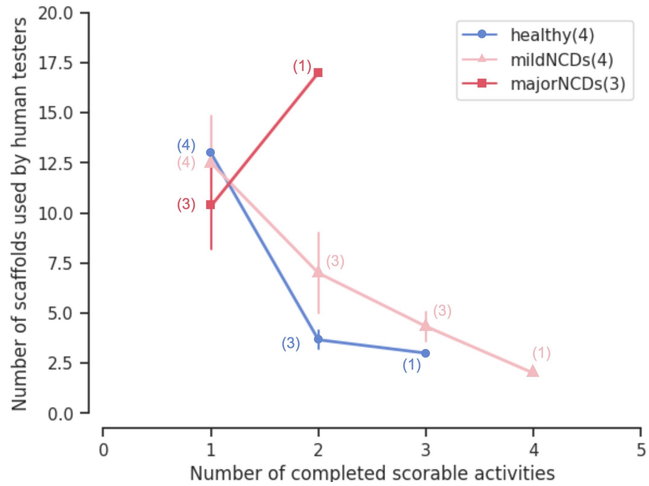


**Figure 3: The number of scaffolds (all types included) used by the human testers for subjects in groups of healthy and mild NCDs declined as the scorable activities were completed gradually. The sample size for each data point is annotated. Most subjects with major NCDs completed only one scorable activity.**

The Zone of Proximal Development (ZPD) theory has been employed to guide caregivers in providing appropriate support to individuals with NCDs, ensuring a balance between adequate assistance and avoiding excessive intervention [75]. Similarly, the frequency of scaffolding provided by a conversational agent (CA) should be tailored according to the user's ZPD status. Based on our coding results, we observed a significant negative Spearman's rank correlation of -.56 ($p<.01^{**}$) between the number of scaffolds used by human testers and the number of completed scorable activities. As illustrated in Figure 3, participants without or with mild NCDs demonstrated improved proficiency in independently completing the task, indicating a change in their ZPD status, with a noticeable decrease in the number of scaffolds provided by testers as the number of completed scorable activities increased. In contrast, subjects with major NCDs needed more scaffolding as the task progressed and many of them purchased few ingredients despite receiving scaffolding. Upon reviewing the recordings, we observed that individuals with major NCDs experienced disorientation and had a tendency to forget about the task at hand. However, these observations may be attributed to individual differences.

Based on our findings, we drew an essential guideline for designing our scaffolding CA: adjust scaffolding frequency according to the ZPD. Specifically, if the user is seen to be progressing towards the zone of actual development in the ZPD spectrum (refer to Fig. 1), indicating an increase in their understanding of the task, the scaffolding frequency can be gradually reduced. However, if repeated attempts at scaffolding prove ineffective, the CA should implement an exit mechanism from the screening test as the user might be in the out-of-reach zone.

## 3.3 Guideline 2: Follow the Three-Level Scaffolding Framework

*3.3.1 Identify Human Testers' Scaffolding Strategies.* Three levels of scaffolding strategies employed by testers were identified based on the theory of scaffolding for individuals with NCDs [26], as outlined in Table 2. Scaffolding is employed at the activity level to enhance subjects' understanding of the overall task process and activity organization during the test. The GSDT consists of four commonly occurring activities: (a) naming the ingredient to buy, (b) locating it on the map, (c) narrating the token's path to it, and (d) purchasing the ingredient. Scaffolding at the activity level typically takes the form of a question that serves as a transition between different activities. For example, to initiate activity (a), the question "What would you like to buy to cook the soup?" is posed, and to transition to activity (b), the question "Where can you buy this in the grocery?" is asked. When the subject successfully purchases the correct ingredient at the correct position in activity (d), they earn a score. This process is repeated until the subject believes they have purchased all the required ingredients or until the tester ends the test due to the subject's out-of-reach zone. The more ingredients the subjects successfully purchase, the higher the score they get on the test. A higher score on the test is achieved by successfully purchasing more ingredients.

Scaffolding occurs at both the action level and the repair level within an activity. The action scaffolding involves the use of continuation and follow-up questions, which are tailored to the subject's responses to the questions at the activity level. Subjects who possess a solid grasp of the task tend to provide correct answers or appropriate commands more frequently. In these instances, testers do not need to provide excessive hints but instead, prompt for continuation or request minimal corrections through follow-up questions, enabling subjects to exercise greater autonomy. However, in the early period of the task, subjects may need more scaffolding at the repair level. Testers may encourage subjects when they signal difficulty or express frustration. Alternatively, they may provide necessary hints within the legitimate scope, such as reminders about the rules or objectives of the task, instead of directly providing the correct answers.

In summary, activity scaffolding facilitates the understanding of the task process by helping transition between activities. Action and repair scaffolding support the action completion within an activity and the interpretation of the task rules or objectives when necessary.

*3.3.2 Categorize Responses of Subjects.* In the GSDT, testers initially operate under the assumption that subjects are within their zone of proximal development (ZPD), as proposed by Vygotsky's theory [71]. This implies that providing scaffolds should effectively enhance subjects' comprehension of the task. Testers progressively update their beliefs regarding the actual development of subjects as they observe the subjects' increasing independence in task completion and evaluate their responses and performances. Alternatively, when subjects become disoriented despite the scaffolding provided by testers, they may be considered to be approaching the out-of-reach zone (ORZ).

To unfold this process, we analyzed the subjects' responses following each scaffolding question and identified four categories, as presented in Table 3. Despite receiving detailed task instructions at the beginning of the test, subjects must develop a practical approach to provide correct answers or commands to testers. During the initial phase of the test, subjects may produce numerous irrelevant responses due to misunderstandings, and they may also seek clarification from the testers. We classify these responses as invalid. Valid responses are classified into three categories: correct answers, wrong answers, and vague answers. For instance, when the token is obstructed by a shelf, instructing the token to turn left or right would be deemed a correct answer. If the command lacks specific information about the direction, it would be classified as a vague answer. In such cases, the tester may pose a follow-up question to inquire about the desired direction. If the subject persists in moving forward, it would be deemed a wrong answer. The tester would then either encourage the subject to reconsider or provide a hint regarding the token being blocked by the shelf. In addition, the subject may also raise a question such as "Am I blocked?" or "Which direction am I facing?" These would be categorized as invalid responses and the tester would provide relevant hints to the subject.

Different response categories are regarded as signs of different ZPD statuses. For instance, continuous wrong answers or invalid responses are signs of being near the ORZ and more correct answers indicate more likelihood of being near the ZAD. The identified response categories imply that the CA should adjust scaffolding strategies according to the response category and the user's ZPD status. For instance, repair scaffolding can be employed for either a wrong answer or an invalid response while continuation is used more often for a correct answer.

*3.3.3 The Scaffolding Framework for the CA during NCD Screening.* Drawing upon the findings of the previous content analysis, we developed a framework for the CA's scaffolding strategies. The NCD screening test is organized as multiple activities, each of which can be viewed as a session of scaffolding, as illustrated in Fig. 4. The scaffolding session typically commences with a scaffolding question at the activity level. Based on the user's response, scaffolding at the action level can be employed for situations that are near the ZAD or the middle ZPD. However, when the user is nearly out of reach of the tester's assistance, i.e., near the ORZ, repair scaffolding is considered. If the user is believed to be in the ORZ, the CA can retry the activity by giving instructions again or terminate the activity

**Table 2: The human testers' scaffolding strategies at three levels.**

| Scaffolding Level | Scaffolding Strategy | Strategy Description | Examples in the GSDT |
|---|---|---|---|
| Activity | Scaffolding Question | Start or restart an activity, or transit between activities with a question. | *What's the first ingredient to buy?*<br>*Which shelf can you find the ingredient on?*<br>*How do you walk there?*<br>*What do you want to buy here?* |
| Action | Continuation | Confirm the subject's response or prompt for the next step. | *Anything else?*<br>*What's next?*<br>*And then?* |
| | Follow-up Questions | Request the specification information or information completion. | *Which vegetable specifically?*<br>*Which side to turn?* |
| Repair | Encourage | Suggest one more attempt when the subject signals difficulties. | *Take one more look.* |
| | Hint | Remind about contextual information, conversation history, task rules, or objectives. | *The token is blocked by the shelf.*<br>*We already bought this one.*<br>*You can tell me whether to go straight, turn left or right.*<br>*We are preparing to cook the soup.* |

**Table 3: The responses of subjects during the NCD screening task are categorized. Their responses reflect their ZPD status either near the zone of actual development (ZAD), the middle zone of proximal development (ZPD), or near the out-of-reach zone (ORZ).**

| ZPD Status | Subject's Response | Examples in the GSDT | Available Scaffolding Strategies |
|---|---|---|---|
| Near the ZAD | Correct Answer | An answer that articulates what to buy next, the shelf location, how to move/stop the token, or the command to purchase. | Continuation or scaffolding questions to proceed to the next activity. |
| The Middle ZPD | Vague Answer | An answer missing necessary information, e.g., a command to turn the token but without a specific direction. | Follow-up questions, hints, or repeat the previous scaffolding questions. |
| | Wrong Answer | An answer with invalid location information of the target shelf, a non-executable command to move the token, or buy ingredients at an invalid position. | Hint, encourage or repeat the previous scaffolding questions. |
| Near the ORZ | Invalid Response | An irrelevant response signaling problems in finding ingredients, misunderstanding the token state, the task rules or objectives. | Hint, encourage or repeat the previous scaffolding questions. |

or the test. If it is believed to be in the ZAD, no further scaffolding is needed and the user completes the activity independently. Upon completion of the activity, e.g., having moved the token to the correct position, the user can proceed to the next activity. Generally, the user determines whether all necessary ingredients have been purchased and decides to end the task.

## 3.4 Prototyping

The ultimate objective of this work is to automate the NCD screening task using the CA. However, a fully automatic CA necessitates a sophisticated system that integrates technologies such as automatic speech recognition, natural language processing, and speech synthesis. In this preliminary study, we concentrate on designing the CA's scaffolding strategies. Consequently, we developed a semi-automatic prototype system with human wizards simulating the conversation behind the CA during the NCD screening. This could also be considered as a Wizard-of-Oz study which is commonly used in the research of CAs for patients with NCDs [56].

