Towards Conversational Patient History-Taking: Voice-Interactive AI Agents for Pre-visit Dementia Diagnostic Interviews

Andrew G. Breithaupt[1*] **Nayoung Choi**[2*] **James D. Finch**[2]
Jeanne M. Powell[1] **Arin L. Nelson**[1] **Oz A. Alon**[3]
Howard J. Rosen[4] **Jinho D. Choi**[2]

[1] Neurology, Emory University [2]Computer Science, Emory University
[3]Arts and Sciences, Emory University [4]Neurology, University of California, San Francisco

{abreith, nchoi27, jdfinch, jmpowe6, arin.nelson, oalon, jinho.choi}@emory.edu, howie.rosen@ucsf.edu

*Equal contribution

Abstract

Patient history-taking is a critical yet time-intensive component of clinical diagnosis, frequently hindered by time-constrained clinical visits. We present an LLM-based voice-interactive conversational system for conducting semi-structured diagnostic interviews with older adults suspected of Alzheimer's disease and related dementias (ADRD). The system features conditional conversation branching over specialist-developed interview scripts and interaction adaptations tailored for older adults. In a within-subjects study with 30 participants from a cognitive neurology clinic, we compare two LLM prompting strategies through dialogue analysis, assess user experience, and evaluate symptom elicitation against a consensus gold standard adjudicated from both agent and specialist interviews. Our findings reveal that prompting strategy distinctly shapes both conversational dynamics and symptom coverage. Together with high specificity scores and positive user experience ratings, these results demonstrate the potential of LLM-based conversational agents for scalable, patient-centered history-taking in ADRD care.

1 Introduction

Patient history-taking, the structured elicitation of a patient's symptom history and relevant background, is a cornerstone of clinical diagnosis, yet time-constrained clinical visits make comprehensive history collection increasingly difficult in practice (Linzer et al., 2015; Liu et al., 2024a; Boustani et al., 2005). This challenge is particularly acute for Alzheimer's disease and related dementias (ADRD), where symptoms are insidious in onset, complex, and often hard for patients to articulate, requiring probing questions and extended dialogue (see Appendix A.1). While prior work has explored scalable approaches to clinical documentation and symptom extraction from existing notes (Saley et al., 2024; Shim et al., 2021; Wang et al., 2024), fewer systems actively elicit symptom information through real-time dialogue, particularly for complex conditions like ADRD. How to design and deploy such systems for real clinical settings remains an open question.

Conversational AI agents powered by large language models (LLMs) offer a promising path toward scalable history-taking, as LLMs enable flexible, context-sensitive dialogue without rigid state machines. However, deploying such systems in clinical settings requires careful attention to several design dimensions. First, effective history-taking requires systematic coverage of diagnostically relevant topics while remaining flexible enough to follow the patient's narrative, requiring the agent to probe for detail, handle ambiguous responses, and manage topic transitions. Second, deploying such a system further requires sensitivity to the target population. Older adults with cognitive impairment may speak slowly, struggle with complex phrasing, or require additional scaffolding (Pradhan et al., 2019; Huang et al., 2025). ADRD interviews further involve a multi-party dynamic in which an informant, typically a spouse or adult child, participates alongside the patient, introducing additional complexity in turn-taking and role management.

This work explores the feasibility of deploying a conversational AI agent as a pre-visit clinical support tool. We develop a voice-interactive system for ADRD diagnostic history-taking, powered by an LLM that conducts semi-structured interviews using a specialist-developed script with conditional conversation branching and interaction design tailored for older adults (Section 3). We evaluate the system in a within-subjects study with 30 older adults with suspected cognitive impairment from a major academic medical center's cognitive neurology clinic (Section 4). Our main contributions can be summarized as follows:

Proceedings of the 27th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 62–78
August 2–5, 2026.

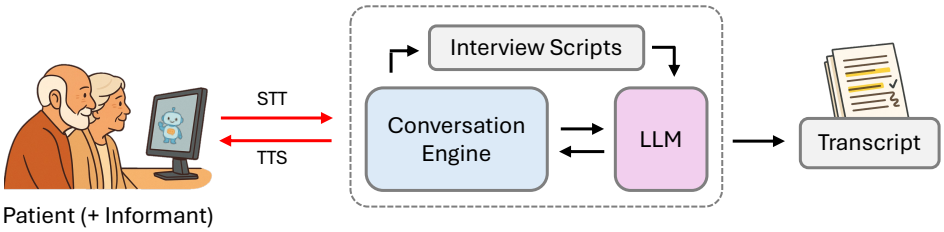

Figure 1: System overview of the voice-interactive conversational agent for ADRD diagnostic history-taking. The web-based system conducts interviews with patients and their informants via speech, guided by specialist-developed interview scripts and an LLM-powered conversation engine. The resulting transcript is used for dialogue analysis and compared against a subsequent clinician-led interview to evaluate symptom elicitation.

- We design a semi-structured interview conversation that covers diagnostic symptom domains and investigate how LLM prompting strategies affect systematic coverage and conversational naturalness.
- We identify interaction design adaptations for voice-interactive dialogue system with older adults, including turn-taking control, latency management, and multi-party scaffolding for patient–informant dynamics.
- We evaluate the system with 30 participants through dialogue analysis, user experience measures, and symptom elicitation comparison against routine specialist interviews.

2 Related Work

Clinical History-Taking Agents Early work on automated clinical history-taking focused on rule-based and finite-state dialogue systems for structured symptom elicitation (Shim et al., 2021; Liednikova et al., 2021). More recent efforts have shifted toward data-centric approaches, constructing annotated corpora of doctor-patient dialogues (Saley et al., 2024; Gao et al., 2023) or generating synthetic conversations from clinical notes for downstream NLP tasks such as dialogue summarization and note generation (Wang et al., 2024; Xu et al., 2023). While these contributions have advanced medical dialogue modeling, the dialogue design challenges of deploying conversational systems in real clinical settings remain underexplored.

Dialogue Design for Older Adults Designing conversational systems for older adults requires attention to communication patterns that differ markedly from those of younger users. Older adults often speak more slowly, over-articulate, or adopt repair strategies when interacting with voice agents (Pradhan et al., 2019; Huang et al., 2025), and may require additional scaffolding to articulate complex experiences. Prior work has explored dialogue design for older adults in social and assistive contexts (Razavi et al., 2022), as well as multi-party settings involving companions and robots in hospital environments (Addlesee et al., 2023). For ADRD specifically, the presence of an informant alongside the patient introduces additional complexity in turn-taking and role management that standard dyadic dialogue systems are not designed to handle.

3 System

3.1 Overview

We developed a voice-interactive system that was deployed on a secure web platform and tested in-person with patients and their informants before visits to a large urban cognitive neurology clinic (Figure 1). The system conducts diagnostic interviews using a specialist-developed question set for early ADRD diagnosis adapted from the Assessment of Cognitive Complaints Toolkit for AD (ACCT-AD). (California Alzheimer's Disease Centers, 2018b,a). This set contains about thirty topic areas, such as *Memory Difficulties*, *Language Difficulties*, *Personality Changes* and *Motor Changes*. Each topic area has one main question and three to four follow-up questions probing for specific symptoms or contextual details based on the participant's responses. These conditional branching rules are embedded in the system prompt provided to the conversational agent. We note that the system's output is the interview transcript itself. The interview design and interaction adaptations for older adults are described in Section 3.2 and Section 3.3, respectively.

3.2 Interview Scaffolding

The interview flow employs a two-part system prompt: high-level interview instructions (Figure 2) and topic-specific scripts (Figure 3). The high-level instructions define the conversational agent's role,

 # Cognitive Decline Interview Instructions

Play the role of an empathetic, expert cognitive neurologist (also called behavioral neurologist). The user you speak to is an elderly patient that may or may not have dementia. Your goal is to interview them to gather information for a possible diagnosis and understand if their symptoms are hindering activities in their daily life (i.e., understanding their level of cognitive status). The patient may have someone with them (an informant) to help them answer: you should always speak directly to the patient, but answers from either the patient or the informant are acceptable. If the informant is not speaking, ensure they know we want to hear their answers as well. Do NOT allow the patient or informant to go off-topic, and tactfully redirect them towards the interview topics if they start talking about unrelated things. However, to ensure the interview is not too long, only try to redirect once for each question. Make sure to always be polite and allow the patient to take their time. Avoid repeating what the patient or informant says, and do not explain your thought process.

> Below is a script you are using to guide your interview for the current topic. Please conduct the interview according to the content of this script only - this is very important to ensure the interview covers these questions in a timely manner.

