npj | biomedical innovations

Article

https://doi.org/10.1038/s44385-024-00004-8

Empowering Alzheimer's caregivers with conversational AI: a novel approach for enhanced communication and personalized support

Check for updates

Wordh Ul Hasan[1], Kimia Tuz Zaman[1], Xin Wang[1], Juan Li[1] ✉, Bo Xie[2,3] & Cui Tao[4]

Alzheimer's disease and related dementias (ADRD) place a significant burden on caregivers. To address this, we developed ADQueryAid, a conversational AI system designed to empower ADRD caregivers. Built on a Large Language Model and enriched with ADRD knowledge, ADQueryAid uses Retrieval-Augmented Generation to provide personalized and informative support. A user study comparing ADQueryAid to a baseline model (ChatGPT 3.5) demonstrated its superior usability, offering contextually relevant information and emotional support. This study highlights the potential of tailored AI systems to enhance the caregiving experience, paving the way for future research on their long-term impact.

Alzheimer's disease and related dementias (ADRD) present a growing challenge in our aging society. As the disease progresses, it places an increasing burden not only on the patients but also on their caregivers, who are often family members or spouses without professional medical training. In the United States, an estimated 11 million people provide unpaid care to someone with Alzheimer's disease or another dementia. This represents about 18% of the adult population (see https://www.alz.org/). Of these caregivers, about 70% are family members, such as spouses, children, or siblings. The remaining 30% are friends, neighbors, or other unpaid volunteers (see https://www.aarp.org/pri/topics/ltss/family-caregiving/caregiving-in-the-united-states/ (2020)). Family caregivers typically provide an average of 53 h of care per week. This is more than twice the amount of care provided by paid home care aides (see https://www.alz.org/). These caregivers provide essential support and services that would otherwise be unavailable or unaffordable. However, these caregivers also face many challenges, such as lack of training, emotional stress, and financial burden. They face the daunting task of managing complex care needs, including symptom management and treatment decision-making. It is important to provide support to unprofessional caregivers so that they can continue to provide care for their loved ones. This support can come in many forms, such as education and training programs, respite care services, and financial assistance.

AI can readily access and share vast amounts of information about diseases, offering caregivers real-time guidance on symptom management, treatment options, and relevant caregiving strategies[1]. Emerging technologies like conversational AI are presenting exciting possibilities for supporting caregivers, both professional and unprofessional. These intelligent virtual assistants offer a range of helpful functions, alleviating burdens and improving the overall care experience. For example, in our previous research, we have developed an AI-powered voice assistant to help informal caregivers manage the daily diet of patients with ADRD and learn food and nutrition-related knowledge[2]. Alloatti et al.[3] conducted research to evaluate ChatGPT's ability to address myths about AD. They found that most healthcare professionals agreed that ChatGPT has potential in providing meaningful explanations and mitigating AD misinformation, but they also noted the need for more detailed and accurate explanations of the disease's mechanisms and treatments. However, the effectiveness of conversational AI technologies is heavily reliant on the user's ability to formulate precise and relevant queries[4,5]. This presents a significant challenge for unprofessional caregivers who may lack the expertise or the vocabulary to ask the right questions or may not even know what information they need.

Consider a scenario where Alice, a 70-year-old with Alzheimer's, starts exhibiting unusual behavior—sudden bursts of anger, withdrawal from social interactions, and difficulty completing familiar tasks. John, her spouse and primary caregiver, feels a rising sense of alarm. John wants to leverage ChatGPT to gain insights into Alice's behavior, but several factors hinder him:

[1]Department of Computer Science, North Dakota State University, Fargo, ND, USA. [2]School of Nursing, The University of Texas at Austin, Austin, TX, USA. [3]School of Information, The University of Texas at Austin, Austin, TX, USA. [4]Department of Artificial Intelligence and Informatics, Mayo Clinic, Jacksonville, FL, USA. ✉e-mail: j.li@ndsu.edu

Lacking medical expertise, John might struggle to translate his observations into concise questions. He might ask general queries like "Why is Alice acting so strange?" instead of specifics like "What could explain Alice's recent aggression and difficulty with tasks?" This leaves the AI with insufficient context to provide targeted guidance.

1. Unfamiliarity with AI communication: John might not be familiar with how to phrase questions effectively for an AI system. He might not realize the importance of providing context, such as Alice's medical history, medications, or typical behavior patterns. This leads to incomplete information and potentially inaccurate responses from the AI.
2. Unclear goals: John might not have a clear goal in mind. Does he want to understand the cause of Alice's behavior? Find ways to manage it? Seek professional help? This lack of clarity makes it difficult for the AI to suggest relevant information or resources.
3. Misinterpretation of responses: Even if the AI provides helpful information, John might misinterpret it due to his limited medical knowledge. He might need additional clarification or guidance on how to apply the information to Alice's-specific situation.