Initially, we designed the CA's scaffolding based entirely on the theories and patterns observed in human-human conversations. However, there may be disparities between human-human and human-CA conversations. Therefore, we utilized an iterative design method, inspired by the design thinking approach for the "first-use study" of information appliances [24], to evaluate and enhance

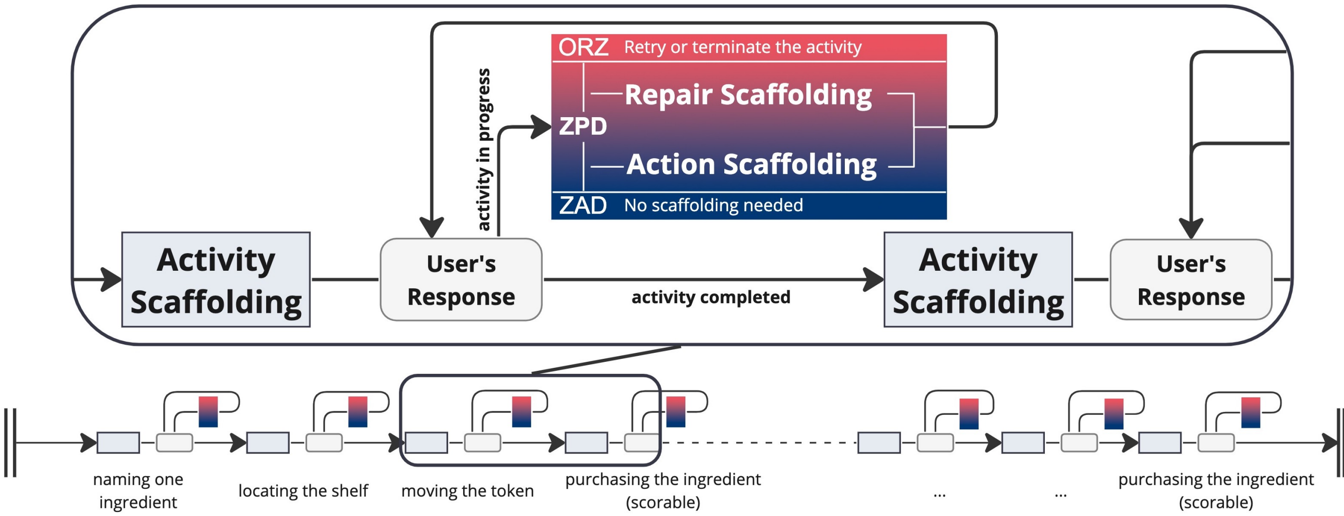


**Figure 4: The CA's scaffolding framework for the dialog task of NCD screening.**

our prototype CA system. This process can also be viewed as the wizard's training in controlling the CA system before the formal user study.

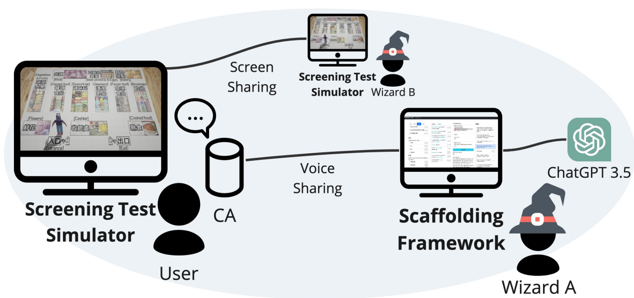


**Figure 5: The architecture of the semi-automatic CA system.**

3.4.1 *System Design.* In our system, two wizards collaborate as Fig. 5 illustrates. Wizard A manages the CA's conversations using an interface that incorporates our scaffolding framework, as illustrated in Figure 6. The interface for wizard A comprises four modules. The first column on the left contains the activity module, which includes four types of scaffolding questions representing the four activities. Below that is the dialog control section, where the wizard can utilize backchannels and delay audio clips to manage response latency and maintain user engagement. The second module is dedicated to categorizing the user's responses. The third module involves consulting ChatGPT using the prompt template, which requires information about the current activity, the user's response, and the token state (see Figure 11). We evaluated the viability of models equipped with vision-language understanding capabilities, such as MiniGPT-4 [84] and BLIP-2 [39], and observed their difficulty in comprehending the spatial information of the grocery map image, leading to multiple instances of hallucinations. Consequently, we opted to embed visual information in the prompt with natural language and employ the currently state-of-the-art model, ChatGPT-3.5. The final module on the right is the scaffolding strategy module, which contains all the available strategies for addressing the user's current response (see supplementary material for all alternatives of scaffolding strategies). For each strategy, we leverage the original utterances spoken by professional human testers in the video recordings and generate paraphrases to increase adaptability. Subsequently, the wizard clicks the "Play" button, allowing the CA to respond to the user with a prerecorded audio clip.

Wizard B collaborates with wizard A by controlling the token's movement in the simulator (see Fig. 10) according to the user's response. For instance, if wizard A deems the user's response to be a correct command to move the token, wizard B moves the token accordingly. The two wizards maintain communication throughout the test. The simulated NCD screening test is presented to the user via screen sharing from wizard B. A microphone captures the user's verbal responses, and a speaker plays the CA's responses based on wizard A's actions.

3.4.2 *Implementation and Test.* Wizard A's interface was developed as a web-based application, using Django[3], the Python web framework, and Bootstrap[4], the website development toolkit. The ChatGPT model we employed was gpt-3.5-turbo. Wizard B utilized the GSDT simulator, which was constructed using Three.js[5], the JavaScript 3D library. Both wizards worked remotely through the online meeting applications, with wizard A sharing audio and wizard B sharing the screen. A 15' laptop was installed at the hospital where participants were recruited. Each participant joined the online meeting with both wizards during the study. This allowed the GSDT simulator to be visible on the laptop screen, and the voice of the conversational agent (CA) could be heard from the laptop speaker.

[3]https://www.djangoproject.com/
[4]https://getbootstrap.com/
[5]https://threejs.org/


**Figure 6: The wizard A's interface for controlling the CA. Please note that all the prompt text for ChatGPT was in English because only the category description was kept in the prompt instead of the original user utterance. Also, ChatGPT returned only the recommended scaffolding strategy in English, instead of a specific response. Cantonese was only used for the CA's response to the user.**

Based on the implementation, we conducted nine rounds of tests. To obtain feedback from multiple perspectives, we invited experts who were familiar with NCD patients to role-play the user in the tests, while two researchers acted as wizards A and B. The CA-user conversations, the operations of wizards A and B, and the recommendations of ChatGPT were all logged during the test.

Aside from technical problems, our test focused on three aspects: the coverage of the scaffolding framework, the prompt for ChatGPT, and the system latency. The researchers could pause the test anytime if issues arose, particularly in cases where wizard A encountered difficulty in finding a suitable scaffolding option or when the recommendations from ChatGPT were considered unreasonable.

*3.4.3 Scaffolding Framework Coverage.* The CA's scaffolding framework was designed to cover as many situations as possible during the NCD screening test. In the content analysis period, we already checked the coverage of the scaffolding framework by utilizing it to scaffold the subject's responses in an extra set of video recordings. However, in this user study, we tested the scaffolding framework in conversations between humans and the CA. While no new strategies were introduced, we incorporated new paraphrases within specific strategy components to enhance the framework's ability to handle diverse contexts. This adjustment was necessary since the system was limited to playing pre-recorded audio instead of generating spontaneous speech.

The experts who role-played the subject were required to simulate the cognitive impairment of an NCD patient to an extreme extent, such as misunderstanding the rules, forgetting about what had already been done, experiencing difficulties verbalizing commands, and so on. In order to handle situations where the user's engagement becomes unattainable and to adhere to the first design guideline, we sought input from the experts and incorporated an exit strategy for the CA. If the repair scaffolding proves unsuccessful after several attempts, the CA initiates a restart, beginning with activity (a), naming the ingredient to be purchased. If even this approach proves ineffective, the CA terminates the test.

*3.4.4 Prompt Engineering.* Whenever wizard A clicked the "Ask GPT" button in the interface depicted in Figure 6, the prompt was incorporated into an HTTP request to the ChatGPT API[6]. The prompt template (see Fig. 11) consisted of three parts: introduction, message history, and prompt. The introduction conveyed the role of ChatGPT as a healthcare worker and provided essential information

[6]https://platform.openai.com/docs/api-reference/chat

about the GSDT. Meanwhile, the message history logged the exchanges between the wizard and ChatGPT. To avoid exceeding the length limit of ChatGPT API, the latest three pairs of prompts and responses were retained. The prompt contained information about the current activity, the token state in the grocery, a description of the user's response category, and all available scaffolding strategies retrieved from our framework. ChatGPT was tasked with selecting a suitable scaffolding strategy to address the user's response, taking into account the provided contextual information, and providing an explanation for its choice.

Inspired by the structured process of prompt engineering in prior work [83], we marked inappropriate strategies recommended by ChatGPT, modified the prompt template, and evaluated during the iterative prototyping. A significant concern arose when ChatGPT provided inappropriate strategy recommendations due to inaccuracies in the token state description. To address this, we revised the token state list. According to the prompt engineering suggestion in prior work [13], the modification was mainly on the natural language description. Another issue was the invalid recommendation, meaning that ChatGPT sometimes recommended a strategy that was not based on our scaffolding framework. For instance, in situations where a token was obstructed by a shelf, our scaffolding framework offered only two viable strategies. However, ChatGPT occasionally recommended an additional strategy, disregarding the constraints imposed by the framework. To address this issue, we implemented two approaches that have been proven effective in previous research [40]: providing examples and explicitly reiterating the constraints within the prompt.

*3.4.5 Latency.* Wizard A had a workload to respond to the user, including categorizing the user's response, updating parameters for prompts and requesting ChatGPT, manually selecting a strategy if ChatGPT's recommendation was inappropriate, and finally playing the response audio. Apart from the wizard's response time, the latency was also influenced by the process of requesting ChatGPT. Under typical circumstances, where the user's response was properly categorized and ChatGPT made appropriate recommendations, the latency ranged from 3 to 8 seconds. However, when the user's actions were unexpected or the wizard had to manually choose an alternative strategy, the latency increased to over 15 seconds.

We tried to minimize the effects of latency by reducing the wizard's information loads and operations. For instance, the keywords were highlighted in the interface and the ChatGPT recommended strategy unfolded automatically. Additionally, wizards were trained to respond faster. However, a more crucial aspect was to mitigate the impact of latency on the user experience. Taking inspiration from previous research [14], we incorporated responses such as backchannels in the lower section of the first column, as depicted in Figure 6, to enhance dialog control. When the wizard heard the user, the backchannel response could be used to acknowledge receipt, and the utterances in the delay section could be utilized to keep the user momentarily engaged.

## 4 EXPLORE: THE EFFECTS OF THE CA'S SCAFFOLDING IN USER STUDY

In this exploratory study, we conducted a first-use study similar to the previous HCI study on prototype design [24]. We did not include any baseline system such as a CA without scaffolding or the human-administered NCD screening. This is because, without verbal scaffolding, people would find it challenging to display their cognitive ability in the dialog task. Additionally, the prototype CA only focused on part of the GSDT, so validating the results of the CA's NCD screening based on human-administered NCD screening would be necessary only when the system is fully automated and covers the GSDT entirely.