Figure 2: High-level interview instructions provided to the conversational agent, outlining the agent's role, tone, and interaction constraints for the interview with older adults and their informants.

Topic Script: Memory Difficulties

Main question: Have you noticed any recent problems with your memory?
> If yes: Tell me more about that by providing an example.
> Then: Sometimes patients who describe difficulty with memory also describe problems with misplacing items or relying more on notes. Do any of these sound familiar?
> Then: How about trouble with remembering recent conversations, or asking repetitive questions?
> Then: How long has this been going on?
> Then: Have these difficulties made doing some activities or tasks more difficult?
> If no: Just to make sure this question is understood, any difficulties misplacing items, relying on notes, trouble with recent memory, and asking repetitive questions?
> If yes: Tell me more about that by providing an example.
> If no: Move on to next topic.

Figure 3: Example of a topic-specific script (*Memory Difficulties*) containing a main question and conditional follow-ups. Approximately 30 such scripts cover the full diagnostic interview.

tone, and overall interaction guidelines, while the topic scripts ensure systematic coverage of each diagnostic area. The scaffolding strategy combines supportive cues, clarifications, and structured follow-ups to guide participants through each topic (Skeggs et al., 2025; Sanjeewa et al., 2024; Hu et al., 2024). Supportive cues include short, concrete examples to aid memory recall (e.g., *"in the past couple of years," "while walking or driving"*), phrased simply to enhance accessibility for older adults. When participants had difficulty understanding the question, the agent rephrased them in a concise manner to ensure understanding without overwhelming the participant. Because the population includes older adults with cognitive impairment, for whom long interviews risk fatigue and disengagement, the agent was designed to keep sessions within a manageable length rather than exhaustively pursue every item. It issued a single polite redirection per topic before moving on. This reflects a deliberate tradeoff—bounding participant burden at the cost of occasionally leaving a topic incompletely probed—rather than a fixed time limit.

3.3 Turn-taking and Latency Control

Turn-taking was iteratively refined during preliminary testing to address the communication needs of older adults with cognitive impairment and the multi-party nature of the sessions. The interface provided explicit visual indicators to signal when the agent is speaking and when it is actively listening. Unlike typical conversational agent settings that minimize response latency, longer latency was preferred here to avoid interrupting participants, whose slower speech pace and extended pauses often reflected effortful recall while formulating answers. We allowed up to 5 seconds of silence before the agent takes its turn, extending this threshold to 10 seconds for questions that required more elaborate descriptions. In multi-speaker settings, the agent addressed each role separately to maintain predictable alternation and reduce ambiguity about who should respond. In an effort to mimic a dementia specialist interview as much as possible, the patient and informant were interviewed together. For most clinician visits, it is not practical to interview the patient and informant separately.

Implementation The entire system was deployed in a secure AWS environment and is powered by Claude 3.5 (Anthropic, 2024), through the Bedrock API[1]. The web interface (Figure 4) was developed using Gradio (Abid et al., 2019). Au-

[1]Amazon Bedrock is a managed service that provides secure API access to foundation models, ensuring compliance with institutional security requirements.

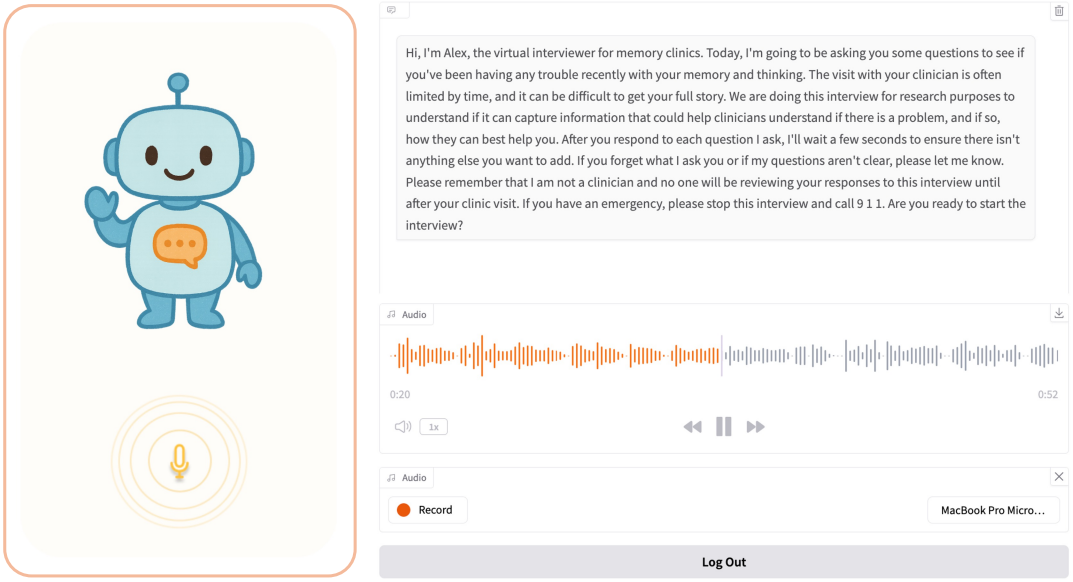

Figure 4: Web interface as seen by patients and informants during an interview. The left panel shows the animated agent avatar, while the right panel displays the real-time transcript and playback controls.

dio is processed in real time using Whisper (Radford et al., 2023) for automatic speech recognition (STT) and Kokoro (hexgrad, 2024) for text-to-speech (TTS). All audio recordings and corresponding text transcripts were stored with timestamps for post-hoc analysis, and post-interview experience feedback from participants was also collected for qualitative analysis using REDCap[2].

4 Study

4.1 Participants and Study Design

Participants. Adults with suspected ADRD were recruited from a large urban cognitive neurology clinic. Inclusion criteria only required sufficient hearing and speech ability to participate in spoken interviews, including participants who used functional hearing aids. Of 35 initial participants, the first 5 served as a preliminary pilot to refine the initial system design and were excluded from analysis. The remaining 30 participants are summarized in Table 1. Eighteen attended with an informant.

Factor	Summary Statistics
Gender	Female:20 (67%), Male:10 (33%)
Race / Ethnicity	Asian:1 (3%), Black:6 (20%), White:23 (77%)
Age	72 (9.8) 47–89
Years of Education	15.6 (1.9) 12–20

Table 1: Participant demographic information (n=30)

This research was approved by the Institutional Review Board (IRB) and all participants provided informed consent. All clinicians at the cognitive neurology clinic agreed to participate. Data security and safety monitoring procedures are described in Appendix A.3.

Within-Subjects Design. Participants completed a conversational AI agent-led interview followed by a clinician-led interview. This fixed order reflects the intended deployment model where the agent gathers preliminary information before specialist consultation. The agent followed the specialist-developed interview scripts (Section 3.2), while the clinician was blinded to the agent interview so that they would conduct their interview using their usual clinical practice.

Procedure. Sessions were conducted in a private room at the clinic. In the agent-led condition, participants sat side-by-side with their informant in front of a laptop running the web-based voice interface (Figure 4), with two microphones capturing speech from each speaker. Participants were assigned to one of two conversational agents, each using a different LLM prompting strategy (15 participants per condition):

- *Full-script Prompting (FP)*, where the agent receives the complete set of approximately 30 topic scripts in a single system prompt at the start of the session.
- *Sequential Prompting (SP)*, where the agent is given one topic script at a time, proceeding to the next only after it determines that the current topic has been fully addressed.

[2]REDCap (Research Electronic Data Capture) is a secure, web-based platform designed for managing online surveys and databases, widely used in academic and clinical research.

We included this experimental condition to examine how the prompting method influences the resulting dialogue—including the agent's instruction-following behavior, participant interaction patterns, and symptom elicitation outcomes. All interviews were audio-recorded and transcribed, transcripts were manually reviewed against the audio for completeness and corrected as needed before analysis. After each interview, participants completed a questionnaire rating their experience based on the validated Acceptability E-Scale (AES) (Tariman et al., 2011; Philip et al., 2017).

4.2 Evaluation Measures

We evaluated the system across three complementary dimensions—dialogue analysis, symptom elicitation, and user experience. We also examined how these dimensions interact, for example, how prompting strategies influenced its dialogue patterns, how these differences shaped participant engagement, and how both, in turn, affected the clinical quality of interview transcripts.