Consequences of these challenges

1. John might miss crucial information about Alice's condition, delaying proper diagnosis and treatment.
2. He might rely on inaccurate or irrelevant information from the AI, leading to misguided care decisions.
3. John's frustration and anxiety might increase due to the AI's unhelpful responses, exacerbating the stress of caregiving.

This paper introduces an innovative approach to optimizing conversational AI for caregivers of individuals with ADRD, focusing on those without formal medical training. Our strategy aims to enhance the interaction between caregivers and AI systems to ensure a more effective and personalized caregiving process.

We achieve this through several key features. Intuitive user interfaces are designed for non-medical caregivers, leveraging advanced prompt engineering techniques to help them articulate their observations and concerns. Personalized assistance integrates patient information, including medical history and daily routines, into the AI's responses for contextually relevant support. We also provide accessible educational resources that offer straightforward explanations of typical ADRD symptoms and caregiving advice. The AI system is goal-oriented, adapting to user objectives, whether understanding behaviors, managing care, or finding resources. We prioritize clear explanations in all AI-generated content to minimize misunderstandings, especially for non-tech-savvy users. To mitigate AI hallucinations, we ground the AI's responses in a curated knowledge graph populated with peer-reviewed medical literature and official guidelines. This ensures the AI's advice is not only relevant and practical but also backed by the most current and reliable ADRD knowledge.

Our approach is unique in its dual emphasis on improving the AI's understanding of caregiver queries and fostering an educational dialog. By contextualizing each query and guiding caregivers through clarifying questions, our system refines the inquiry process and imparts valuable knowledge. By addressing the existing gaps in AI application for ADRD caregiving, we demonstrate how conversational AI can be adapted to be more user-friendly and effective for individuals lacking medical expertise… Our evaluation demonstrates that tailoring conversational AI systems to the specific needs of ADRD caregivers, as exemplified by ADQueryAid, can lead to improved usability and a more positive user experience compared to general-purpose conversational AI.

Results

The evaluation results demonstrate ADQueryAid's enhanced usability and user experience compared to a general-purpose conversational AI model, particularly in providing personalized and contextually relevant information. The scenario-based use cases and the user study with AD Advocacy and Support Groups highlight the system's potential to address the unique challenges faced by caregivers, especially those with limited medical knowledge and literacy.

Evaluating caregiving challenges using scenario-based use cases

To illustrate ADQueryAid's effectiveness in addressing real-world caregiving challenges, we present the case of Alice, a 72-year-old woman diagnosed with mid-stage Alzheimer's disease, and her husband Bob, her primary caregiver. Alice enjoys gardening, painting, and listening to classical music, particularly Mozart. She has recently been prescribed a new antidepressant medication and has experienced some dental issues, leading to a preference for softer foods. Bob, who has low literacy, often struggles to articulate his concerns precisely.

Case 1: Maintaining engagement: Bob expresses to ADQueryAid that Alice has become increasingly bored and difficult to engage. The system, accessing Alice's profile in the knowledge graph, suggests activities like listening to Mozart, painting landscapes, and light gardening tasks. It also recommends social interaction with her close friends, Cindy and Mary, known to be supportive and understanding of Alice's condition.

Case 2: Addressing decreased appetite: Bob raises concerns about Alice's reduced appetite. ADQueryAid, integrating information about Alice's dietary needs and medical conditions, suggests offering smaller, more frequent meals with softer textures, incorporating her favorite foods like mashed potatoes and applesauce. The system also advises engaging Alice in enjoyable activities, such as tending to her beloved rose bushes or listening to classical music, to potentially stimulate her appetite. Further, ADQueryAid reminds Bob to inform Alice's primary care physician, Dr. Emily Davis, about her eating habits for potential medication adjustments.

Case 3: Managing caregiver stress: Bob expresses frustration with the challenges of caregiving. ADQueryAid offers empathetic support, acknowledging the emotional toll of caregiving. It reminds Bob of Alice's love for gardening and suggests spending time together tending to their garden. It also encourages him to seek support from friends like Cindy and Mary or to consider joining a local support group for caregivers of individuals with Alzheimer's.

Case 4: Articulating concerns: Bob messages ADQueryAid, saying, "Alice is not herself lately. She's doing odd things and seems confused." ADQueryAid, recognizing Bob's difficulty in expressing his concerns, initiates a guided conversation. It asks simple questions to clarify his observations: "Can you tell me more about what Alice is doing that seems odd? " As Bob responds, providing examples of Alice's behavior in plain language, ADQueryAid gradually helps him articulate his concerns more clearly. It maps his descriptions to relevant terms in the knowledge graph, such as "memory loss," "confusion," and "changes in behavior." This allows the system to identify potential issues like medication side effects or worsening dementia and provide tailored advice, using simple language that Bob can easily understand.

These cases demonstrate how ADQueryAid leverages personalized information, knowledge retrieval, reasoning capabilities, and empathetic communication to support caregivers like Bob. By addressing specific challenges, including difficulties in articulating concerns, ADQueryAid showcases its potential to enhance the caregiving experience and improve the quality of life for both caregivers and individuals with Alzheimer's. We have included the detailed interaction transcripts for each of the four cases described as Supplementary Materials.