### 4.1 Participants

In our study, we recruited a total of 15 participants (M=66.3, SD=6.69) from a collaborative local hospital, with an age range of 55 to 76. Of the participants, 9 were females. Participants were evenly distributed into three groups: major NCDs, mild NCDs, and healthy. None of them had completed the GSDT before the study. This study obtained approval from the University Research Ethics Committee's Institutional Review Board (IRB). We collaborated with a local hospital and obtained approval from the hospital to conduct experiments on their facility. Additionally, the whole study was closely monitored by clinicians from the hospital on-site. Prior to the study, all participants were instructed about the purpose, content, and process of the study and they consented to take part in the study with audio and video recorded. All personal information of the participants was appropriately anonymized to ensure confidentiality.

**Figure 7: The user study setting. A clinician monitored the user study (partly on the right) to ensure the participant's (on the left) safety but any forms of assistance for the test were not provided. The wizards operated remotely so the participant could not see them.**

### 4.2 Measurements

Throughout the study, we recorded all the conversations between the CA and the participants, as well as all the operations of the wizards. Based on the logged data, we could quantitatively and qualitatively analyze how the CA's scaffolding supported participants during NCD screening and evaluate ChatGPT's performance. Following the study, we gathered feedback from both clinical experts and the participants.

We invited three clinical experts from a research department at a local hospital to evaluate the CA's scaffolding (see Table 4). The invited experts and those who administered the test in the initial

**Table 4: The background information of clinical experts we invited to evaluate the CA, including their clinical experience of cognitive assessment.**

| | Background | Clinical Experience |
|---|---|---|
| Expert 1 | Neuroscience + Psychology | 5 years |
| Expert 2 | Psychology | 2 years |
| Expert 3 | Neuroscience + Psychology | 1 year |

video data for content analysis were from the same research team. Though with different years of experience in cognitive assessment, they were trained by the same procedure and had a similar amount of practice in administering the GSDT. Each expert was randomly assigned the data of 4-6 participants (1-2 hours of evaluation) and finished the evaluation online with a coauthor's assistance. They were paid 38.5 USD for each hour of evaluation. The coauthor played the video recordings and instructed the expert to label each response categorized by the wizards, as well as each scaffolding strategy employed by the CA and recommended by ChatGPT, as "appropriate" or "inappropriate". This assessment aimed to determine whether the categorization of the participant's response or the chosen strategy was suitable within the given context. Additionally, the experts were asked to provide reasons for any identified instances of inappropriateness.

To evaluate from multiple perspectives, we collected feedback from the participants with a post-study survey about usability and willingness (see Fig. 8). We designed the 11-item questionnaire in the survey based on usability [68, 86] and willingness [64, 81] evaluation for previous CAs. Participants were requested reasons for each answer.

**Table 5: The average counts of each participant's responses categorized by the CA in the user study and their percentage.**

| ZPD Status | Category | Mean Count(SD) | Pct. |
|---|---|---|---|
| Near the ZAD | Correct Answer | 14.93 (5.47) | 66.45% |
| The middle ZPD | | 4.33 (4.29) | 19.27% |
| | Vague Answer | 0.8 (1.15) | 3.56% |
| | Wrong Answer | 3.53 (4.22) | 15.71% |
| Near the ORZ | Invalid Response | 3.2 (3.47) | 14.24% |
| Total | | 22.46 (11.73) | |

**Table 6: The CA's average scaffolding counts and percentage at different levels in the user study.**

| Scaffolding Level | Count (SD) | Pct. | Scaffolding Strategy | Count (SD) | Pct. |
|---|---|---|---|---|---|
| Activity | 8.07 (3.26) | 31.86% | Scaffolding Question | 8.07 (3.26) | 31.86% |
| Action | 11 (5.73) | 43.42% | Continuation | 9.6 (4.7) | 37.89% |
| | | | Follow-up Question | 1.4 (1.4) | 5.53% |
| Repair | 6.27 (6.41) | 24.75% | Encourage | 0.87 (1.6) | 3.43% |
| | | | Hint | 5.4 (5.33) | 21.32% |
| Total | 25.33 (13.03) | | | | |

# 5 RESULTS

## 5.1 Summary of the CA's Use of Scaffolding

Throughout the user study, every participant completed at least one ingredient purchase successfully. Notably, participant P2, who was in the healthy group, purchased the highest number of ingredients, specifically five. On average, each participant acquired 2.87 (SD=1.13) ingredients. We summarized the CA's categorization results of participant responses in Table 5 and the CA's usage of scaffolding in Table 6. The data across participant groups (NCD levels) is in the supplementary material. During the user study, the CA averagely categorized 22.46 (SD=11.73) responses from each participant and provided scaffolding 25.33 (SD=13.03) times for each participant. Repeated or paraphrased scaffolding strategies were used to keep participants engaged and help them understand the scaffolding better, so the number of scaffolding was slightly greater than the number of categorized responses.

## 5.2 The CA Provided Appropriate Scaffolding (RQ1)

Overall, 89.45% of the scaffolding strategy adoption was regarded as appropriate by the clinical experts. Additionally, the clinical experts evaluated 98.54% of the response categorization as appropriate. The few instances of inappropriate response categorization were due to operational errors by the wizards, such as clicking the wrong button. In response to RQ1, which focuses on how to design a CA that can implement scaffolding strategies used by clinical professionals, we presented a design process for the CA that incorporates the scaffolding strategies derived from the earlier content analysis. The findings demonstrated that the CA, equipped with the scaffolding framework, successfully generated appropriate scaffolding strategies. This outcome suggests that the CA, developed through the proposed process, effectively exercises the scaffolding strategies employed by clinical professionals, thereby addressing RQ1.

Furthermore, the experts identified various instances of inappropriate scaffolding. One type involved situations where the information provided by the CA's scaffolding was occasionally inadequate. Although the adopted strategy sufficed to continue the task, it would have been preferable if the necessary information had been explicitly provided. For instance, Fragment 1 illustrated that P10 encountered difficulty locating pictures of tomatoes on the grocery map. In such cases, it would have been beneficial for the CA to provide an explicit hint that no such pictures might exist, as the pictures of shelves (refer to Fig. 10) were solely for reference purposes.

Clinical experts advised that the CA should use more straightforward strategies for participants with major NCDs. As shown

**Fragment 1: The information provided by the CA's scaffolding was insufficient here. Instead of encouragement, the CA should remind P10 that the picture was only for reference and that the token could go anywhere P10 thought that there were tomatoes.**

| | |
|---|---|
| | (P10 cannot find tomatoes in the grocery.) |
| P10: | Tomatoes, eggs. I can't find it here. |
| CA: | Where can you find it? (A scaffolding question) |
| ChatGPT: | The user's response indicates that they do not understand the task rules. Instead of assuming that the user knows the task rules, it is important to explain or reclaim the task rules to ensure that the user understands how to give correct commands. This will help provide clarity and guide the user in the right direction. |
| P10: | Tomatoes. I can't find where I can buy tomatoes. |
| CA: | No worries. Try one more time to find it. (Encouragement) |
| P10: | Can't find it here. |

**Fragment 2: The strategy should be more straightforward for P15 who had major NCDs. The CA would better ask which direction to turn when P15 wanted to turn the token.**

| | |
|---|---|
| | (The token is at the entrance.) |
| CA: | Please tell me how to walk there. |
| P15: | Go to... Go to the entrance. |
| CA: | The token is here. (A hint of contextual information) |
| P15: | Turn that way. |
| P15: | I don't know how to say. |
| CA: | The token is currently facing forward. (A hint of contextual information) |
| P15: | Go forward. |

in Fragment 2, the strategy of providing contextual information about the token was not clear enough to guide the participants with major NCDs. Instead, the experts suggested choosing the strategy to directly clarify the task rules or request the missing information, such as requesting the specific direction to turn. Generally, the CA could try one scaffolding strategy two times or three by repeating or rephrasing it. If it did not help, more straightforward strategies could be considered.

Last, it is also suggested that the CA should consider the participant's emotional state and adjust the scaffolding strategy as well. For example, when P3 started to narrate the path to move the token instead of answering where to find the pork, this could be a sign of impatience according to the experts' inferring. The CA could request the shelf name again as logged in Fragment 3 but it would be better to move the token as P3 told before that.

## 5.3 The Feasibility of ChatGPT for Scaffolding (RQ1)

During the study, ChatGPT recommended the scaffolding strategy 252 times in total, 16.8 (SD=8.56) times per participant. Although only 54.27% of these recommendations were adopted by the wizards,

**Fragment 3: The CA should notice the impatience of P3. So the CA's hint here was unnecessary.**

| | |
|---|---|
| CA: | What would you like to buy next? (A scaffolding question) |
| P3: | Buy pork. |
| CA: | Where can you find it? (A scaffolding question) |
| P3: | Turn right and go forward. |
| CA: | Tell me the shelf name of the ingredient. (A hint of the rule) |
| P3: | Meat. |

**Fragment 4: ChatGPT could avoid human error.**

| | |
|---|---|
| CA: | What is the first ingredient you would like to buy? (A scaffolding question) |
| P3: | Vegetable. |
| CA: | Which shelf can you find it on? (A scaffolding question) |
| ChatGPT: | The user's response was vague, and I can ask for clarification on the specific ingredient the user wants to buy. This will help me get a clearer understanding of the task confirmation. |

68.25% were evaluated as appropriate by the clinical experts. This result suggested our design process to integrate ChatGPT in the CA for automating scaffolding was promising, which answered RQ1. However, we identified notable issues of ChatGPT performance from the data log during the study.

Initially, we observed a lack of flexibility in ChatGPT's recommendations within specific scenarios. Despite our framework offering multiple available strategies for ChatGPT to select from, it consistently recommended the same strategy. For instance, when a participant issued a command to purchase a single ingredient while positioned correctly in front of the corresponding shelf, ChatGPT consistently advised using a strategy to confirm the token's position. Although this recommendation was generally appropriate for such situations, the wizards employed more versatile strategies, such as proceeding to the next task (purchasing the next ingredient), inquiring about additional ingredients needed, or providing a general prompt for the subsequent step.

Then the second issue was the insensitivity to contextual information. Selecting an appropriate scaffolding strategy relied on contextual information, including the state of the token. Nevertheless, the state of the token appeared to have a limited influence on ChatGPT's generation. For instance, when the participant consistently provided accurate responses to move the token, the token's state transitioned between "stands at the entrance," "is moving forward," or "stops and needs to turn." Ideally, ChatGPT should have recommended strategies according to the token state. However, it kept picking the same continuation strategy to prompt for the next step regardless of how the token state changed. The experts noted that while this strategy was acceptable for the majority of participants, it would be more suitable to offer precise information, particularly for individuals with significant NCDs. For example, an additional strategy could be employed to confirm when the token should stop while in motion. Another aspect of contextual information, namely the message history, had minimal impact on ChatGPT's recommendations, as no recommendation reasons provided by ChatGPT appeared to be grounded in the message history.