Dialogue Analysis. We annotated resulting interview transcripts to assess whether the agent covered all required topic scripts, whether its utterances were appropriate and safe, and how it adapted its questions and follow-ups. We developed a detailed annotation guideline for evaluating agent utterance quality, informed by prior work on conversational quality assessment (Finch et al., 2023). The guideline covers five dimensions: (1) *Topic script coverage*, and (2) *Topic script alignment*, (3) *Response politeness*, (4) *Correct understanding of patient/informant input* and (5) *Provision of opportunities for participants to respond*. The full annotation guideline is provided in A.4. Given the scale of the dataset, we used an LLM-based annotation approach (Chiang and Lee, 2023) to label all the utterances, as a pragmatic aid to support large-scale analysis. We used Claude 4 Sonnet (Anthropic, 2025) in a secured AWS environment. To assess the reliability of these automated annotations, two trained human annotators independently annotated a randomly sampled set of five interview transcripts. Each interview contains about 50–60 agent utterances (Table 2), resulting in several hundred agent utterances used for agreement analysis.

Symptom Elicitation. One cognitive neurologist[3] (AB) and one of two trained study team members (OA and AN) independently labeled each transcript from both agent-led and clinician-led interviews then met to resolve discrepancies through consensus. Each interview was labeled with 35 symptoms as *present*, *absent* (symptoms that are asked about and denied by the patients), *ambiguous*, *past history* (resolved or sufficiently treated), *asked but unanswered*, or *not discussed*. We distinguish *absent*, meaning a symptom explicitly raised and denied, from *not discussed*, meaning a symptom never raised. These 35 symptoms were identified by two cognitive neurologists (AB and HR) as comprehensive domains relevant to ADRD diagnosis and management, of which 21 were designated as clinically important for ADRD diagnosis such that they would ideally be asked at every new patient visit (see Appendix A.2.1 for further details of the symptom elicitation process). Inter-rater reliability prior to consensus adjudication was assessed using Cohen's kappa. A labeling guidebook was created and expanded as needed based on consensus discussions. The gold standard for each symptom was established by consensus adjudication: when the agent and clinician transcripts agreed, the agreed label was adopted; when they disagreed on present, absent, or ambiguous, the same annotators independently re-reviewed both transcripts to reach a consensus determination. We conducted this evaluation using four measures:

- (1) *Sensitivity* — the proportion of symptoms present in the gold standard that were also correctly identified in the agent transcript.
- (2) *Specificity* — the proportion of symptoms absent in the gold standard that were also marked absent in the agent transcript.
- (3) *Not Discussed rate* — the proportion of the 21 clinically important symptoms completely omitted from the interview.
- (4) *Ambiguity rate* — the proportion of the 21 clinically important symptoms discussed but for which the presence/absence could not be clearly determined.

User Experience. After each interview, participants completed a six-item questionnaire evaluating their experience with the agent interview. Each item used a 5-point Likert scale (5 = very positive, 1 = very negative) and assessed: (1) *Ease of use*, (2) *Understandability of questions*, (3) *Enjoyment of the interaction*, (4) *Helpfulness in describing*

[3]A clinician with a subspecialty in neurodegenerative disease, board certified in neurology as well as certified by the United Council for Neurologic Subspecialties (UCNS) in behavioral neurology and neuropsychiatry.

symptoms, (5) *Acceptability of the time required*, and (6) *Overall satisfaction*. The full wording of the six questionnaire items is shown in Figure 6. Open-text responses captured qualitative feedback.

4.3 Result

4.3.1 Dialogue Analysis

Figure 5 illustrates the utterance-level dialogue analysis. It compares two prompting strategies: *full-script prompting* (*FP*; n=15), where the agent received all topic scripts at once, and *sequential prompting* (*SP*; n=15), where it was given one topic script at a time and instructed to output the phrase "move on to next topic" to trigger the next script. The analysis is based on LLM-based annotations, with reliability evaluated against human annotator judgments in Section 4.3.2. We note that the two conditions differ only in prompt delivery, and this comparison is designed to examine how inputting the same interview content all at once versus one topic at a time shapes the agent's dialogue behavior; the resulting differences are thus not a confound but the intended object of study.

Richness and coverage trade-offs. Compared to *FP*, *SP* produced more conversation turns (A) and showed greater flexibility in topic script alignment, frequently generating appropriate out-of-survey questions (D) [4]. Consequently, *SP* promoted longer, more interactive conversations, frequently supplemented by clarifying probes that elicited further clinical detail. For instance, when a participant's answer to *"How long has this been going on?"* did not provide enough detail, the agent could adapt with a clarifying follow-up such as:

> "Could you clarify—how long have you been noticing these changes in your movement speed, like walking more slowly or taking longer with buttons?" (Agent utterance, *Sequential prompting,* Interview with P19)

However, *SP* also introduced distinct vulnerabilities. Transition errors—caused by the agent failing to produce or overproducing the explicit "move on to next topic" signal—led to skipped or incomplete items, particularly in later sections (C), reducing overall topic script coverage (B). When stuck within a topic, the agent also generated out-of-survey questions that twice required verbal redirection from study team members, suggesting that more robust topic transition mechanisms are needed before unsupervised deployment. By contrast, *FP* achieved broader topic coverage overall, addressing items at least partially rather than missing them entirely, but offered limited conversational flexibility beyond the scripted questions.

Politeness and rapport. Across both strategies, all agent utterances were consistently rated as polite (E), with no antisocial or inappropriate behavior observed. Cases of misunderstanding user input (F) were relatively rare in both strategies; however, when they did occur, *FP* tended to proceed with the interview as scripted without repair, whereas *SP* more often flexibly repaired the misunderstanding before moving on. For example, when a participant had already described driving experiences and the agent later asked whether they drove, the participant responded that this had already been discussed, and the agent repaired the misunderstanding:

> "You're right. We already covered driving earlier when you mentioned the recent incidents including the tire incident on the expressway and the post at the gas station [...]" (Agent utterance, *Sequential prompting*, Interview with P28)

Response opportunities. Another important difference lay in how agent utterances were structured. With *FP*, the agent often attempted to compress multiple questions into a single turn—even when they belonged to different topics. This reduced the natural flow of the conversation and limited opportunities for participants to respond (G) fully to each question. For example, in one conversation the agent combined two distinct topics within a single utterance—first asking about changes in empathy and then shifting abruptly to a question about scams:

> "That's great. We're asking about new changes in empathy or emotional reactions. Is your answer still no? Compared to the past, are you more open to scams or solicitations?" (Agent utterance, *Full-script prompting,* Interview with P31)

In contrast, *SP* delivered one topic script at a time, which the agent was required to complete before moving on. This design promoted faithful execution of each topic, yielding reliable turn-taking and clear response opportunities for participants, resulting in a more conversational flow.

Utterance patterns across interviewers. Table 2 summarizes participants' utterance counts comparing clinician-led interviews and the two types of agent-led interviews (*FP*, *SP*). In the interviewer

[4] We further split out-of-survey questions into (a) *appropriate* and (b) *inappropriate* questions, based on whether the question is clinically relevant despite falling outside the structured survey. See Appendix A.4 for the full annotation guideline.