Performance validation through user study

The user study with 20 participants from AD Advocacy and Support Groups further validated ADQueryAid's superior performance compared to the base GPT-3.5 model.

Table 1 provides a snapshot of the demographic characteristics of the 20 participants involved in our study. The participants were primarily recruited through online channels, which may explain the high rate of tech

Table 1 | Demographic characteristics of participants ($n = 20$)

Characteristic	Frequency (n)	Percentage (%)
Age (years)		
<20	0	0%
20–29	1	5%
30–39	7	35%
40–49	3	15%
50–59	2	10%
60+	7	35%
Gender		
Male	8	40%
Female	12	60%
Caregiver duration (years)		
<1	7	35%
1–3	3	15%
>3	10	50%
Caregiver hours (per week)		
0–9	11	55%
10–19	2	10%
20–29	3	15%
30–39	2	10%
40+	2	10%
Tech proficiency		
Yes	16	80%
No	4	20%
ADRD training		
Yes	3	15%
No	17	85%

Table 2 | CUQ metrics comparison between ChatGPT 3.5 and ADQueryAid

Metric	ChatGPT 3.5	ADQueryAid
CUQ Score	72.5 ± 19.2	83.8 ± 17.2
Lowest Score	43.8	42.2
Highest Score	96.9	100
Median Score	75.0	90.6

proficiency (80%). They were mostly female (60%) and within the age range of 30–59 (50%), with a significant portion aged 60 and above (35%). Most participants had been caregivers for more than three years (50%) and worked between 0 and 9 h per week in that role (55%). However, only a small percentage (15%) had received formal training in ADRD.

This demographic information is crucial as it helps us understand the specific context and needs of the individuals for whom our mobile app is designed. The data suggest a target user base of primarily middle-aged to older adults, who are likely to have experience as caregivers and a high degree of comfort with technology. The low percentage of participants with ADRD training highlights the potential value of our app in providing accessible and personalized dietary guidance to caregivers who may not have extensive medical knowledge in this area.

The Chatbot Usability Questionnaire (CUQ) is a tool specifically designed to measure the usability and user experience of chatbot systems. Scores are calculated on a scale of 0 to 100, with higher scores indicating better perceived usability. Table 2 presents the CUQ metrics for both systems, with ADQueryAid achieving a significantly higher mean CUQ score of 83.8 compared to 73.3 for GPT-3.5. This indicates that participants generally perceived ADQueryAid as more user-friendly and effective. The higher median score (90.6 vs. 75.0) and slightly lower standard deviation further suggest that ADQueryAid provided a more consistently positive user experience.

A detailed analysis of the 16 CUQ questions revealed ADQueryAid's superior usability compared to the ChatGPT 3.5 model (Table 3). This superiority was evident in several key aspects:

1. Personality (Q1): Participants rated ADQueryAid (4.4 ± 1.0) significantly higher than ChatGPT 3.5 (3.5 ± 1.2) in terms of having a realistic and engaging personality, a crucial factor for fostering user engagement in caregiver support systems.
2. Robotic perception (Q2): ADQueryAid (1.7 ± 1.0) was perceived as significantly less robotic than ChatGPT 3.5 (2.4 ± 1.4), indicating a more natural and responsive interaction.
3. Input recognition (Q10): ADQueryAid outperformed ChatGPT 3.5 in recognizing user inputs, suggesting a more accurate understanding of caregiver queries.
4. Informativeness (Q11): ADQueryAid (4.5 ± 0.6) also scored higher in providing useful, appropriate, and informative responses compared to ChatGPT 3.5 (3.7 ± 0.8). This is essential for empowering caregivers with relevant information to make informed decisions.

We conducted a paired t test to statistically analyze the differences between ADQueryAid and the base GPT-3.5 model across the 16 CUQ questions. Prior to conducting the paired t test, we visually assessed the normality of the distribution of the differences between paired scores using a histogram. The results, presented in Table 4, reveal several key findings:

ADQueryAid's advantages:

1. Personality: ADQueryAid was perceived as significantly more realistic and engaging (Q1).
2. Purpose: ADQueryAid better communicated its scope and purpose (Q5).
3. Response Quality**:** ADQueryAid provided significantly more useful and relevant responses while generating fewer irrelevant responses (Q11, Q12).

These findings highlight ADQueryAid's strengths in key areas crucial for user engagement and satisfaction:

1. Friendliness: Both systems were perceived as equally welcoming and friendly.
2. Navigation: Both systems were considered easy to navigate.
3. Error handling: Both systems performed similarly in handling errors and mistakes.

These results suggest that while ADQueryAid excels in certain aspects, both systems offer comparable performance in other areas of usability.

Overall, the paired t test analysis underscores the importance of tailoring conversational AI systems for specific user needs. ADQueryAid's superior performance in key areas like personality, purpose clarification, and response quality indicates the potential of specialized AI systems in providing more effective support to caregivers. These findings also offer valuable insights for further development, highlighting areas where both systems can be improved to better meet the needs of ADRD caregivers.