Moreover, we noticed ChatGPT's hallucinatory reasoning for its recommendation. For instance, ChatGPT recommended one strategy but in its reasoning description, it thought another strategy was more appropriate. The original ChatGPT generation: *"GPT Recommends: P4 (continuation or prompt for the next command). In this situation, the user has provided an executable command to move the token. Instead of continuation or prompting for the next command, it would be* ***more appropriate*** *to provide contextual information e.g., token info or previous commands. This can help the user understand the current state of the token and provide them with the necessary information to make further decisions."* Similarly, ChatGPT recommended one strategy though regarded as unnecessary: *"Strategy SQc-1 (asking for a specific direction to turn or go) is the best option in this situation because the token is already facing the correct shelf in activity (c). Asking for a specific direction to turn or go would* ***not*** *be necessary as the user has already given the executable command to stop the token appropriately."* ChatGPT seemed to take the token state into consideration but made an incorrect reasoning.

Despite its flaws, ChatGPT possessed notable strengths. First, ChatGPT was more straightforward than the wizard in some cases. In Fragment 1, had the wizard followed ChatGPT's recommendation, the CA would have been able to provide P10 with the suggestion that the position of tomatoes could be deduced from the shelf names rather than relying on the shelf pictures (refer to Fig. 10). Occasionally, human wizards made mistakes by overlooking critical information; however, ChatGPT had the ability to circumvent such human errors. For example in Fragment 4, the wizard overlooked P3's vague answer but ChatGPT insisted on asking for clarification of "vegetable", which was evaluated as more appropriate by the experts.

## 5.4 Participants Perceived the CA's Scaffolding as Helpful in Task Understanding (RQ2)

Based on the post-study survey results as shown in Fig. 8, we found that participants generally considered the CA's scaffolding as easy to understand and useful to facilitate their task understanding. The CA's scaffolding was perceived as not "too much" during the task. The system latency was acceptable for most participants. Though participants showed a willingness to complete NCD screening tasks with the guidance of the CA, they preferred the companionship of family members. P18 and P19 were concerned about any unpredictable problems so they might need a family member for help. P3 found the dialog task was kind of less interesting so it would be more fun to complete it with family members. However, P10 also mentioned the concern that family members were too busy to help her with the NCD screening test. Most participants did not mind making mistakes in front of the CA because the CA showed much patience as mentioned by P18, except P10 said she was embarrassed when making mistakes. When confronted with difficulty, participants thought that the CA had plenty of patience and sometimes

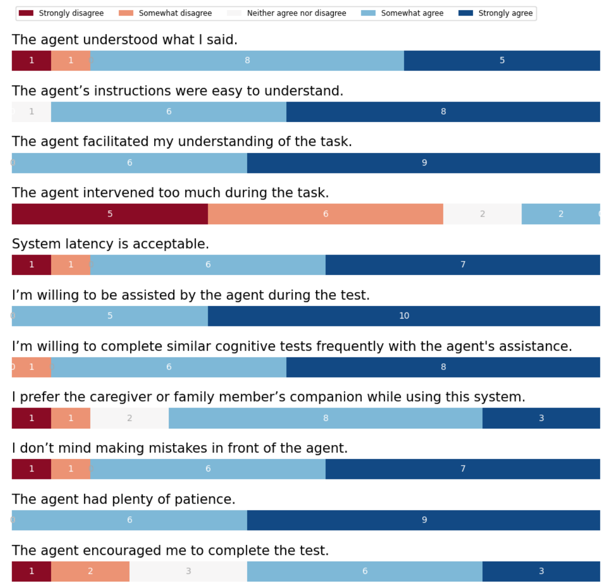


**Figure 8: The post-study survey result. It was well explained to each participant that the "agent"' here refers to the CA they talked with in the study.**

encouraged them to complete the task. In summary, the result answered RQ2 about what the effects of the CA's scaffolding were on participants' task understanding from a subjective perspective.

## 5.5 The CA's Scaffolding Usage Decreased as Tasks Progressed (RQ2)

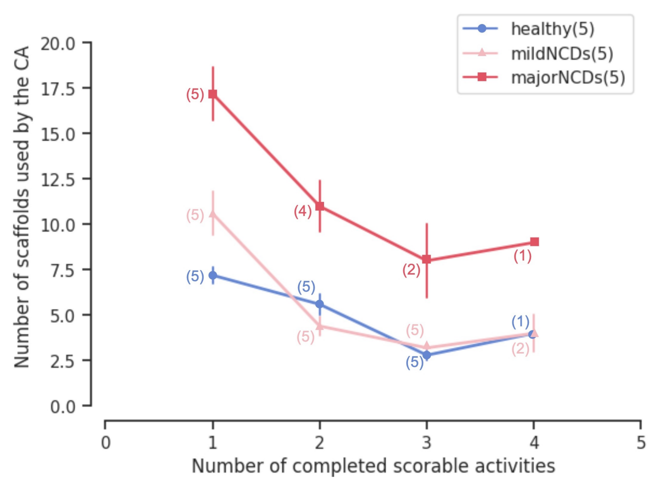


**Figure 9: The number of used scaffolds (all types included) by the CA decreased as the scorable activities were completed one by one. The sample size for each data point is annotated. Participants with major NCDs generally completed fewer scorable activities.**

To address RQ2 objectively, we conducted an analysis of the system log regarding the usage of the CA's scaffolding for the participant's completing each scorable activity throughout the study. We identified a similar decline in scaffolding use as shown in Fig. 9. A Spearman's rank correlation of -.62 (p<.001***) was found between the quantity of scaffolds provided by the CA and the number of completed scorable activities by the participants. This implies that the CA's scaffolding supported participants' comprehension of the task, similar to how the scaffolding provided by human testers enabled participants to grasp the interpretation of contextual information and articulate accurate answers or commands to accomplish the task. Furthermore, this indicated a transition in the participants' zone of proximal development (ZPD) status, shifting from the vicinity of the out-of-reach zone (ORZ) to the proximity of the zone of actual development (ZAD). The number of used scaffolds varied more for participants with major NCDs, and they generally needed more scaffolding during the task. Also, they tended to score less as only two of them purchased the third ingredient. Nonetheless, the decline pattern exhibited a divergence from the observed pattern in human testers' scaffolding for subjects with major NCDs, as illustrated in Figure 3. One potential explanation is that the larger sample sizes of data points in this case could have mitigated the impact of individual differences. Further validation through a comprehensive large-scale study incorporating statistical tests is warranted.

The utilization of scaffolding experienced a slight increase during the final scorable activity as the CA sought confirmation from the participants regarding their decision to conclude the task. For instance, in the later stage of the task, there was an escalation in the frequency of required scaffolding for a particular healthy participant (P2). After checking the log data, we found that it was because the CA was confirming with P2 if the task was finished but P2 forgot which ingredients had been purchased. In summary, the findings imply that the CA's scaffolding facilitated participants' comprehension of the task and promoted their independence, thus addressing RQ2. Moreover, the patterns of scaffolding usage were consistent with the patterns shown in Fig 3, suggesting that the CA's scaffolding had similar effects as the clinical professionals' scaffolding, which also answered RQ1.

## 5.6 Patterns of How CA's Strategies at Different Scaffolding Levels Facilitated Task Understanding (RQ2)

The quantitative analysis revealed a general tendency of growing task understanding of participants affected by the CA's scaffolding strategies. In this section, we further investigate the patterns of how those strategies at different scaffolding levels, i.e., levels of activity, action, and repair, facilitate task understanding of participants based on qualitative analysis.

We prepared scaffolding questions at the activity level for the CA to structure the GSDT. As we mentioned earlier, the task of purchasing each ingredient could be divided into four activities so four types of scaffolding questions were asked sequentially to facilitate the participant's understanding of the task process. We found that in the later period of the task, many participants had understood the process so they could manage the transition between activities independently without the CA's scaffolding questions. This suggested that the scaffolding questions at the activity level improved participants' comprehension of the task process so they could accomplish the whole task, i.e., purchasing all the needed ingredients one by one. For example, after P2 (Healthy) had purchased two ingredients, he could manage the activities independently. However, most participants with major NCDs needed scaffolding at the activity level throughout the whole task.

**Fragment 5: Providing no scaffolding might be regarded as a denial of the previous command by P18 (with major NCDs).**

| | |
|---|---|
| P18: | Tofu. |
| P18: | Can I point the way? [pointing on the screen] Tofu! |
| CA: | … [for about 3 seconds while the wizard is operating] |
| P18: | Oh, I did it wrong... |

**Fragment 6: The follow-up question strategy at the action level facilitated task understanding in a more specific way by probing for missing information.**

| | |
|---|---|
| P15: | Turn. |
| CA: | Which direction to turn? (A follow-up question) |
| P15: | … |
| CA: | Left or right? (A follow-up question) |
| P15: | Left. |

Scaffolding at action and repair levels handled more specific situations. Recall that one typical strategy at the action level was the continuation which helped users confirm their correct understanding or responses so they could proceed to the next step. Interestingly, in the absence of a prompt for continuation, participants might interpret the silence as a rejection of their prior responses, as demonstrated in Fragment 5. Another scaffolding strategy at the action level, namely, the follow-up question, served a distinct purpose by soliciting the missing information from participants. For example, P15 (with major NCDs) already gave a command to turn but lacked the specific direction information (see Fragment 6. Therefore, the CA applied two follow-up questions at the action level of scaffolding to help P15 understand the necessity of direction information.

Scaffolding strategies at the repair level were frequently used during the early period of the GSDT when participants did not know much of the task rule, i.e., near the out-of-reach zone. For example, participants usually did not notice when the token was blocked by a shelf at first so the hint as in Fragment 7 was given as the repair scaffolding.

To summarize, the activity scaffolding aided in the smooth transition between activities and enhanced participants' comprehension of the task process. The action scaffolding facilitated the execution of particular actions within each activity, while the repair scaffolding served as a direct reminder of task rules or objectives as required. These results helped answer RQ2 more specifically with conversational fragments from the user study data.

**Fragment 7: A hint was given in the early period of the test when participants did not know much of the task rule, i.e., near the out-of-reach zone.**

```
      (The token is blocked by a shelf.)
P12:  Go forward...to buy vegetable...
 CA:  It's blocked and cannot go there. (A hint)
P12:  So turn to right.
```

# 6 DISCUSSION

## 6.1 Result Summary

In general, the clinical experts validated 89.45% of the CA-adopted scaffolding strategies and 68.25% of the ChatGPT-recommended scaffolding strategies as appropriate. The results suggested that the CA's scaffolding framework we designed with the proposed design process was effective in producing appropriate scaffolding strategies. Though there was room for improvement, the embedding of ChatGPT was proved feasible in the dialog task scenario for NCD screening. Moreover, the participants evaluated the CA as easy to use and helpful in task understanding. The results provided empirical evidence of the effectiveness of our design process, which answered RQ1.