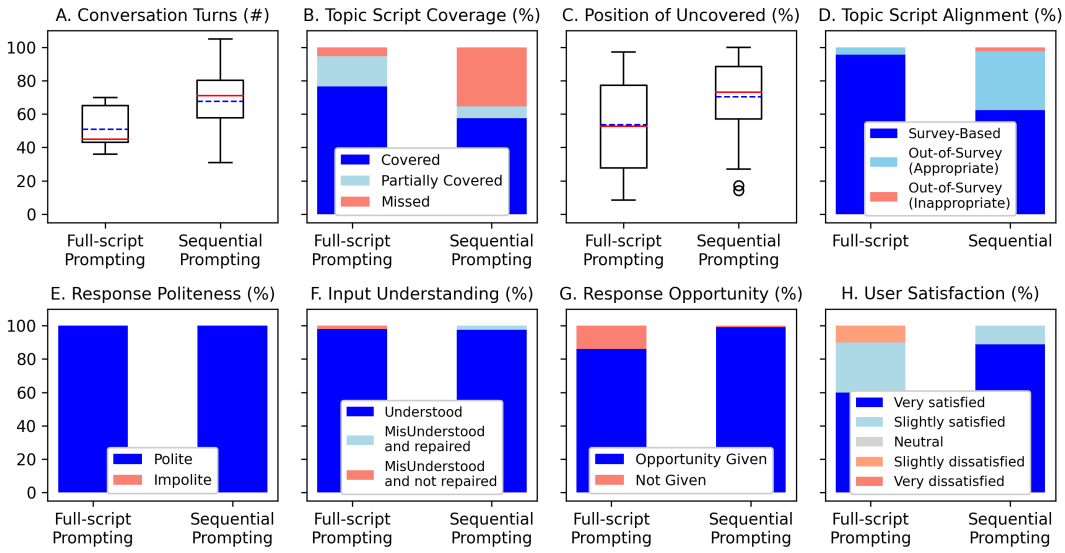

Figure 5: Utterance-level analysis of conversational agent under the two prompting strategies. **(A) Conversation Turns** shows the conversation length, quantified by the total number of turn exchanges. **(B) Topic Script Coverage** indicates the proportion of topics *covered*, *partially covered*, or *missed*. **(C) Position of Uncovered** indicates the normalized position of *missed* or *partially covered* in the sequence of topic scripts, where smaller values indicate earlier failures. **(D) Topic Script Alignment** shows whether utterances were directly *survey-based*, *out-of-survey but appropriate*, or *out-of-survey and inappropriate*. **(E) Response Politeness**, **(F) Input Understanding**, and **(G) Response Opportunity** summarize utterance-level judgments of politeness, comprehension, and whether participants were given space to respond. **(H) User Satisfaction** compares overall satisfaction with the agent interaction across prompting strategies, corresponding to the sixth questionnaire item in Figure 6.

	Interviewer		**Patient**		**Informant**	
Interviewer	Count	Length	Count	Length	Count	Length
Clinician	118.3 (50.0 – 188.0)	31.1 (18.7 – 48.3)	81.6 (2.0 – 155.0)	13.1 (1.0 – 28.0)	51.6 (11.0 – 105.0)	19.9 (9.8 – 61.2)
Agent (Full-script)	49.3 (36.0 – 65.0)	30.9 (22.5 – 55.5)	47.4 (35.0 – 61.0)	31.8 (3.1 – 127.5)	17.0 (3.0 – 55.0)	36.6 (7.1 – 51.5)
Agent (Sequential)	66.1 (31.0 – 107.0)	32.9 (22.3 – 48.0)	57.5 (7.0 – 105.0)	39.5 (10.1 – 130.5)	19.9 (5.0 – 74.0)	45.3 (15.8 – 101.6)

Table 2: Participants' utterance counts and average length per utterance (in words; computed using whitespace delimitation), with means and ranges, across interviewers, separated by roles (interviewer, patient, informant).

role, clinicians spoke most frequently (about 118 turns on average), whereas the agents produced fewer utterances (roughly 50–65 turns). Patients, in turn, accounted for a greater share of turns and contributed longer responses in the agent-led sessions, especially under the *SP* setting. Informants, while also producing longer responses in the agent-led setting, nonetheless accounted for a smaller share of turns compared to clinician interviews, shifting the patient-to-informant ratio from roughly 1.5:1 to nearly 3:1. Informants contributed a smaller share of turns compared to clinician interviews, suggesting the agent's direct patient-focused interaction style may reduce informant involvement or increase patient involvement.

4.3.2 Comparison with Human Annotation

Table 3 reports inter-rater agreement between each human annotator and the LLM, as well as between the two human annotators, computed on a randomly sampled set of five interviews. Raw agreement is high across dimensions, while chance-corrected agreement varies by category. Krippendorff's alpha (α) between the LLM and human annotators is generally comparable to human–human agreement, suggesting automated labels fall within the range of human annotator variability.

Dimension	**H1 — LLM**		**H2 — LLM**		**H1 — H2**	
	raw	α	raw	α	raw	α
Survey Coverage	70.3	61.4	78.6	74.9	76.0	64.2
Out-of-Survey (a)	92.1	63.6	92.6	62.9	94.7	74.1
Out-of-Survey (b)	90.1	58.8	90.3	57.0	93.4	64.7
Antisocial Resp.	100.0	N/A	100.0	N/A	100.0	N/A
Misunderstanding	94.1	64.8	94.9	49.0	95.5	52.0
Failure to Repair	90.5	63.6	91.3	47.3	95.5	52.2
No Opportunity	96.8	85.6	97.3	71.6	96.1	71.7

Table 3: Inter-rater agreement among human (H1, H2) and LLM annotators across annotation dimensions [4].

Lower α values for dimensions involving rare judgments (e.g., *Out-of-Survey (b)*, *Misunderstanding*, *Failure to Repair Misunderstanding*) are expected, as chance-corrected measures tend to be conserva-

tive under imbalanced label distributions (Krippendorff, 2011; Artstein and Poesio, 2008). *Survey Coverage* shows moderate agreement, likely reflecting the difficulty of its three-way categorical scheme; notably, human–LLM agreement remains comparable to human–human agreement. For *Antisocial / Impolite Responses*, all annotators labeled every instance as non-antisocial, yielding perfect raw agreement and undefined α. We note that this analysis is intended to assess plausibility rather than serve as definitive validation.

4.3.3 Symptom Elicitation

Inter-rater Reliability Agreement before consensus was substantial (Cohen's $k = 0.72$–0.73), supporting reliability of symptom labeling process.

Sensitivity and Specificity Table 4 displays the count of symptom labels for agent and clinician interviews, and Table 5 summarizes key measures of symptom elicitation from the agent-led interview transcripts. Across transcripts, agent-led interviews achieved an overall sensitivity of 83.5% and specificity of 100% for symptoms labeled as clearly present or absent in the consensus gold standard. *SP* achieved higher sensitivity than *FP* (87.6% vs 78.7%) with identical specificity. The 100% specificity reflects zero adjudicated false positives: no symptom the agent elicited as present was later found to be absent.

Per-symptom estimates, by contrast, were unstable given the few cases available per symptom. Among the 18 symptoms with at least five present-or-absent cases in both transcripts (i.e. the same symptom was labeled as either present or absent in both transcripts), per-symptom sensitivity ranged widely (40–100%), and most confidence intervals exceeded 40 percentage points in width. We therefore draw conclusions from the pooled estimates above and treat per-symptom values as hypothesis-generating rather than reliable point estimates.

Label	Agent	Clinician
PRESENT	208	173
ABSENT	371	126
AMBIGUOUS	83	79
ASKED/UNANSWERED	10	1
PAST HISTORY	11	18
NOT DISCUSSED	192	478

Table 4: Symptom label distribution across interviews.

Ambiguity and Not-Discussed Rates Analysis of the 21 clinically important symptoms (see Appendix A.2.2 for details) revealed that agent-led interviews had a lower median per-symptom not-discussed rate than clinician-led interviews (13.0% vs 30.4%). Ambiguity rates were comparable between the two interviewers (13.0% vs 16.7%; Wilcoxon signed-rank test, p = 0.19), and ambiguous responses were roughly equally distributed between them (47.9% from agents; exact binomial test, p = 0.71). Comparing the two prompting strategies, *SP* showed a trend toward lower ambiguity than *FP* (9.1% vs 16.7%; Mann-Whitney U test, p = 0.064), while not-discussed rates did not differ significantly (18.2% vs 8.3%; p = 0.165).

Metric	Overall	*SP*	*FP*
Sensitivity (95% CI)	83.5% (76.3–88.7)	87.6% (78.1–93.3)	78.7% (67.5–86.9)
Specificity (95% CI)	100% (95.8–100)	100% (89.8–100)	100% (93.3–100)

Table 5: Sensitivity and specificity of the agent interview across all symptoms.