Discussion

The evaluation of ADQueryAid, encompassing both scenario-based use cases and a user study with AD Advocacy and Support Groups, yielded compelling evidence of its potential to significantly enhance the caregiving experience for individuals with ADRD. This comprehensive evaluation approach allowed us to assess ADQueryAid's performance from multiple angles, providing a holistic understanding of its strengths and areas for further refinement.

The scenario-based use cases served as a practical demonstration of ADQueryAid's capabilities in addressing diverse real-world caregiving challenges. Through these simulated scenarios, we observed the system's

Table 3 | Usability comparison (mean ± SD) between ADQueryAid and ChatGPT 3.5 based on CUQ questions

Question	ChatGPT 3.5 (mean ± SD)	ADQueryAid (mean ± SD)
Q1: The chatbot's personality was realistic and engaging	3.5 ± 1.2	4.4 ± 1.0
Q2: The chatbot seemed too robotic	2.4 ± 1.4	1.7 ± 1.0
Q3: The chatbot was welcoming during initial setup	4.3 ± 1.1	4.1 ± 1.1
Q4: The chatbot seemed very unfriendly	1.9 ± 1.1	1.7 ± 0.8
Q5: The chatbot explained its scope and purpose well	3.7 ± 1.0	4.3 ± 0.6
Q6: The chatbot gave no indication as to its purpose	2.2 ± 1.2	1.7 ± 0.9
Q7: The chatbot was easy to navigate	4.2 ± 1.0	4.5 ± 0.9
Q8: It would be easy to get confused when using the chatbot	2.1 ± 1.1	1.8 ± 1.0
Q9: The chatbot understood me well	3.8 ± 1.1	4.5 ± 0.8
Q10: The chatbot failed to recognize a lot of my inputs	2.0 ± 1.0	1.5 ± 0.8
Q11: Chatbot responses were useful, appropriate and informative	3.7 ± 0.8	4.5 ± 0.6
Q12: Chatbot responses were irrelevant	2.4 ± 0.9	1.7 ± 1.0
Q13: The chatbot coped well with any errors or mistakes	3.8 ± 1.2	4.2 ± 1.0
Q14: The chatbot seemed unable to handle any errors	2.2 ± 1.0	1.7 ± 0.9
Q15: The chatbot was very easy to use	4.4 ± 0.6	4.6 ± 0.7
Q16: The chatbot was very complex	1.9 ± 1.1	1.6 ± 1.0

Table 4 | Paired *t* test results for ADQueryAid vs. GPT-3.5

Question	T-statistic	*P* value	ADQueryAid
Q1: The chatbot's personality was realistic and engaging	2.391	$P = 0.02$	Statistically significant difference, favoring ADQueryAid.
Q5: The chatbot explained its scope and purpose well	2.238	$P = 0.037$	Statistically significant difference, favoring ADQueryAid.
Q11: Chatbot responses were useful, appropriate, and informative	3.106	$P = 0.005$	Statistically significant difference, favoring ADQueryAid.
Q12: Chatbot responses were irrelevant	−2.595	$P = 0.017$	Statistically significant difference, favoring ADQueryAid.
Other Questions (Q2, Q3, Q4, Q6, Q7, Q8, Q9, Q10, Q13, Q14, Q15, Q16)	–	–	No statistically significant difference between ADQueryAid and GPT-3.5.

ability to provide tailored and contextually relevant support to caregivers. In the case of Alice and Bob, ADQueryAid effectively addressed issues such as maintaining engagement, managing dietary changes, and mitigating caregiver stress. The system leveraged personalized patient information from the knowledge graph, combined with its knowledge base and reasoning capabilities, to offer specific and actionable recommendations. In addition, ADQueryAid's ability to guide caregivers through clarifying questions proved invaluable in scenarios where concerns were difficult to articulate, ensuring a thorough understanding of the caregiver's needs and providing appropriate guidance.

The user study with 20 participants further validated ADQueryAid's effectiveness and superior performance compared to the base GPT-3.5 model. The blind, randomized design ensured unbiased feedback, allowing for a direct comparison of user experiences. Quantitative analysis using the CUQ revealed that ADQueryAid achieved a significantly higher mean score than GPT-3.5, indicating a more positive user experience overall. Notably, ADQueryAid excelled in areas such as having a more engaging and realistic personality, being less robotic, accurately recognizing user inputs, and providing more useful and informative responses.

Qualitative feedback from participants aligned with the quantitative findings. Caregivers expressed appreciation for ADQueryAid's personalized approach, empathetic tone, and ability to offer relevant information tailored to their specific needs. They highlighted the system's potential to alleviate the burden of caregiving by providing accessible information, practical suggestions, and emotional support.