Based on the conversation log between the CA and participants, we identified the decline of the CA's scaffolding usage as the task progressed gradually, suggesting that the CA's scaffolding strategies facilitated participants' task understanding. We also explored the effects of the CA's strategies at different levels on participants' task understanding. The CA's scaffolding strategies at the activity level helped participants understand the task process and activity transition. Scaffolding strategies of the repair level provided participants with specific information such as the rules or objectives of the task in NCD screening. Strategies of the action level worked in a minimized way to confirm the correct responses of participants or give necessary prompts. The results provided both quantitative and qualitative perspectives to reveal the effects of CA's scaffolding on the participant's task understanding, which answered RQ2. Thus, we propose implications for the future design of CAs in the dialog task scenario for NCD screening.

## 6.2 Design Implications

*6.2.1 Integrating the Three-level Scaffolding Framework for the CA.* Drawing on existing theories, such as the scaffolding-based CA for learning, Sara [77], which was motivated by the theories of social constructivism and the ZPD, to inform scaffolding design is a well-established approach. When developing speech generation algorithms of the CA for NCD screening, researchers integrated both theories and practices of backchanneling [14]. Inspired by the previous works, we combined theories and experts' practices of scaffolding in this work. Moreover, we extracted human strategies to motivate the design of the scaffolding framework for the CA. Building on an examination of scaffolding dynamics between professional human testers and subjects, as well as between the CA and the participants, this subsection proceeds to explore the integration of the three-level scaffolding framework into CA design within this specific scenario.

We identified human scaffolding strategies from the data of human-administered NCD screening tasks as mentioned in section 3.3. To better organize identified scaffolding strategies for the CA design, we integrated the three-level scaffolding theory for NCD patients [26]. The theory conceptualized the scaffolding for NCD patients with three levels, including activity, action, and repair levels. The results of our user study also explored their effects on participants' task understanding. According to the original theory, scaffolding at the activity level should assist NCD patients in framing, reframing, and recalling the ongoing activity [26]. Similarly in this study, the CA's scaffolding strategies at the activity level facilitated participants' understanding of the task process and transition between activities. When participants encountered difficulties, the CA employed scaffolding questions to facilitate the resumption of the activity. Once the participant fully understood the task process, the CA dynamically adapted the scaffolding questions, enabling the reordering of activity sequences, including the option to skip certain activities. Within the GSDT framework, the shopping task involves collaborative activities between the CA and the user, which can be supported through activity scaffolding.

Scaffolding at the action level is possible in constructing and performing joint actions [26]. In this study, various joint actions happened especially in the activity of narrating the process of moving the token. Each step of the token's moving was the result of both the participant's narrative and the CA's scaffolding. To carry out "just enough" scaffolding [75], the strategy of continuation was frequently used by the CA when the participant was comparatively capable, i.e., near the zone of actual development. This was also to maximize the participant's contribution to the task. However, when the participant exhibited limited proficiency in performing the actions, typically during the early stages of the task or in the presence of major NCDs, additional support was provided through scaffolding strategies, such as follow-up questions The scaffolding questions at the activity level were basically open-ended questions such as "How do you get there to buy it?" If the participant failed to verbalize the instruction to move the token step by step, the scaffolding strategy simplified it by raising close-ended follow-up questions such as "Turn left or turn right?" Older adults, especially those with major NCDs, encounter less difficulty in responding to closed-ended questions [46]. Additionally, family members of older adults with NCDs often utilize this strategy [59].

Scaffolding at the repair level is summoned when any identified troubles or problems are produced and this can be indicated by either the speaker or the listener [26]. In this study, analogous patterns were found when the CA hinted at the identified trouble such as when the token was blocked as Fragment 7 showed. Sometimes the participant signaled or pointed out the problem, e.g., P3 noticed the token was blocked and told the CA, and then the CA confirmed. Repairs were organized collaboratively with the CA's scaffolding and the user's improved understanding of the task. More importantly, the CA should organize scaffolding at the three levels collaboratively to deal with various situations in different periods of the task.

*6.2.2 Considerations about the Zone of Proximal Development.* Another scaffolding theory incorporated into our design is the theory of the zone of proximal development (ZPD) [71]. Existing CAs enabled with scaffolding skills used the theory as an assumption, e.g., assuming the CA's scaffolding was within the learner's ZPD so it could help close the knowledge gap, however, the CA's scaffolding strategies were not dynamic according to the learner's ZPD status [2, 77]. In this work, the CA was also assumed to scaffold the users in their ZPD. But since our CA was designed with a variety of scaffolding strategies to deal with different situations, three types of ZPD status were introduced in our CA's scaffolding framework as listed in Table 3. Prior research on problem-solving tasks has revealed the dynamic adjustment of support by teachers or parents to align with learners' ZPD status [54, 70]. Thus, we further divided the ZPD into three subzones, including parts near the ZAD, in the middle ZPD, and near the ORZ. Note that all three subzones were still inside the ZPD so the assumption of working in the ZPD was retained. However, this brought more dynamics to the CA's scaffolding strategies to align with the users' ZPD status.

The ZPD status of users is not identical to their cognitive ability. For instance, individuals with major NCDs could also approach the ZAD during the later stages of the task, while healthy users might find themselves near the ORZ at the task's outset when task information has not been adequately provided. However, there were greater possibilities in general for participants with major NCDs to get near ORZ based on observation in this study. Correct answers, especially multiple successive correct answers, usually indicate the user's ZPD status near the ZAD. So, strategies such as continuations and scaffolding questions to move on to the next activity were often used by the CA in this study, which adhered to the "just enough" principle of scaffolding for NCD patients [75]. During the users' mid-ZPD phase, this study frequently observed vague and incorrect answers. These types of responses reflected that the users were at least trying to answer the previous question or perform the task. Therefore, the CA applied strategies such as follow-up questions or hints to provide necessary information or suggestions to facilitate the user's understanding of the task. Unlike the scaffolding CA designed for learning which provided detailed explanations and correct answers when the user made mistakes [77], our CA for the NCD screening scenario was not allowed to do that because of the potential risk of affecting the screening results. Mistakes and unfinished tasks are important and taken into account in the final scoring of the GSDT [21]. Even when the participants were near the ORZ and kept making invalid responses in the study, the CA's hints were still within the legitimate range. If the CA's final attempt to restart the task proved unsuccessful, the CA would terminate the task. While not every participant completed the entire task, which involved purchasing all the ingredients, the minimum number of purchased ingredients in this study was one.

*6.2.3 Automation of NCD screening Tasks.* Though the CA system in this study was semi-automated and controlled by the human wizards, the integration with the large language model, ChatGPT, could provide implications for designing a fully automated CA for dialog tasks of NCD screening in the future. The model ChatGPT exhibited effectiveness in recommending appropriate strategies based on our scaffolding framework and prompt design without model training or fine-tuning in the study. We took advantage of ChatGPT's zero-shot learning ability [58] to accomplish fast prototyping of the CA. Since ChatGPT 3.5 did not have the ability of visually understanding, we embedded all the visual information, e.g., the token's positions and movements, in the prompt template as shown in Fig. 11. However, this approach might cause information loss and further result in inappropriate scaffolding strategy recommendations. Though the ideal solution was to implement language models with vision-language understanding ability, available models [39, 84] at present had difficulty comprehending spatial information in the grocery map image. To address this limitation, a rule-based module could be introduced to summarize the visual information in the grocery, given that the token's state can be readily captured by the test simulator. For example, "the target shelf is on the right of the token", "the token needs to go straight all the way to the end of this aisle to find tomatoes".

We found that the message history parameter acceptable by ChatGPT API contributed little to the generation. Adding a module to summarize key actions that happened previously, instead of relying on the array of message history, might help improve. We tested this idea initially and presented examples in the supplementary material. Our results implied that the token state information was important but had limited effects on the generation. One possible reason was that the token state was not significantly highlighted or emphasized in the prompt. According to the implication of prior work [13], the token state information should be explicitly emphasized or repeated by the natural language in the prompt.

## 6.3 Generalizability

Though this work focused on a specific task, GSDT [21], it can be generalized in three forms. First, the CA's scaffolding strategies may be generalized in other NCD screening tests. Scaffolding for NCD patients is usually deployed verbally [26] and many cognitive tests contain dialog parts, e.g., the Cookie Picture description task in DementiaBank [7]. Table 7 lists examples of the CA's scaffolding strategy application in other cognitive tests. Second, our scaffolding framework has generalizability in the design of CAs for other daily activities of older adults. CAs have been deployed in family and hospital settings [10, 64, 80]. Meanwhile, the CA's scaffolding skills would be critical in NCD patients' daily activity support [20, 25, 26, 32]. Thus, our scaffolding framework may support the organization and completion of these activities for the CA and the user. Third, our design process may be generalized in designing other types of conversational skills for CAs. CAs have been equipped with various conversational skills such as backchanneling [14] and active listening [79]. Our design process rooted in existing theories and practical data on conversational skills was validated as effective. Moreover, embedding the large language model in iterative prototyping could also facilitate the design of CAs' various conversational skills based on natural language.

# 7 LIMITATIONS

There are limitations in this work. First, the sample size was limited. Future work may conduct user studies with a larger sample size and statistical tests to further verify the results. Second, the

[7] https://dementia.talkbank.org/protocol/

**Table 7: The possible application of the CA's scaffolding strategies in other cognitive tests, based on the administration scripts of MoCA [50], ACE-R [47] and DementiaBank.**

| Scaffolding Level | Scaffolding Strategy | Examples in Other Tests |
|---|---|---|
| Activity | Scaffolding Question | *Can you name this animal?* (MoCA)<br>*Can you describe everything happening in this picture?* (DementiaBank)<br>*Now can you name as many animals as possible, beginning with any letter?* (ACE-R) |
| Action | Continuation | *What's the next (animal)?* (MoCA)<br>*What happened next?* (DementiaBank) |
| | Follow-up Questions | *Tell me the exact date and day of the week.* (MoCA)<br>*Tell me which city this place is in.* (MoCA) |
| Repair | Encourage | *Take a look (point to the picture).* (DementiaBank) |
| | Hint | *It's a part of the body.* (MoCA)<br>*The animal names don't have to begin with "P".* (ACE-R)<br>*OK, I'll give you some hints: was the name X, Y or Z?* (ACE-R) |

prototype of CA was semi-automated with the human wizards' operations. Future work may explore the implementation of the fully automated CA to administer the NCD screening test and verify the validity of the screening result. The semi-automation also caused the issue of system latency. Though participants and clinical experts considered the latency acceptable in this study, further investigation and solutions for the latency issue are needed because it might influence the spatial navigation performance of people with memory decrements [31]. Third, this work focused on one specific task, the GSDT. The generalizability of other dialog tasks for NCD screening needs to be further evaluated as mentioned in section 6.3. Last, NCD screening tests are advised to be conducted on a routine basis [53], however, the participants in this work only finished one time of the test in the hospital setting. Future work may explore the long-term use of the CA for NCD screening in other settings such as home environments. The involvement of family members might also affect the conversational dynamics of the CA [79].