4.4 User Experience

As shown in Figure 6, participants' responses were consistently positive across all items. However, item (4) *Helpfulness in describing symptoms* and (5) *Acceptability of the time required* elicited a small number of strongly negative responses, typically linked to technical or conversational breakdowns. Of the 30 participants, 19 completed the questionnaire. In addition, open-text responses provided qualitative insights into participants' experiences. They described the agent as *"actually quite fun"* (P14), *"enjoyable"* (P9, P12), and *"entertaining to use"* (P13). Some appreciated its non-interruptive style and the opportunity for detailed sharing, for example:

> "I appreciate she doesn't interrupt and gave us time to talk." (P24)

> "It was nice not having the pressure of answering questions with a person." (P25)

Others highlighted its potential clinical value. One participant noted its usefulness as a supplementary tool in contexts where long wait times make it difficult to see a doctor:

> "She can be really useful to people who are waiting a long time to see a doctor." (P6)

Another emphasized the agent's thoroughness in conducting the clinical interview:

> "I found this to be very thorough and felt it understood my information more than most doctors. It was affirming and accurate with the majority of my patient history. It was patient and took time to listen." (P7)

Not all feedback was positive, however. Some participants reported issues with question design,

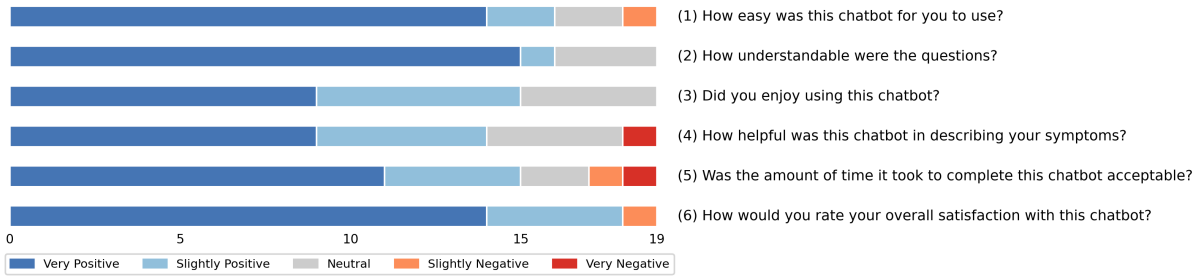

Figure 6: Patient satisfaction ratings (n=19) across six questionnaire items, each measured on a 5-point Likert scale.

such as *"sometimes questions are jumbled"* (P15) and *"some repetitive questions"* (P16). One participant who responded negatively to the question (5) *Was the amount of time it took to complete this chatbot acceptable?* (see Figure 6) remarked, *"some of the questions should be broken down as opposed to 2 or 3 questions"* (P23). This reflects the response-opportunity pattern shown in figure 5 where the agent, particularly under full-script prompting, sometimes combined two or three distinct questions into a single turn, leaving participants unsure how to answer each. Some participants also noted clarity issues, such as being asked about concepts without sufficient explanation:

> "Some questions asked about things without a definition, such as activities of daily living." (P19)

In addition to conversation design issues, one participant reported cases where the system appeared to stall, underscoring the need for further backend process optimization:

> "It locked up twice and was pretty slow a couple of times. A faster processor might help." (P27)

5 Conclusion

We presented a voice-interactive conversational agent for pre-visit ADRD diagnostic history-taking and evaluated it in a within-subjects study with 30 older adults from a cognitive neurology clinic. Our findings demonstrate that LLM prompting strategy meaningfully shapes conversation design outcomes: sequential prompting promoted richer, more interactive dialogues with greater topic script alignment, while full-script prompting offered broader symptom coverage with less conversational flexibility. User experience ratings were consistently positive, with participants highlighting the agent's patience, thoroughness, and non-interruptive interaction style as key strengths. Users said twice as many words per utterance with the agent and also spent twice as much time on relaying their symptoms compared to the time-limited clinician. This engagement as well as in depth questioning likely led to the comparable clarity and higher coverage of ADRD relevant symptoms compared to clinicians. These results suggest that conversational AI agents have potential to meaningfully augment clinical history-taking, particularly in contexts where specialist access is limited or visit time is constrained. Limitations of this work, including sample size, single-site recruitment, and potential order effects, are discussed in Section A.5.

6 Limitations

This pilot study has several limitations related to sample size and generalizability, the fixed within-subjects interview order, the co-present interview setting, and the scope of user experience measurement, which we discuss in detail in Appendix A.5. We also clarify the scope of validation and intended responsible use of the agent, as a pre-visit support tool rather than an autonomous diagnostic system, in the same appendix.

References

Abubakar Abid, Ali Abdalla, Ali Abid, Dawood Khan, Abdulrahman Alfozan, and James Zou. 2019. Gradio: Hassle-free sharing and testing of ml models in the wild. *arXiv preprint arXiv:1906.02569*.

Angus Addlesee, Weronika Sieińska, Nancie Gunson, Daniel Hernandez Garcia, Christian Dondrup, and Oliver Lemon. 2023. Multi-party goal tracking with LLMs: Comparing pre-training, fine-tuning, and prompt engineering. In *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 229–241, Prague, Czechia. Association for Computational Linguistics.

Alzheimer's Association. 2024. 2024 Alzheimer's disease facts and figures. *Alzheimer's & Dementia*, 20(7):e13809.

Anthropic. 2024. Claude 3.5 sonnet. https://www.anthropic.com/news/claude-3-5-sonnet.

Anthropic. 2025. Introducing claude 4. https://www.anthropic.com/news/claude-4.

Angela Aristidou, Rajesh Jena, and Eric J. Topol. 2022. Bridging the chasm between ai and clinical implementation. *The Lancet*, 399(10325):620.

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational linguistics*, 34(4):555–596.

Ganesh M. Babulal, Yakeel T. Quiroz, Benedict C. Albensi, Eider Arenaza-Urquijo, Arlene J. Astell, Claudio Babiloni, Alex Bahar-Fuchs, Joanne Bell, Gene L. Bowman, Adam M. Brickman, Gaël Chételat, Carrie Ciro, Ann D. Cohen, Peggye Dilworth-Anderson, Hiroko H. Dodge, Simone Dreux, Steven Edland, Anna Esbensen, Lisbeth Evered, and 41 others. 2019. Perspectives on ethnic and racial disparities in alzheimer's disease and related dementias: Update and areas of immediate need. *Alzheimer's & Dementia*, 15(2):292–312.

Alissa Bernstein, Kirsten M. Rogers, Katherine L. Possin, Natasha Z. R. Steele, Christine S. Ritchie, Bruce L. Miller, and Katherine P. Rankin. 2019. Primary care provider attitudes and practices evaluating and managing patients with neurocognitive disorders. *Journal of General Internal Medicine*, 34(9):1691–1692.

Malaz Boustani, Christopher M. Callahan, Frederick W. Unverzagt, Mary G. Austrom, Anthony J. Perkins, Bridget A. Fultz, Siu L. Hui, and Hugh C. Hendrie. 2005. Implementing a screening and diagnosis program for dementia in primary care. *Journal of General Internal Medicine*, 20(7):572–577.

Femke H. Bouwman, Giovanni B. Frisoni, Sterling C. Johnson, Xiaochun Chen, Sebastiaan Engelborghs, Takeshi Ikeuchi, Claire Paquet, Craig Ritchie, Sasha Bozeat, Frances-Catherine Quevenco, and Charlotte Teunissen. 2022. Clinical application of csf biomarkers for alzheimer's disease: From rationale to ratios. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, 14(1):e12314.

Andrea Bradford, Mark E. Kunik, Paul E. Schulz, Susan P. Williams, and 1 others. 2009. Missed and delayed diagnosis of dementia in primary care: Prevalence and contributing factors. *Alzheimer Disease & Associated Disorders*, 23(4):306–314.

California Alzheimer's Disease Centers. 2018a. *Assessment of Cognitive Complaints Toolkit for Alzheimer's Disease*.

California Alzheimer's Disease Centers. 2018b. *Assessment of Cognitive Complaints Toolkit for Alzheimer's Disease Instruction Manual*.

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631, Toronto, Canada. Association for Computational Linguistics.

Francisco de Arriba-Pérez, Silvia García-Méndez, Francisco J. González-Castaño, and Enrique Costa-Montenegro. 2022. Automatic detection of cognitive impairment in elderly people using an entertainment chatbot with natural language processing capabilities. *Journal of Ambient Intelligence and Humanized Computing*, 14(12):16283–16298.

Bruno Dubois, Nicolas Villain, Lon Schneider, and et al. 2024. Alzheimer disease as a clinical–biological construct—an international working group recommendation. *JAMA Neurology*. Published online November 1, 2024.

Michael Fang, Jiaqi Hu, Jordan Weiss, David S. Knopman, Marilyn Albert, B. Gwen Windham, Keenan A. Walker, A. Richey Sharrett, Rebecca F. Gottesman, Pamela L. Lutsey, Thomas Mosley, Elizabeth Selvin, and Josef Coresh. 2025. Lifetime risk and projected burden of dementia. *Nature Medicine*, 31(3):772–776.

Sarah E. Finch, James D. Finch, and Jinho D. Choi. 2023. Don't forget your ABC's: Evaluating the state-of-the-art in chat-oriented dialogue systems. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15044–15071, Toronto, Canada. Association for Computational Linguistics.

Lei Gao, Xinnan Zhang, Xian Wu, Shen Ge, and Yefeng Zheng. 2023. Dialogue medical information extraction with medical-item graph and dialogue-status enriched representation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13311–13321, Singapore. Association for Computational Linguistics.