The paired *t* test analysis confirmed the statistical significance of ADQueryAid's superior performance in several key usability aspects, including personality, purpose clarification, and response quality. These findings underscore the importance of tailoring conversational AI systems to the specific needs of caregivers and demonstrate the potential of ADQueryAid in providing effective and personalized support for individuals navigating the complexities of ADRD care.

While our evaluation methodology provided valuable insights into ADQueryAid's effectiveness and usability, it is important to acknowledge its limitations.

The CUQ, while a valuable tool for assessing user experience, is primarily designed for evaluating basic chatbots and may not fully capture the nuances of more advanced conversational AI systems like ADQueryAid and ChatGPT. For instance, the CUQ does not specifically address aspects such as the quality of personalized recommendations, the accuracy of medical information, or the ability to handle complex and nuanced queries. These are crucial aspects of ADQueryAid's functionality, and a more specialized evaluation framework would be needed to fully assess its capabilities.

Furthermore, the user study involved a relatively small sample size of 20 participants, which may limit the generalizability of the findings. While the participants were recruited from relevant online communities, a larger and more diverse sample would be needed to draw more robust conclusions about ADQueryAid's effectiveness across a wider range of caregiver populations.

The reliance on self-reported survey data also presents a limitation. While the CUQ provides valuable insights into user perceptions, it may not fully capture the actual user behavior and interaction patterns with the system. In-depth interviews or observational studies could provide a more nuanced understanding of how caregivers interact with ADQueryAid and how it impacts their caregiving experience.

While acknowledging the limitations of the current evaluation, this study offers promising initial insights into the potential of ADQueryAid to enhance the usability and user experience for caregivers of individuals with

ADRD, compared to a general-purpose conversational AI model. Future research will focus on developing more comprehensive evaluation frameworks, expanding the sample size and diversity, and incorporating qualitative research methods to gain a deeper understanding of the impact of ADQueryAid on the caregiving journey.

The findings of this study align with and extend previous research on the application of conversational AI systems in healthcare and caregiving. The existing literature on the use of ChatGPT in medical education and consultation demonstrates its potential to provide logical justification and informational context for answers, which are beneficial for medical education and clinical decision-making. For instance, ChatGPT has been evaluated for its capability to answer questions from the United States Medical Licensing Examination (USMLE), performing at or near the passing threshold for all three exams without specialized training[6,7]. This suggests that AI models can be valuable tools for learning and knowledge assessment in medical education. While ChatGPT has shown promise in providing medical information and patient education, its responses require careful oversight to ensure accuracy[8,9]. This aligns with our findings that ADQueryAid, a customized AI system tailored for ADRD caregivers, outperformed the general-purpose ChatGPT 3.5 in providing contextually relevant and supportive responses. This emphasizes the value of specialized AI systems in delivering precise information tailored to specific user needs.

Research has also explored the potential of ChatGPT as a virtual assistant in healthcare, capable of formulating interpretable responses to clinical questions[10]. Similar to these findings, our study demonstrated ADQueryAid's ability to provide emotionally supportive and contextually relevant advice, significantly enhancing the caregiving experience. This suggests that the benefits observed with ChatGPT in clinical settings can be extended to caregiving contexts through tailored AI systems like ADQueryAid. The importance of prompt engineering in optimizing AI performance has been highlighted in previous research. Studies by Zuccon and Koopman[11] and White et al.[12] have shown that the knowledge embedded in prompts can critically impact the accuracy of ChatGPT's responses in healthcare. Our work extends this understanding by demonstrating ADQueryAid's superior performance, attributed in part to its advanced prompt engineering techniques. This underscores the necessity of prompt engineering in enhancing the performance of AI systems, particularly in specialized applications like caregiving.

Furthermore, research by Perry et al.[13], Dang et al.[14], and Qin and Eisner[15] has emphasized the need for structured approaches in prompt engineering to elicit accurate and relevant information from language models. Our findings align with this research, as ADQueryAid's effectiveness is partially due to its use of structured and well-designed prompts that facilitate professional-level responses, even when used by non-professionals.

In conclusion, this study builds upon prior research by demonstrating the superior support that ADQueryAid provides to ADRD caregivers compared to ChatGPT 3.5. It highlights the importance of specialization, prompt engineering, and tailored responses in enhancing AI usability and effectiveness in caregiving. Future research should further refine these models and integrate them into broader caregiving support systems to maximize their impact.

Methods

Our methodology centers on leveraging prompt engineering and knowledge graphs to develop a customized Conversational AI system—ADQueryAid. To integrate both user-specific details and specialized knowledge on ADRD, we employ plug-and-play (PnP) modules, enhancing ChatGPT's ability to comprehend and respond to users' unique situations. These PnP modules, structured as knowledge graphs and Internal file system, grant ChatGPT access to a wealth of context-specific and medical information, thereby refining its responses to cater to individual user needs and medical inquiries. Furthermore, we design interfaces that facilitate ChatGPT's interaction with external services and databases, expanding its capabilities to include real-time data retrieval, transaction processing, and access to specialized systems. Figure 1 shows the system architecture. This figure illustrates the architecture of a voice-activated AI assistant system, beginning with user voice input converted to text via a speech-to-text (STT) engine. The text is then preprocessed and analyzed by the natural language understanding (NLU) component to identify intent and extract entities. A knowledge graph is queried to provide relevant context, while the context management system tracks conversation history. Advanced AI reasoning algorithms, powered by OpenAI's GPT-4, generate contextually accurate responses, which are then converted back to speech through a text-to-speech (TTS) engine, completing the interaction. In the following sections, we explain the details of each of the major system components.