## 8 CONCLUSION

Conversational agents (CA) show great potential in aiding digital cognitive assessment and enabling its scalability for early detection of NCDs. This work explored designing the CA's scaffolding strategies which were critical in dialog tasks for people with potential NCDs. Based on existing scaffolding theories and clinical practices in human-administered NCD screening (the GSDT) data of 12 subjects, we proposed a framework for the CA to provide appropriate scaffolding strategies that were aligned to the user's ZPD status. Through an iterative process, we built up a semi-automated CA prototype empowered by ChatGPT based on the scaffolding framework. A mixed-method user study was conducted with 15 participants including five with major NCDs and five with mild NCDs. The CA administered the GSDT with the participants and provided scaffolding to guide them. The results showed that about 89% of the CA-adopted scaffolding strategies were evaluated as appropriate by the clinical experts. The participants perceived the CA as easy to use and showed a willingness to use it for NCD screening. During the dialog task, the CA's scaffolding facilitated the participant's task understanding. Finally, we proposed implications for the future design of CAs for NCD screening in terms of integrating the user's ZPD status and the three-level scaffolding framework, as well as considerations of automation.

## ACKNOWLEDGMENTS

This project is supported by the Hong Kong SAR Research Grants Council's Theme-based Research Grant Scheme (Project No. T45-407/19N). The assistance provided by Professor Vincent Mok's team (with special thanks to Wing Lam Kwan and Man Ho So), Dr. Ka Ho Wong, and Hang Yu Wong is greatly appreciated.

## REFERENCES

[1] Asangaedem Akpan, Maturin Tabue-Teguo, and Bertrand Fougère. 2019. Neurocognitive Disorders: Importance of Early/Timely Detection in Daily Clinical Practice. *Journal of Alzheimer's disease: JAD* 70, 2 (2019), 317—322. https://doi.org/10.3233/jad-180381

[2] Patricia Albacete, Pamela Jordan, Sandra Katz, Irene-Angelica Chounta, and Bruce M. McLaren. 2019. The Impact of Student Model Updates on Contingent Scaffolding in a Natural-Language Tutoring System. In *Artificial Intelligence in Education*, Seiji Isotani, Eva Millán, Amy Ogan, Peter Hastings, Bruce McLaren, and Rose Luckin (Eds.). Springer International Publishing, Cham, 37–47.

[3] Jesús B. Alonso-Hernández, María Luisa Barragán-Pulido, José Manuel Gil-Bordón, Miguel Ángel Ferrer-Ballester, and Carlos M. Travieso-González. 2021. Using a Human Interviewer or an Automatic Interviewer in the Evaluation of Patients with AD from Speech. *Applied Sciences* 11, 7 (2021), 3228.

[4] Rabeea'h W. Aslam, Vickie Bates, Yenal Dundar, Juliet Hounsome, Marty Richardson, Ashma Krishan, Rumona Dickson, Angela Boland, Joanne Fisher, Louise Robinson, and Sudip Sikdar. 2018. A systematic review of the diagnostic accuracy of automated tests for cognitive impairment. *International Journal of Geriatric Psychiatry* 33, 4 (2018), 561–575. https://doi.org/10.1002/gps.4852

[5] Alzheimer's Association. 2023. 2023 Alzheimer's disease facts and figures. *Alzheimer's & Dementia* 19, 4 (2023), 1598–1695. https://doi.org/10.1002/alz.13016

[6] James T. Becker, François Boiler, Oscar L. Lopez, Judith Saxton, and Karen L. McGonigle. 1994. The Natural History of Alzheimer's Disease: Description of Study Cohort and Accuracy of Diagnosis. *Archives of Neurology* 51, 6 (Jun 1994), 585–594. https://doi.org/10.1001/archneur.1994.00540180063015

[7] David Bissig, Jeffrey Kaye, and Deniz Erten-Lyons. 2020. Validation of SATURN, a free, electronic, self-administered cognitive screening test. *Alzheimer's & Dementia: Translational Research & Clinical Interventions* 6, 1 (Jan. 2020), e12116. https://doi.org/10.1002/trc2.12116

[8] Som S. Biswas. 2023. Role of Chat GPT in Public Health. *Annals of Biomedical Engineering* 51, 5 (May 2023), 868–869. https://doi.org/10.1007/s10439-023-03172-7

[9] Eduardo Carrasco, Gorka Epelde, Aitor Moreno, Amalia Ortiz, Igor Garcia, Cristina Buiza, Elena Urdaneta, Aitziber Etxaniz, Mari Feli González, and Andoni

Arruti. 2008. Natural Interaction between Avatars and Persons with Alzheimer's Disease. In *Computers Helping People with Special Needs*, Klaus Miesenberger, Joachim Klaus, Wolfgang Zagler, and Arthur Karshmer (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 38–45.

[10] Clare Carroll, Catherine Chiodo, Adena Xin Lin, Meg Nidever, and Jayanth Prathipati. 2017. Robin: Enabling Independence For Individuals With Cognitive Disabilities Using Voice Assistive Technology. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI EA '17)*. Association for Computing Machinery, New York, NY, USA, 46–53. https://doi.org/10.1145/3027063.3049266

[11] Breda Cullen, Brian O'Neill, Jonathan J Evans, Robert F Coen, and Brian A Lawlor. 2007. A review of screening tests for cognitive impairment. *Journal of Neurology, Neurosurgery & Psychiatry* 78, 8 (2007), 790–799. https://doi.org/10.1136/jnnp.2006.095414

[12] Louise Cummings. 2019. Describing the Cookie Theft picture: Sources of breakdown in Alzheimer's dementia. , 153–176 pages. https://doi.org/10.1075/ps.17011.cum

[13] Paul Denny, Viraj Kumar, and Nasser Giacaman. 2023. Conversing with Copilot: Exploring Prompt Engineering for Solving CS1 Problems Using Natural Language. In *Proceedings of the 54th ACM Technical Symposium on Computer Science Education V. 1* (Toronto ON, Canada) *(SIGCSE 2023)*. Association for Computing Machinery, New York, NY, USA, 1136–1142. https://doi.org/10.1145/3545945.3569823

[14] Zijian Ding, Jiawen Kang, Tinky Oi Ting HO, Ka Ho Wong, Helene H Fung, Helen Meng, and Xiaojuan Ma. 2022. TalkTive: A Conversational Agent Using Backchannels to Engage Older Adults in Neurocognitive Disorders Screening. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, Article 304, 19 pages. https://doi.org/10.1145/3491102.3502005

[15] Eva Eggenberger, Katharina Heimerl, and Michael I. Bennett. 2013. Communication skills training in dementia care: a systematic review of effectiveness, training content, and didactic methods in different care settings. *International Psychogeriatrics* 25, 3 (2013), 345–358. https://doi.org/10.1017/S1041610212001664

[16] Shahla Farzana, Mina Valizadeh, and Natalie Parde. 2020. Modeling Dialogue in Conversational Cognitive Health Screening Interviews. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*. European Language Resources Association, Marseille, France, 1167–1177. https://aclanthology.org/2020.lrec-1.147

[17] Cleusa P. Ferri, Martin Prince, Carol Brayne, Henry Brodaty, Laura Fratiglioni, Mary Ganguli, Kathleen Hall, Kazuo Hasegawa, Hugh Hendrie, Yueqin Huang, Anthony Jorm, Colin Mathers, Paulo R. Menezes, Elizabeth Rimmer, and Marcia Scazufca. 2005. Global prevalence of dementia: a Delphi consensus study. *The Lancet* 366, 9503 (2005), 2112–2117. https://doi.org/10.1016/S0140-6736(05)67889-0

[18] Kathleen C Fraser, Jed A Meltzer, and Frank Rudzicz. 2016. Linguistic features identify Alzheimer's disease in narrative speech. *Journal of Alzheimer's Disease* 49, 2 (2016), 407–422. https://doi.org/10.3233/JAD-150520

[19] S Gauthier, C Webster, S Servaes, J Morais, and P Rosa-Neto. 2022. World Alzheimer Report 2022. *Alzheimer's Disease International: London, UK* (2022).

[20] Trude Gjernes and Per Måseide. 2019. Framing and scaffolding as relational caregiving in an institution for people living with dementia. *Journal of Aging Studies* 49 (Jun 2019), 39–45. https://doi.org/10.1016/j.jaging.2019.04.001

[21] Xianmin Gong, Patrick CM Wong, Helene H Fung, Vincent CT Mok, Timothy CY Kwok, Jean Woo, Ka Ho Wong, and Helen Meng. 2022. The Hong Kong Grocery Shopping Dialog Task (HK-GSDT): A Quick Screening Test for Neurocognitive Disorders. *International Journal of Environmental Research and Public Health* 19, 20 (2022), 13302.

[22] Harold Goodglass and Edith Kaplan. 2001. *BDAE: The Boston diagnostic aphasia examination.*

[23] R. Chulaka Gunasekara, David Nahamoo, Lazaros C. Polymenakos, David Echevería Ciaurri, Jatin Ganhotra, and Kshitij P. Fadnis. 2019. Quantized Dialog – A general approach for conversational systems. *Computer Speech & Language* 54 (2019), 17–30. https://doi.org/10.1016/j.csl.2018.06.003

[24] Björn Hartmann, Scott R. Klemmer, Michael Bernstein, Leith Abdulla, Brandon Burr, Avi Robinson-Mosher, and Jennifer Gee. 2006. Reflective Physical Prototyping through Integrated Design, Test, and Analysis. In *Proceedings of the 19th Annual ACM Symposium on User Interface Software and Technology (UIST '06)*. Association for Computing Machinery, New York, NY, USA, 299–308. https://doi.org/10.1145/1166253.1166300 event-place: Montreux, Switzerland.