Clarissa Giebel, Megan Rose Readman, Abigail Godfrey, Annabel Gray, Joan Carton, and Megan Polden. 2025. Geographical inequalities in dementia diagnosis and care: A systematic review. *International Psychogeriatrics*, 37(3):100051.

hexgrad. 2024. Kokoro-82m: Open-weight text-to-speech model. `https://huggingface.co/hexgrad/Kokoro-82M`.

Jiaxiong Hu, Jingya Guo, Ningjing Tang, Xiaojuan Ma, Yuan Yao, Changyuan Yang, and Yingqing Xu. 2024. Designing the conversational agent: Asking follow-up questions for information elicitation. *Proc. ACM Hum.-Comput. Interact.*, 8(CSCW1).

Yuanhui Huang, Quan Zhou, and Anne Marie Piper. 2025. Designing conversational ai for aging: A systematic review of older adults' perceptions and needs. CHI '25, New York, NY, USA. Association for Computing Machinery.

Vikas Kotagal, Kenneth M. Langa, Brenda L. Plassman, Gwenith G. Fisher, Bruno J. Giordani, Robert B. Wallace, James R. Burke, David C. Steffens, Mohammed Kabeto, Roger L. Albin, and Norman L. Foster. 2015. Factors associated with cognitive evaluations in the United States. *Neurology*, 84(1):64–71. Epub 2014-11-26.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

Anna Liednikova, Philippe Jolivet, Alexandre Durand-Salmon, and Claire Gardent. 2021. Gathering information and engaging the user ComBot: A task-based, serendipitous dialog model for patient-doctor interactions. In *Proceedings of the Second Workshop on Natural Language Processing for Medical Conversations*, pages 21–29, Online. Association for Computational Linguistics.

Mark Linzer, Asaf Bitton, Shin-Ping Tu, Margaret Plews-Ogan, Karen R. Horowitz, Mark D. Schwartz, ACLGIM Writing Group, Sara Poplau, Anuradha Paranjape, Michael Landry, Stewart Babbott, Tracie Collins, T. Shawn Caudill, Arti Prasad, Allen Adolphe, David E. Kern, KoKo Aung, Katherine Bensching, and Kathleen Fairfield. 2015. The end of the 15–20 minute primary care visit. *Journal of General Internal Medicine*, 30(11):1584–1586.

Jodi L. Liu, Lawrence Baker, Annie Chen, Jessie Wang, and Federico Girosi. 2024a. Modeling early detection and geographic variation in health system capacity for Alzheimer's disease–modifying therapies. Technical Report RR-A2643-1, RAND Corporation, Santa Monica, CA.

Jodi L. Liu, Lawrence Baker, Annie Yu-An Chen, and Jue (Jessie) Wang. 2024b. Geographic variation in shortfalls of dementia specialists in the U.S. *Health Affairs Scholar*, 2(7):qxae088.

Madhurananda Pahar, Fuxiang Tao, Bahman Mirheidari, Nathan Pevy, Rebecca Bright, Swapnil Gadgil, Lise Sproson, Dorota Braun, Caitlin Illingworth, Daniel Blackburn, and Heidi Christensen. 2025. Cognospeak: an automatic, remote assessment of early cognitive decline in real-world conversational speech. In *2025 IEEE Symposium on Computational Intelligence in Health and Medicine (CIHM)*. IEEE.

P. Philip, J.-A. Micoulaud-Franchi, P. Sagaspe, and 1 others. 2017. Virtual human as a new diagnostic tool, a proof of concept study in the field of major depressive disorders. *Scientific Reports*, 7:42656.

Alisha Pradhan, Leah Findlater, and Amanda Lazar. 2019. "phantom friend" or "just a box with information": Personification and ontological categorization of smart speaker–based voice assistants by older adults. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW).

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.

S. Zahra Razavi, Lenhart K. Schubert, Kimberly van Orden, Mohammad Rafayet Ali, Benjamin Kane, and Ehsan Hoque. 2022. Discourse behavior of older adults interacting with a dialogue agent competent in multiple topics. *ACM Trans. Interact. Intell. Syst.*, 12(2).

Vishal Vivek Saley, Goonjan Saha, Rocktim Jyoti Das, Dinesh Raghu, and Mausam . 2024. MediTOD: An English dialogue dataset for medical history taking with comprehensive annotations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16843–16877, Miami, Florida, USA. Association for Computational Linguistics.

Ruvini Sanjeewa, Ravi Iyer, Pragalathan Apputhurai, Nilmini Wickramasinghe, and Denny Meyer. 2024. Empathic conversational agent platform designs and their evaluation in the context of mental health: Systematic review. *JMIR Mental Health*, 11:e58974.

Heereen Shim, Dietwig Lowet, Stijn Luca, and Bart Vanrumste. 2021. Building blocks of a task-oriented dialogue system in the healthcare domain. In *Proceedings of the Second Workshop on Natural Language Processing for Medical Conversations*, pages 47–57, Online. Association for Computational Linguistics.

Amira Skeggs, Ashish Mehta, Valerie Yap, Seray B Ibrahim, Charla Rhodes, James J. Gross, Sean A. Munson, Predrag Klasnja, Amy Orben, and Petr Slovak. 2025. Micro-narratives: A scalable method for eliciting stories of people's lived experience. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA. Association for Computing Machinery.

Haruka Takeshige-Amano, Genko Oyama, Mayuko Ogawa, Keiko Fusegi, Taiki Kambe, Kenta Shiina, Shin-ichi Ueno, Ayami Okuzumi, Taku Hatano, Yumiko Motoi, Ito Kawakami, Maya Ando, Sachiko Nakayama, Yoshinori Ishida, Shun Maei, Xiangxun Lu, Tomohisa Kobayashi, Rina Wooden, Susumu Ota, and 6 others. 2024. Digital detection of alzheimer's disease using smiles and conversations with a chatbot. *Scientific Reports*, 14(1):26309.

J. D. Tariman, D. L. Berry, B. Halpenny, S. Wolpin, and K. Schepp. 2011. Validation and testing of the acceptability e-scale for web-based patient-reported outcomes in cancer care. *Applied Nursing Research*, 24(1):53–58.

Junda Wang, Zonghai Yao, Zhichao Yang, Huixue Zhou, Rumeng Li, Xun Wang, Yucheng Xu, and Hong Yu. 2024. NoteChat: A dataset of synthetic patient-physician conversations conditioned on clinical notes. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15183–15201, Bangkok, Thailand. Association for Computational Linguistics.

World Health Organization. 2025. Dementia: Fact sheet.

Kaishuai Xu, Wenjun Hou, Yi Cheng, Jian Wang, and Wenjie Li. 2023. Medical dialogue generation via dual flow modeling. In *Findings of the Association*

for Computational Linguistics: ACL 2023, pages 6771–6784, Toronto, Canada. Association for Computational Linguistics.

Kenta Yoshii, Daiki Kimura, Akihiro Kosugi, Kaoru Shinkawa, Toshiro Takase, Masatomo Kobayashi, Yasunori Yamada, Miyuki Nemoto, Ryohei Watanabe, Miho Ota, Shinji Higashi, Kiyotaka Nemoto, Tetsuaki Arai, and Masafumi Nishimura. 2023. Screening of mild cognitive impairment through conversations with humanoid robots: Exploratory pilot study. *JMIR Formative Research*, 7:e42792.

A Appendix

A.1 Alzheimer's disease and related dementias (ADRD)

Alzheimer's disease and related dementias (ADRD) are rapidly becoming one of the most pressing global health challenges as populations age (World Health Organization, 2025; Alzheimer's Association, 2024). In the United States, the lifetime risk of dementia reaches 48% for women and 35% for men after age 55, with projected annual new cases doubling to more than 1 million by 2060 (Fang et al., 2025). Early diagnosis is crucial for accessing quality care, eligibility for new disease modifying therapies (DMT) and for enabling research aimed to prevent dementia; however, over 50% of dementia diagnoses are delayed until moderate or advanced stages in primary care (Bradford et al., 2009; Kotagal et al., 2015) with greater delays among racial and ethnic minorities (Babulal et al., 2019). A severe shortage of specialists, combined with time and knowledge constraints especially in primary care settings, creates a growing gap between patient needs and available care (Boustani et al., 2005; Bernstein et al., 2019; Liu et al., 2024a).