System design and development

To materialize this vision, we undertook the development of ADQueryAid, employing a multi-faceted approach that encompasses knowledge foundation construction, user intent identification, response generation, and prototype implementation.

Knowledge foundation construction

ADQueryAid's effectiveness is rooted in its comprehensive knowledge foundation, which integrates structured and unstructured data to provide nuanced support for ADRD caregiving.

At the heart of ADQueryAid lies a meticulously structured knowledge graph, underpinned by a well-defined ontology. Figure 2 depicts a top-level knowledge graph structured as an ontology, categorizing and linking

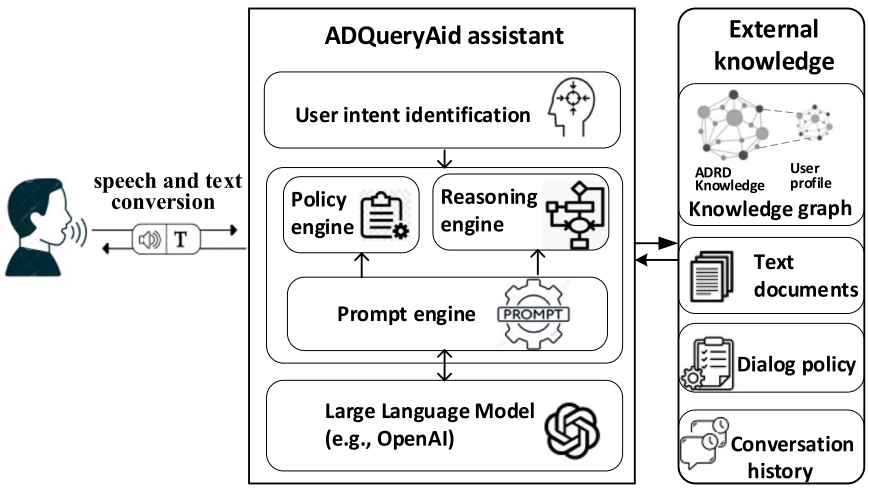

Fig. 1 | ADQueryAid system architecture for personalized ADRD care.

Fig. 2 | Ontology-driven knowledge graph for enhancing AI-assisted ADRD care.

various healthcare concepts. The ontology includes broad categories such as Concerns, Diagnostics, Diseases, Drugs, Events, Patient Care, Person, Profiles, Risk Factors, Stages of Dementia, Symptoms, Treatments, and Types of Dementia. Each category contains specific sub-concepts, like Mental Health under Concerns and Imaging Tests under Diagnostics. The Profiles section details patient and health-related profiles, while the owl: topObjectProperty section outlines object properties defining relationships between concepts, such as hasMedicalHistory and hasSymptoms. This structured approach facilitates efficient data retrieval and reasoning in AI applications for personalized patient care and management. This ontology serves as the blueprint, defining the fundamental concepts and relationships that comprise the knowledge base. It outlines the various entities relevant to ADRD care, including patients, caregivers, symptoms, diagnoses, treatments, and best practices. Furthermore, it establishes the relationships between these entities, such as the association between patients and their caregivers, the progression of Alzheimer's stages, and the link between symptoms and their potential treatments.

The ontology was constructed using a systematic and iterative process involving domain definition, knowledge acquisition, conceptualization, and implementation, as detailed in our previous work[16]. This process included close collaboration with domain experts in ADRD, a thorough review of relevant literature and guidelines, iterative refinement based on expert feedback, and the utilization of ontology engineering tools. The complete ontology is publicly available via GitHub (https://github.com/Intelligent-System-Lab/ADQueryAid).

This ontological framework is then instantiated with specific data, creating a rich and interconnected knowledge graph. Patient profiles encompass sociodemographic details, health information like Alzheimer's stage, and relevant capabilities. Caregiver profiles include educational backgrounds, schedules, and preferences, facilitating personalized support. The graph incorporates comprehensive ADRD-related data, encompassing diagnosis processes, treatment options, symptom descriptions, and best practices.

Complementing the knowledge graph is a vast repository of unstructured text documents from authoritative sources like the Alzheimer's Association. This repository contains care guides, practical resources, educational materials, and in-depth information on diagnosis and treatment, further enhancing the system's ability to provide comprehensive guidance.