[25] Lars-Christer Hydén. 2011. Narrative collaboration and scaffolding in dementia. *Journal of Aging Studies* 25, 4 (Dec 2011), 339–347. https://doi.org/10.1016/j.jaging.2011.04.002

[26] Lars-Christer Hydén and Mattias Forsblad. 2017. *Collaborative Remembering in Dementia: A Focus on Joint Activities.* Oxford University Press, 0. https://doi.org/10.1093/oso/9780198737865.003.0025

[27] Lars-Christer Hydén. 2014. Cutting Brussels sprouts: Collaboration involving persons with dementia. *Journal of Aging Studies* 29 (2014), 115–123. https://doi.org/10.1016/j.jaging.2014.02.004

[28] Eun-Young Kang, Sung Ju Jee, Cuk-Seong Kim, Kwang-Sun Suh, Alex W. K. Wong, and Jae Young Moon. 2018. The feasibility study of Computer Cognitive Senior Assessment System-Screen (CoSAS-S) in critically ill patients with sepsis. *Journal of Critical Care* 44 (2018), 128–133. https://doi.org/10.1016/j.jcrc.2017.10.005

[29] Sandra Katz, Patricia Albacete, Irene-Angelica Chounta, Pamela Jordan, Bruce M. McLaren, and Diego Zapata-Rivera. 2021. Linking Dialogue with Student Modelling to Create an Adaptive Tutoring System for Conceptual Physics. *International Journal of Artificial Intelligence in Education* 31, 3 (Sep 2021), 397–445. https://doi.org/10.1007/s40593-020-00226-y

[30] J. F. Kelley. 1984. An Iterative Design Methodology for User-Friendly Natural Language Office Information Applications. *ACM Trans. Inf. Syst.* 2, 1 (jan 1984), 26–41. https://doi.org/10.1145/357417.357420

[31] Roy P. C. Kessels, Amy van Doormaal, and Gabriele Janzen. 2011. Landmark Recognition in Alzheimer's Dementia: Spared Implicit Memory for Objects Relevant for Navigation. *PLOS ONE* 6, 4 (April 2011), e18611. https://doi.org/10.1371/journal.pone.0018611

[32] Jacqueline Kindell, John Keady, Karen Sage, and Ray Wilkinson. 2017. Everyday conversation in dementia: a review of the literature to inform research and practice. *International Journal of Language & Communication Disorders* 52, 4 (2017), 392–406. https://doi.org/10.1111/1460-6984.12298

[33] Emre Kokmen, Glenn E. Smith, Ronald C. Petersen, Eric Tangalos, and Robert C. Ivnik. 1991. The Short Test of Mental Status: Correlations With Standardized Psychometric Testing. *Archives of Neurology* 48, 7 (Jul 1991), 725–728. https://doi.org/10.1001/archneur.1991.00530190071018

[34] Alexandra König, Nicklas Linz, Johannes Tröger, Maria Wolters, Jan Alexandersson, and Phillipe Robert. 2018. Fully Automatic Speech-Based Analysis of the Semantic Verbal Fluency Task. *Dementia and Geriatric Cognitive Disorders* 45, 3-4 (06 2018), 198–209. https://doi.org/10.1159/000487852

[35] E. E. Kravtsova. 2009. The Cultural-Historical Foundations of the Zone of Proximal Development. *Journal of Russian & East European Psychology* 47, 6 (2009), 9–24. https://doi.org/10.2753/RPO1061-0405470601

[36] Mohammad Amin Kuhail, Nazik Alturki, Salwa Alramlawi, and Kholood Alhejori. 2023. Interacting with educational chatbots: A systematic review. *Education and Information Technologies* 28, 1 (Jan 2023), 973–1018. https://doi.org/10.1007/s10639-022-11177-3

[37] Hedda Lausberg and Han Sloetjes. 2009. Coding gestural behavior with the NEUROGES-ELAN system. *Behavior Research Methods* 41, 3 (Aug 2009), 841–849. https://doi.org/10.3758/BRM.41.3.841

[38] Yoonjoo Lee, John Joon Young Chung, Tae Soo Kim, Jean Y Song, and Juho Kim. 2022. Promptiverse: Scalable Generation of Scaffolding Prompts Through Human-AI Hybrid Knowledge Graph Annotation. 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 96, 18 pages. https://doi.org/10.1145/3491102.3502087

[39] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. arXiv:2301.12597 [cs.CV]

[40] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-Train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *ACM Comput. Surv.* 55, 9, Article 195 (jan 2023), 35 pages. https://doi.org/10.1145/3560815

[41] Wendy J Lorentz, James M Scanlan, and Soo Borson. 2002. Brief Screening Tests for Dementia. *The Canadian Journal of Psychiatry* 47, 8 (2002), 723–733. https://doi.org/10.1177/070674370204700803 PMID: 12420650.

[42] Tivadar Lucza, Kázmér Karádi, János Kállai, Rita Weintraut, József Janszky, Attila Makkos, Sámuel Komoly, and Norbert Kovács. 2015. Screening Mild and Major Neurocognitive Disorders in Parkinson's Disease. *Behavioural Neurology* 2015 (May 2015). https://doi.org/10.1155/2015/983606

[43] Francesca Lunardini, Matteo Luperto, Marta Romeo, Nicola Basilico, Katia Daniele, Domenico Azzolino, Sarah Damanti, Carlo Abbate, Daniela Mari, Matteo Cesari, Nunzio Alberto Borghese, and Simona Ferrante. 2020. Supervised Digital Neuropsychological Tests for Cognitive Decline in Older Adults: Usability and Clinical Validity Study. *JMIR Mhealth Uhealth* 8, 9 (Sep 2020), e17963. https://doi.org/10.2196/17963

[44] Saturnino Luz, Fasih Haider, Sofia de la Fuente, Davida Fromm, and Brian MacWhinney. 2020. Alzheimer's Dementia Recognition Through Spontaneous Speech: The ADReSS Challenge. In *Proc. Interspeech 2020*. 2172–2176. https://doi.org/10.21437/Interspeech.2020-2571

[45] Hyuma Makizako, Hiroyuki Shimada, Hyuntae Park, Takehiko Doi, Daisuke Yoshida, Kazuki Uemura, Kota Tsutsumimoto, and Takao Suzuki. 2013. Evaluation of multidimensional neurocognitive function using a tablet personal computer: Test–retest reliability and validity in community-dwelling older adults. *Geriatrics & Gerontology International* 13, 4 (2013), 860–866. https://doi.org/10.1111/ggi.12014

[46] Lisa Mikesell. 2009. Conversational Practices of a Frontotemporal Dementia Patient and His Interlocutors. *Research on Language and Social Interaction* 42, 2 (2009), 135–162. https://doi.org/10.1080/08351810902864552

[47] Eneida Mioshi, Kate Dawson, Joanna Mitchell, Robert Arnold, and John R. Hodges. 2006. The Addenbrooke's Cognitive Examination Revised (ACE-R): a brief cognitive test battery for dementia screening. *International Journal of Geriatric Psychiatry* 21, 11 (2006), 1078–1085. https://doi.org/10.1002/gps.1610

[48] Bahman Mirheidari, Daniel Blackburn, Traci Walker, Markus Reuber, and Heidi Christensen. 2019. Dementia detection using automatic analysis of conversations. *Computer Speech & Language* 53 (2019), 65–79. https://doi.org/10.1016/j.csl.2018.07.006

[49] Carmen Moret-Tatay, Isabel Iborra-Marmolejo, María José Jorques-Infante, José Vicente Esteve-Rodrigo, Carla H. A. Schwanke, and Tatiana Q. Irigaray. 2021. Can Virtual Assistants Perform Cognitive Assessment in Older Adults? A Review. *Medicina* 57, 12 (2021). https://doi.org/10.3390/medicina57121310

[50] Ziad S. Nasreddine, Natalie A. Phillips, Valérie Bédirian, Simon Charbonneau, Victor Whitehead, Isabelle Collin, Jeffrey L. Cummings, and Howard Chertkow. 2005. The Montreal Cognitive Assessment, MoCA: A Brief Screening Tool For Mild Cognitive Impairment. *Journal of the American Geriatrics Society* 53, 4 (2005), 695–699. https://doi.org/10.1111/j.1532-5415.2005.53221.x

[51] Keiichi Onoda, Tsuyoshi Hamano, Yoko Nabika, Atsuo Aoyama, Hiroyuki Takayoshi, Tomonori Nakagawa, Masaki Ishihara, Shingo Mitaki, Takuya Yamaguchi, Hiroaki Oguro, Kuninori Shiwaku, and Shuhei Yamaguchi. 2013. Validation of a new mass screening tool for cognitive impairment: Cognitive Assessment for Dementia, iPad version. *Clinical Interventions in Aging* 8 (2013), 353–360. https://doi.org/10.2147/CIA.S42342

[52] World Health Organization et al. 2021. Global status report on the public health response to dementia. (2021).

[53] S. T. Pendlebury, S. P. Klaus, M. Mather, M. de Brito, and R. M. Wharton. 2015. Routine cognitive screening in older patients admitted to acute medicine: abbreviated mental test score (AMTS) and subjective memory complaint versus Montreal Cognitive Assessment and IQCODE. *Age and Ageing* 44, 6 (10 2015), 1000–1005. https://doi.org/10.1093/ageing/afv134

[54] Deborah Pino-Pasternak, David Whitebread, and Andrew Tolmie. 2010. A Multidimensional Analysis of Parent–Child Interactions During Academic Tasks and Their Relationships With Children's Self-Regulated Learning. *Cognition and Instruction* 28, 3 (Jul 2010), 219–272. https://doi.org/10.1080/07370008.2010.490494

[55] W. James Potter and Deborah Levine-Donnerstein. 1999. Rethinking validity and reliability in content analysis. *Journal of Applied Communication Research* 27, 3 (Aug 1999), 258–284. https://doi.org/10.1080/00909889909365539

[56] Margherita Rampioni, Vera Stara, Elisa Felici, Lorena Rossi, and Susy Paolini. 2021. Embodied Conversational Agents for Patients With Dementia: Thematic Literature Analysis. *JMIR Mhealth Uhealth* 9, 7 (Jul 2021), e25381. https://doi.org/10.2196/25381

[57] Leon Reicherts, Gun Woo Park, and Yvonne Rogers. 2022. Extending Chatbots to Probe Users: Enhancing Complex Decision-Making Through Probing Conversations. In *Proceedings of the 4th Conference on Conversational User Interfaces* (Glasgow, United Kingdom) *(CUI '22)*. Association for Computing Machinery, New York, NY, USA, Article 2, 10 pages. https://doi.org/10.1145/3543829.3543832

[58] Laria Reynolds and Kyle McDonell. 2021. Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI EA '21)*. Association for Computing Machinery, New York, NY, USA, Article 314, 7 pages. https://doi.org/10.1145/3411763.3451760

[59] Danielle N. Ripich, Elaine Ziol, Thomas Fritsch, and Ellen J. Durand. 2000. Training Alzheimer's Disease Caregivers for Successful Communication. *Clinical Gerontologist* 21, 1 (2000), 37–56. https://doi.org/10.1300/J018v21n01_05

[60] Nicole Ruggiano, Ellen L Brown, Lisa Roberts, C Victoria Framil Suarez, Yan Luo, Zhichao Hao, and Vagelis Hristidis. 2021. Chatbots to Support People With Dementia and Their Caregivers: Systematic Review of Functions and Quality. *J Med Internet Res* 23, 6 (3 Jun 2021), e25006. https://doi.org/10.2196/25006

[61] Steven R Sabat. 1991. Facilitating conversation via indirect repair: A case study of Alzheimer's disease. *Georgetown Journal of Languages and Linguistics* 2, 3 (1991), 284–96.