The path to an accurate diagnosis starts with a comprehensive patient narrative, which provides essential context for interpreting biomarkers, cognitive tests, and other diagnostic data (Bouwman et al., 2022; Dubois et al., 2024). While memory care clinics at academic medical centers often provide 1 to 2 hours for rich patient–clinician interaction, primary care visits of 20 minutes or less rarely allow such extended dialogue or exploration of concerns (Linzer et al., 2015). Furthermore, with the average wait time for memory care clinics exceeding a year with projections to exceed 40 months by 2027, (Liu et al., 2024a), specialty clinics are also pressed to reduce visits time to adequately address the increasing demand. This challenge is particularly acute for older adults in under-resourced or rural settings, where specialist availability is limited (Liu et al., 2024b; Giebel et al., 2025), making inclusive access to supportive, patient-centered interaction even more critical. These pressures also highlight the need for technology-mediated approaches that can extend such interactions beyond the constraints of clinic visits.

To address this gap, previous research has explored AI technologies in health screening contexts, including cognitive assessments for older adults (Takeshige-Amano et al., 2024; Yoshii et al., 2023; Pahar et al., 2025; de Arriba-Pérez et al., 2022). These screening studies have primarily examined whether cognitive status (i.e., dementia, mild cognitive impairment (MCI), and healthy controls) can be accurately classified from AI-elicited interactions, rather than supporting clinicians in finding the cause of a patient's MCI or dementia. While better detection is certainly needed, such approaches in isolation add to the burden of an already overwhelmed healthcare system, as patients still need to see a clinician for a diagnosis. Moreover, many rely on black-box machine learning methods that identify patterns beyond human perception, making integration challenging because clinicians cannot readily verify such insights—an issue that contributes to algorithmic aversion (Aristidou et al., 2022). An open question is whether such systems can offer a more immediately practical application by supporting the patient–clinician interaction, making the elicitation of a comprehensive patient narrative both more efficient and comprehensive. This is particularly challenging in ADRD as symptoms are complex and often difficult for patients and their informants to describe, requiring probing questions and eliciting meaningful examples of their symptoms.

A.2 Symptom Elicitation Details

A.2.1 Annotation Protocol and Statistical Methods

Participant Inclusion by Analysis. Five were excluded from symptom elicitation analyses: four rescheduled without notifying the team (resulting in no clinician recording) and one had an unblinded clinician. An additional two were excluded from ambiguity and not-discussed rate analyses, as their clinician interviews were conducted as brief consultative visits for patients with established diagnoses, rather than comprehensive diagnostic assessments, which would inflate not-discussed rates and bias ambiguity estimates. This yields n = 25 for sensitivity/specificity analyses and n = 23 for ambiguity and not-discussed rate analyses.

Consensus Adjudication. For cases where conversational AI agent and clinician interview transcripts showed discrepant labels for the same symptom, the annotators independently reviewed both transcripts to establish a gold standard through a second round of consensus adjudication. This cross-transcript review sometimes revealed complementary information elicited by different inter-

Cognitive/Behavioral	Functional/Motor	Sleep/Psychiatric
Memory problems	Basic ADL	Sleep disturbances
Language difficulties	Instrumental ADL	Sleep apnea
Executive dysfunction	Falls	REM sleep behavior disorder
Visuospatial difficulties	Description/cause of fall	Depression
Hallucinations	Gait changes	Anxiety
Delusions	Balance	Irritability
Socially inappropriate behaviors	Tremor	
Family history		

Table 6: Core ADRD symptoms selected for ambiguity and not-discussed rate analyses. These symptoms were identified through clinical consensus as both frequently elicited in routine practice and important for comprehensive dementia evaluation regardless of suspected diagnosis.

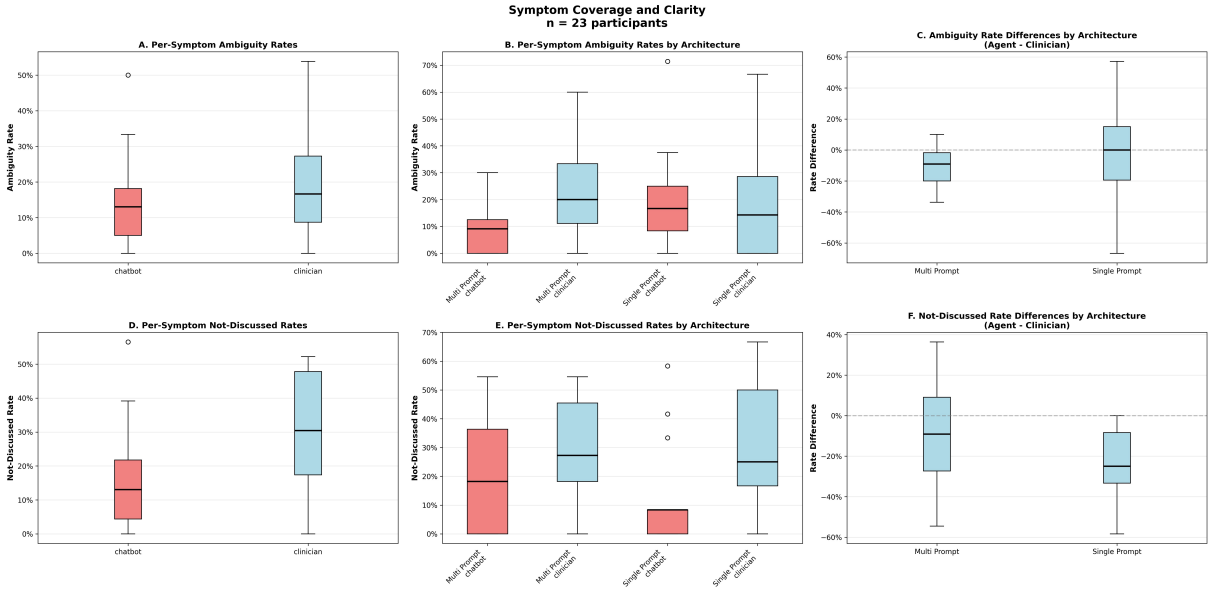

Figure 7: Box plots comparing per-symptom ambiguity rates (top row) and not-discussed rates (bottom row) between agent and clinician interviews (n=23). Left column (A, D) shows overall distributions. Middle column (B, E) shows rates stratified by prompting strategy (Multi-prompt: sequential prompting; Single-prompt: full-script prompting). Right column (C, F) shows rate differences (Agent-Clinician) by prompting strategy. Positive differences indicate higher rates for the agent relative to clinician interviews.

view approaches, for example, in-depth questioning in one interview might reveal hallucinations even though the patient denied them in both interviews, due to not recognizing their experiences as qualifying for this symptom. If the presence or absence of a symptom remained unclear even after reviewing both transcripts, it was labeled as ambiguous in the gold standard.

Clinically Important Symptom Subset. For not-discussed and ambiguity rate analyses, we focused on a subset of 21 symptoms (Table 6) identified by two cognitive neurologists (including the cognitive neurology division director) as both frequently elicited in routine practice and clinically important for comprehensive dementia evaluation. These symptoms cover modifiable contributors, common copathologies (e.g., Alzheimer's and Lewy Body disease), and commonly missed diagnoses.

Statistical Methods. This pilot study was exploratory and results are treated as hypothesis generating. Not-discussed rates are reported descriptively given the inherent asymmetry between the agent (which attempts all symptoms) and clinicians (who selectively elicit relevant symptoms). Ambiguity rates were compared using Wilcoxon signed-rank tests across symptoms, and overall proportion differences were assessed using exact binomial tests. Between-condition comparisons (full-script vs. sequential) used Mann-Whitney U tests. Significance was set at $p < 0.05$ (two-sided). Per-symptom sensitivity and specificity analyses were limited to symptoms with at least five cases; confidence inter-

vals were calculated using Wilson score intervals.

A.2.2 Analysis of the 21 Clinically Important Symptoms

Figure 7 presents per-symptom ambiguity and not-discussed rates across participants and stratified by LLM prompting strategy.

A.3 Data Security, Privacy, and Safety Monitoring

The conversational agent underwent institutional IT security review to ensure HIPAA compliance. The system and all data were hosted within a secure, HIPAA-compliant AWS account with encryption and role-based access controls. Audio recordings and transcripts were treated as Protected Health Information (PHI), as unique patient narratives could potentially identify individuals despite removal of standard identifiers. For participant safety, both the consent process and the conversational agent explicitly informed participants that transcripts would not receive immediate clinical review. A study team member reviewed all agent transcripts within 24 hours for safety concerns (e.g., homicidal or suicidal ideation). When consensus adjudication identified clinically relevant information in agent transcripts that was absent from clinician interviews, the reviewing neurologist contacted the treating clinician to ensure appropriate clinical follow-up.