ADQueryAid enhances the caregiving experience through two advanced knowledge retrieval mechanisms. The first mechanism, graph data retrieval, leverages the inherent structure of the knowledge graph for efficient and precise retrieval of information. Utilizing graph query languages like SPARQL or Cypher, ADQueryAid can quickly traverse the relationships between nodes and edges in the graph to pinpoint relevant information. This enables rapid response times for direct queries, a critical factor in real-time conversational AI. The interconnected nature of the knowledge graph is a key asset in this process. By exploiting these relationships, ADQueryAid can infer implicit connections between concepts, leading to more comprehensive and insightful responses. For example, a query about a specific Alzheimer's symptom could not only retrieve direct information about that symptom but also related information about potential causes, treatments, and management strategies, all linked within the graph.

The second mechanism, text data retrieval, employs a Retrieval-Augmented Generation (RAG) workflow to access unstructured text documents. This broadens the system's informational reach beyond the

knowledge graph, enriching responses with contextually relevant and authoritative information from a vast repository. Powered by vector search technology, this approach understands and responds to complex queries, even when answers are not readily available in structured data. By combining these retrieval mechanisms, ADQueryAid delivers a dynamic support system balancing efficiency and depth. Each interaction is personalized and contextually informed, drawing from both structured and unstructured knowledge sources. This ensures proactive, anticipatory guidance rooted in a thorough understanding of ADRD care.

User intent identification

ADQueryAid prioritizes accurate identification of caregiver intent within their queries, a crucial step towards personalized guidance. To achieve this, the system leverages both the structured ADRD knowledge graph and the classification abilities of a large language model (LLM).

User input is enriched through contextualization against the knowledge graph, allowing the reasoning engine to collaborate with the prompt engine to refine the query and generate a more targeted prompt for the LLM. This process involves a multi-faceted analysis of the user's input:

1. Knowledge graph alignment: Upon receiving a user query, ADQueryAid first attempts to map key concepts in the query to nodes and relationships within the structured ADRD knowledge graph. In our current system, we leverage the existing capabilities of the LLM for this mapping, categorizing the intent into predefined areas such as diagnosis, treatment, symptoms, etc.
2. Targeted clarification: If the initial query lacks specificity or clarity, ADQueryAid engages the user in further refining the query. Guided by the identified intent category and the associated knowledge graph context, it strategically elicits additional details or presents clarifying options tailored to the domain knowledge.
3. Prompt engineering for clarifying questions: We employ prompt engineering techniques to guide the LLM in formulating these clarifying questions. The prompts are carefully crafted to be:

a. Context-aware: Incorporating relevant patient and caregiver information from the knowledge graph.
b. Intent-specific: Tailored to the identified intent category (e.g., questions about symptoms would focus on symptom details, duration, etc.).
c. User-friendly: Framed in simple language accessible to caregivers with limited medical knowledge.
d. Neutral and open-ended: Prioritize open-ended questions that avoid leading language or assumptions. Encourage users to describe their observations and concerns in their own words.

The decision to ask clarifying questions is primarily driven by several factors, including low intent classification confidence, incomplete context, heuristic rules and patterns, and conversation history analysis. When the initial intent classification yields a low confidence score, it suggests ambiguity in the query, prompting ADQueryAid to seek further clarification. Similarly, if the knowledge graph lacks sufficient information about the patient or caregiver, ADQueryAid prompts for additional details to personalize its response. Additionally, predefined heuristic rules help identify common patterns of generic or unspecific queries that often require clarification. Furthermore, by analyzing the conversation history, ADQueryAid can detect if the user repeatedly asks broad or vague questions without providing specific details, which might indicate a difficulty in articulating their concerns or a lack of understanding of the system's capabilities.

For example, if a user asks, "My dad is having trouble lately," ADQueryAid might analyze the query as follows:

1. Few keywords: "trouble," "lately"
2. Ambiguous entity: "trouble"
3. Low intent classification confidence

Based on this analysis, ADQueryAid would recognize the need for clarification and generate targeted follow-up questions.

We utilize a few-shot learning approach with LLM to dynamically generate clarifying questions. By providing the LLM with carefully curated examples of neutral, open-ended prompts, we guide it to produce similar outputs without relying on rigid, pre-programmed rules. This approach allows the system to adapt to various user inputs flexibly and reduce the risk of leading the user.

For example, in the following Python code, we provide the LLM with a few examples:

```
# Policy: Empathy (using sentiment
analysis)
sentiment = analyze_sentiment(user_query)
    if sentiment == "negative":
        prompt += "Acknowledge the
caregiver's feelings and provide
reassurance."
# Policy: Language Tailoring
if context.get("caregiver_education_level")
== "low":
        prompt += "Use simple, easy-to-
understand language. Avoid medical jargon."
```

By combining these techniques and employing thoughtful prompt engineering, ADQueryAid strives to accurately identify user intent and facilitate effective communication, even when faced with initially vague or ambiguous queries.

Response generation

The response generation hinges on sophisticated prompt engineering, where the system strategically crafts prompt that leverage its diverse knowledge sources to elicit the most relevant and informative responses from the LLM. This process seamlessly integrates knowledge graph data retrieval, RAG, conversation history, policy considerations, and reasoning logic.