[62] Anne Snowdon, Abdulkadir Hussein, Robert Kent, Lou Pino, and Vladimir Hachinski. 2015. Comparison of an Electronic and Paper-based Montreal Cognitive Assessment Tool. *Alzheimer Disease & Associated Disorders* 29, 4 (2015), 325–329. https://doi.org/doi:10.1097/WAD.0000000000000069

[63] Vera Stara, Michiel de Jong, Elisa Felici, Daniel Bolliger, Edith Birrer, Viviane von Döllen, Lorena Rossi, and Marcel Heerink. 2020. The Design Adaptation of the Virtual Assistant Anne for Moderate Dementia Patients and Their Formal Caregivers in Protected Environment Tests. In *Advances in Human Factors and Ergonomics in Healthcare and Medical Devices*, Nancy J. Lightner and Jay Kalra (Eds.). Springer International Publishing, Cham, 270–279.

[64] Vera Stara, Benjamin Vera, Daniel Bolliger, Susy Paolini, Michiel de Jong, Elisa Felici, Stephanie Koenderink, Lorena Rossi, Viviane Von Doellen, and Mirko di Rosa. 2021. Toward the Integration of Technology-Based Interventions in the Care Pathway for People with Dementia: A Cross-National Study. *International Journal of Environmental Research and Public Health* 18, 19 (2021). https://doi.org/10.3390/ijerph181910405

[65] Alan B. Stevens, Catherine A. King, and Cameron J. Camp. 1993. IMPROVING PROSE MEMORY AND SOCIAL INTERACTION USING QUESTION ASKING READING WITH ADULT DAY CARE CLIENTS. *Educational Gerontology* 19, 7 (1993), 651–662. https://doi.org/10.1080/0360127930190706

[66] Chiara F. Tagliabue, David Bissig, Jeffrey Kaye, Veronica Mazza, and Sara Assecondi. 2023. Feasibility of Remote Unsupervised Cognitive Screening With SATURN in Older Adults. *Journal of Applied Gerontology* 42, 9 (2023), 1903–1910. https://doi.org/10.1177/07334648231166894

[67] David F. Tang-Wai, Eric E. Smith, Marie-Andrée Bruneau, Amer M. Burhan, Atri Chatterjee, Howard Chertkow, Samira Choudhury, Ehsan Dorri, Simon Ducharme, Corinne E. Fischer, Sheena Ghodasara, Nathan Herrmann, Ging-Yuek Robin Hsiung, Sanjeev Kumar, Robert Laforce Jr, Linda Lee, Fadi Massoud, Kenneth I. Shulman, Michael Stiffel, Serge Gauthier, and Zahinoor Ismail. 2020. CCCDTD5 recommendations on early and timely assessment of neurocognitive disorders using cognitive, behavioral, and functional scales. *Alzheimer's & Dementia: Translational Research & Clinical Interventions* 6, 1 (2020), e12057. https://doi.org/10.1002/trc2.12057

[68] Stergios Tegos, Stavros Demetriadis, and Anastasios Karakostas. 2015. Promoting academically productive talk with conversational agent interventions in collaborative learning settings. *Computers & Education* 87 (2015), 309–325. https://doi.org/10.1016/j.compedu.2015.07.014

[69] Stergios Tegos, Georgios Psathas, Thrasyvoulos Tsiatsos, Christos Katsanos, Anastasios Karakostas, Costas Tsibanis, and Stavros Demetriadis. 2020. Enriching Synchronous Collaboration in Online Courses with Configurable Conversational Agents. In *Intelligent Tutoring Systems*, Vivekanandan Kumar and Christos Troussas (Eds.). Springer International Publishing, Cham, 284–294.

[70] Janneke van de Pol, Monique Volman, Frans Oort, and Jos Beishuizen. 2015. The effects of scaffolding in the classroom: support contingency and student independent working time in relation to student achievement, task effort and appreciation of support. *Instructional Science* 43, 5 (Sep 2015), 615–641. https://doi.org/10.1007/s11251-015-9351-z

[71] Lev Semenovich Vygotsky and Michael Cole. 1978. *Mind in Society: Development of Higher Psychological Processes*. Harvard University Press. 86 pages.

[72] Sarah E. Wallace, Elena V. Donoso Brown, Andrea D. Fairman, Koren Beardshall, Anna Olexsovich, Alicia Taylor, and James B. Schreiber. 2017. Validation of the Standardized Touchscreen Assessment of Cognition with neurotypical adults. *NeuroRehabilitation* 40, 3 (2017), 411–420. https://doi.org/10.3233/NRE-161428

[73] Pierre Wargnier, Samuel Benveniste, Pierre Jouvelot, and Anne-Sophie Rigaud. 2018. Usability assessment of interaction management support in LOUISE, an ECA-based user interface for elders with cognitive impairment. *Technology and Disability* 30, 3 (2018), 105–126.

[74] Pierre Wargnier, Adrien Malaisé, Julien Jacquemot, Samuel Benveniste, Pierre Jouvelot, Maribel Pino, and Anne-Sophie Rigaud. 2015. Towards attention monitoring of older adults with cognitive impairment during interaction with an embodied conversational agent. In *2015 3rd IEEE VR International Workshop on Virtual and Augmented Assistive Technology (VAAT)*. 23–28. https://doi.org/10.1109/VAAT.2015.7155406

[75] Stephen Wey. 2006. Working in the zone–a social ecological framework for dementia rehabilitation. *Assistive Technology in Dementia Care. London: Hawker Publications* (2006), 81–99.

[76] Rainer Winkler, Claudio Büchi, and Matthias Söllner. 2019. Improving Problem-Solving Skills with Smart Personal Assistants: Insights from a Quasi Field Experiment. In *International Conference on Interaction Sciences*.

[77] Rainer Winkler, Sebastian Hobert, Antti Salovaara, Matthias Söllner, and Jan Marco Leimeister. 2020. Sara, the Lecturer: Improving Learning in Online Education with a Scaffolding-Based Conversational Agent. 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, 1–14. https://doi.org/10.1145/3313831.3376781

[78] David Wood, Jerome S. Bruner, and Gail Ross. 1976. THE ROLE OF TUTORING IN PROBLEM SOLVING*. *Journal of Child Psychology and Psychiatry* 17, 2 (1976), 89–100. https://doi.org/10.1111/j.1469-7610.1976.tb00381.x

[79] Ziang Xiao, Michelle X. Zhou, Wenxi Chen, Huahai Yang, and Changyan Chi. 2020. If I Hear You Correctly: Building and Evaluating Interview Chatbots with Active Listening Skills.

[80] Teng Le Xin, Amad Arshad, and Zailan Arabee bin Abdul Salam. 2021. AlzBot-Mobile App Chatbot for Alzheimer's Patient to be Active with Their Minds. In *2021 14th International Conference on Developments in eSystems Engineering (DeSE)*. 124–129. https://doi.org/10.1109/DeSE54285.2021.9719410

[81] Albert Deok-Young Yang, Yeo-Gyeong Noh, and Jin-Hyuk Hong. 2021. Topic Recommendation to Expand Knowledge and Interest in Question-and-Answer Agents. *Applied Sciences* 11, 22 (2021). https://doi.org/10.3390/app112210600

[82] Zi Ye, Shoukang Hu, Jinchao Li, Xurong Xie, Mengzhe Geng, Jianwei Yu, Junhao Xu, Boyang Xue, Shansong Liu, Xunying Liu, and Helen Meng. 2021. Development of the Cuhk Elderly Speech Recognition System for Neurocognitive Disorder Detection Using the Dementiabank Corpus. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 6433–6437. https://doi.org/10.1109/ICASSP39728.2021.9413634

[83] J.D. Zamfirescu-Pereira, Richmond Y. Wong, Bjoern Hartmann, and Qian Yang. 2023. Why Johnny Can't Prompt: How Non-AI Experts Try (and Fail) to Design LLM Prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 437, 21 pages. https://doi.org/10.1145/3544548.3581388

[84] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. *arXiv preprint arXiv:2304.10592* (2023).

[85] Tamara Zubatiy, Kayci L Vickers, Niharika Mathur, and Elizabeth D Mynatt. 2021. Empowering Dyads of Older Adults With Mild Cognitive Impairment And Their Care Partners Using Conversational Agents. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 657, 15 pages. https://doi.org/10.1145/3411764.3445124

[86] Dilawar Shah Zwakman, Debajyoti Pal, and Chonlameth Arpnikanondt. 2021. Usability Evaluation of Artificial Intelligence-Based Voice Assistants: The Case of Amazon Alexa. *SN Computer Science* 2, 1 (Jan. 2021), 28. https://doi.org/10.1007/s42979-020-00424-4

# A APPENDIX


**Figure 10: The simulated GSDT test.**

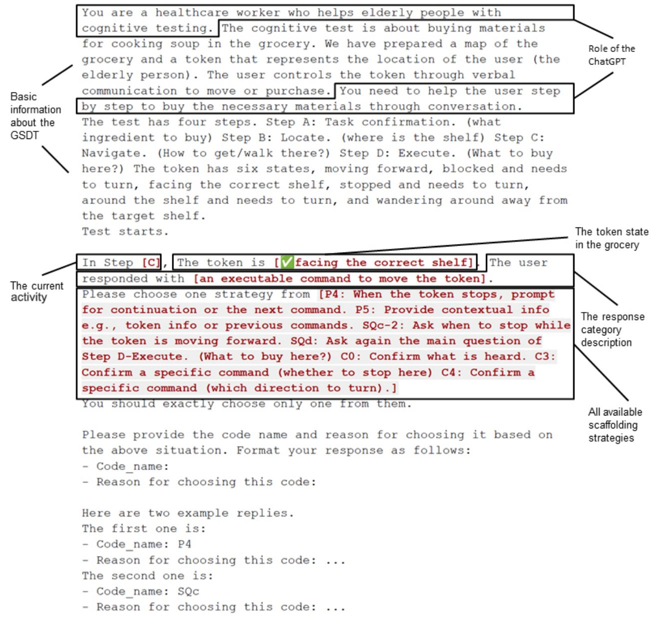


**Figure 11: An instance of the prompt template for ChatGPT's recommended scaffolding strategy.**