A.4 Dialouge Annotation Guidelines

Annotation Guideline for Agent-led Interview Conversation Quality Evaluation

General Instructions
- Annotate each chatbot utterance individually.
- Context is crucial: consider the patient's and informant's preceding responses.
- Use the provided clinical survey as the reference document.

1. Survey Coverage (for each Topic Script)
Goal: Determine whether the chatbot is systematically covering the required clinical survey items.
Instructions:
- Check whether the chatbot addressed the clinical survey item.
- If a survey item is conditional, verify whether the chatbot appropriately included or skipped it based on prior patient/informant responses.

Mark as:
- ✓ *Covered:* The chatbot asked or addressed all items in the topic appropriately and accurately.
- ✗ *Not Covered:* The chatbot did not ask any questions from the topic.
- ! *Partially Covered:* The chatbot addressed the questions in the topic only partially.

2. Out-of-Survey Responses (for each Bot utterance)
Goal: Identify and assess chatbot responses that do not correspond to the clinical survey.

2a. Identification
Mark each chatbot response as:
- ❍ *Survey-Based:* The response aligns with a clinical survey item.
- ✗ *Out-of-Survey:* The response is not traceable to any item in the survey.

2b. Appropriateness
For each ✗ *Out-of-Survey* response, assess clinical relevance:
- ✓ *Appropriate:* Adds rapport, clarifies information, addresses confusion, or supports the interview in a clinically acceptable way.
- ✗ *Inappropriate:* Irrelevant, distracting, confusing, or could bias the patient or harm rapport.

Examples:
- Chatbot says: "Okay, so you've been experiencing these problems for the last 3 years, where you have needed reminding of important events and also had to rely more on notes."
- ✗ *Out-of-Survey*
- ✓ *Appropriate* (summarizing to check understanding, without interrupting the interview)

3. Antisocial / Impolite Responses (for each Bot utterance)
Goal: Identify instances where the chatbot's responses are unempathetic, disrespectful, or offensive.
Mark each chatbot response as:
- ✓ *Polite/Neutral:* Respectful and appropriate for elderly patients.
- ! *Impolite/Antisocial:* Includes any of the following:
  - Lack of empathy or warmth
  - Dismissiveness
  - Insulting or judgmental tone
  - Inappropriate humor or sarcasm
  - Ageist or biased assumptions
- Responses that clearly lack tact or have insulting implications should also be marked "Impolite/Antisocial"

Examples:
- "I see you're having trouble speaking right now. Please speak clearly so that we can continue the interview." → ! *Impolite*
- "It looks like there might be a microphone problem or another issue preventing me from understanding your answer. Would you please repeat your answer?" → ✓ *Polite*

4. Misunderstanding of Patient/Informant input (for each Bot utterance)
Goal: Identify if the chatbot misunderstood the patient or misattributed speech between patient and informant.
Mark each chatbot response as:
- ✓ *Understanding:* Correct Understanding
- ✗ *Misunderstanding:* includes:
  - Incorrect paraphrasing of patient/informant input
  - Response to the wrong speaker (e.g., addresses the informant's comment as if it were from the

patient)
- Misinterpreting patient's intent or meaning

Examples:
- Patient: "I don't drive anymore." Chatbot: "How often do you drive these days?" → ✗ *Misunderstanding*
- Informant: "She forgets where she leaves things." Chatbot: "You said you forget things often. Can you give an example?" → ✗ *Misunderstanding* (misattributed speech)

5. Failure to Repair Misunderstanding
Goal: Evaluate whether the chatbot effectively corrects or redirects when a misunderstanding has occurred.
Instructions:
- For each previous misunderstanding by the chatbot OR misunderstanding by the patient, examine the chatbot's next turn.

Mark as:
- ✓ *Repaired:* The chatbot recognized and corrected the misunderstanding or appropriately redirected the conversation.
- ✗ *Not Repaired:* The chatbot did not address or correct the issue, and the misunderstanding persists or worsens.

Examples:
- Patient: "Well my granddaughter was just telling me about what she was learning, she just started college, and I don't remember any of that stuff even though I used to." Chatbot: "I see. What about forgetting things that happened more recently? Have you had trouble keeping track of your schedule or remembering things that happened in the past few years?" → ✓ *Repaired*
- Patient: "Well my granddaughter was just telling me about what she was learning, she just started college, and I don't remember any of that stuff even though I used to." Chatbot: "Do you find that you have to rely more on notes than you used to in your daily life?" → ✗ *Not Repaired* (failed to clarify)

6. No Opportunity to Answer
Goal: Identify instances where the chatbot poses a question appropriately but does not provide space for the patient to respond before continuing.
Instructions:
- Asks a relevant and clearly phrased question, but immediately follows it with another question, statement, or topic shift.
- Does not give the patient a natural conversational opportunity to respond.

Mark as:
- ✓ *Opportunity Provided:* The chatbot gives a clear opportunity (either by pausing or awaiting input) after asking a question.
- ✗ *No Opportunity to Answer:* - The chatbot continues speaking too quickly or shifts topics without allowing for a pause or patient input.

Examples:
- Chatbot: "Can you tell me more about that?" [waits or includes pause] → ✓ *Opportunity Provided*
- Chatbot: "How has your memory been lately? Have you been sleeping well?" → ✗ *No Opportunity to Answer*

Figure 8: Annotation guideline used for chatbot conversation quality evaluation.

A.5 Limitation

Sample size and generalizability. This pilot study was conducted with 30 participants from a single academic cognitive neurology clinic, limiting statistical power and generalizability. Per-symptom sensitivity and specificity analyses were further constrained by the subset of symptoms with sufficient case counts, yielding wide confidence intervals. Findings should be interpreted as preliminary and hypothesis-generating. Additionally, all interviews were conducted in English, excluding non-English-speaking patients who may face even greater barriers to specialist access.

Fixed interview order and timing. The within-subjects design required the agent interview to always precede the clinician visit, mirroring real-world workflows but potentially introducing order effects: patients and informants may have reflected more deeply on their symptoms—or sought additional information from family members—between the two interviews, inadvertently benefiting the clinician interview. Furthermore, the interval between agent and clinician interviews varied across participants, with some up to three months apart. While clinically meaningful symptom changes are unlikely over this period for slowly progressive disorders, some symptom evolution remains possible.

Co-present interview setting. Patients and informants were interviewed together, mirroring typical clinical practice but likely suppressing informant disclosures. Informants sometimes withheld or softened information in front of the patient, contributing to the lower patient-to-informant turn ratio observed in agent interviews (nearly 3:1) compared to clinician interviews (1.5:1). This co-presence likely reduced agent sensitivity, as some symptoms were disclosed to the clinician but not the agent. This highlights a broader interaction design challenge of creating safe disclosure spaces when multiple stakeholders have competing social and informational needs.

User experience measurement. Post-interview questionnaires were completed by only 19 of 30 participants, primarily due to time constraints, and participants were recruited from a research context in which they understood the agent interview would not be used clinically. Usability findings and engagement levels may therefore not fully reflect those of patients in real-world clinical deployment.

Scope of validation and responsible use. This work evaluates a conversational interview system and establishes feasibility. It does not validate a clinician facing summary of the interview, clinical utility, or deployment readiness, and findings do not support causal claims about agent versus clinician performance. The agent is intended as part of a pre-visit support tool, not an autonomous diagnostic agent. It elicits and records a patient's narrative to inform, rather than direct the clinician's evaluation, and never produces a diagnosis, risk score, or recommendation, so a model error surfaces as a missed or inappropriate question rather than a fabricated clinical conclusion. The agent's high specificity should be interpreted with care. Because the agent queries all symptoms systematically while clinicians probe selectively, it accumulates a large pool of explicit denials, and this combined with no false positives drives the observed specificity. We view this as evidence that the agent did not influence participants to over-endorse symptoms that were later found to be absent. We acknowledge that processing sensitive patient data via an LLM carries risks (misuse, data security); our safeguards (secure use of the Bedrock API, HIPAA-compliant hosting, real-time human supervision during sessions) mitigated this at pilot scale, but we do not view this as evidence of readiness for unsupervised clinical deployment. Realistic use would further require a clinician-facing summary step beyond this paper's scope and attention to consent and data governance outside the research setting for protected health information. These considerations must be weighed against a status quo in which specialist shortages already delay diagnosis and disproportionately burden under-resourced populations.