Extracted entities and relationships from the knowledge graph are incorporated directly into the prompt. This provides the LLM with crucial context about the patient, caregiver, and ADRD-related concepts. For instance, if the query involves a specific symptom, the prompt would include details about the patient's diagnosis, stage of Alzheimer's, and relevant medical history. Information retrieved from the unstructured text repository via RAG is strategically woven into the prompt. This ensures that the LLM has access to the most up-to-date and comprehensive information on ADRD care, even if it's not explicitly captured in the structured knowledge graph.

Key points from previous interactions are summarized and included in the prompt. This allows the LLM to maintain context and generate responses that are coherent and relevant to the ongoing conversation. For instance, if the caregiver has previously expressed concerns about a particular behavior, the prompt would remind the LLM of this concern, ensuring that the response addresses it appropriately. The prompt explicitly instructs the LLM to adhere to the predefined policies governing the interaction. This ensures that the generated response is not only informative but also empathetic, safe, and culturally sensitive. For example, the prompt might direct the LLM to use simple language if the caregiver has limited health literacy or to prioritize safety concerns if the patient is at risk.

This interplay of components allows ADQueryAid to dynamically refine caregiver requests. Each interaction is informed, contextually relevant, and increasingly attuned to the caregiver's needs, ultimately improving the caregiving experience and patient outcomes.

The response generation process follows a structured path, represented by the conversational state tuple $(q, c, k, p, h) \rightarrow r$, where:

q: The caregiver's immediate query.

c: Background information, including patient-specific details from the knowledge graph.

k: Relevant ADRD knowledge from the knowledge graph and external documents.

p: A set of predefined rules guiding the system's interactions, ensuring ethical standards and caregiving best practices.

h: Conversation history for seamless continuity.

r: The generated response, synthesizing the gathered information.

The prompt explicitly instructs the LLM to adhere to a set of predefined policies that govern the interaction. These policies, which serve as the "predefined rules" mentioned earlier, ensure that the generated responses are not only informative but also empathetic, safe, and culturally sensitive. They guide the system's behavior and responses, shaping the overall user experience. Some of the key policies include:

1. Goal orientation: The system strives to identify and address the caregiver's primary goal for the conversation.
2. Brevity: Responses are kept concise and to-the-point to promote comprehension, particularly in voice-based interactions.
3. Language tailoring: The complexity and terminology of the responses are adapted based on the caregiver's education level and proficiency, as inferred from the knowledge graph.
4. Iterative clarification: Focused follow-up questions are used to pinpoint the exact nature of the query, enabling more precise and relevant responses.
5. Empathy and support: Responses integrate expressions of understanding and empathy for the challenges caregivers face, providing emotional support alongside information.
6. Safety protocols: The system is designed to identify situations where the caregiver or patient may be at risk and prompt action toward emergency or support resources.

For example, if a caregiver asks, "My mom is wandering at night," the system would consider:

q: "My mom is wandering at night."

c: Mom has mid-stage Alzheimer's, lives with caregiver, no recent medication changes.

k: ADRD knowledge graph provides information on wandering, risks, and management strategies.

p: Empathy, Safety, Resource Provision.

r: "Wandering can be concerning. Is your mom safe when this happens? I have some tips to minimize risk, I can email you the links to safety resource. Would you like strategies to help redirect her at night?"

Prototype implementation

We developed the ADQueryAid prototype using OpenAI's API, specifically GPT-4 models, and the Python FAST API framework. Each GPT assistant was assigned a specific role and set of tasks, adhering to predefined policies. Neo4j was used to construct the knowledge graph, providing context for the assistants. Queries were facilitated through the Python FastAPI server, allowing efficient retrieval of relevant information. Conversation history was maintained in the form of thread messages, ensuring contextual continuity for personalized interactions. Unstructured data, such as Alzheimer's-related information and caregiving guidelines, was stored within the files section of the GPT assistant, complementing the structured knowledge from the graph. The front end was developed using the Jinja2 template engine, creating a user-friendly interface for caregivers.

Evaluation methodology

Ethics approval. This study was reviewed and approved by the institutional review board of NDSU. The IRB Protocol number is IRB0005069. We employed a dual-pronged approach to evaluate ADQueryAid, ensuring a comprehensive assessment of its capabilities and usability.

Scenario-based use case studies

First, we conducted scenario-based use case studies, simulating real-world caregiving situations to gauge the system's performance in addressing diverse challenges and information needs. By immersing ADQueryAid in these scenarios, we could observe its responses in context, assessing their accuracy, relevance, and overall effectiveness in meeting caregiver needs. This approach offered valuable insights into the system's strengths and weaknesses, guiding our refinement efforts.

User study with ADRD caregivers.

1. Recruitment process: Participants were recruited through online Alzheimer's advocacy and support communities, such as SubReddit (r/Alzheimers, r/caregivers, r/CaregiverSupport), ALZConnected, Facebook Groups (Walk to End Alzheimer's- Fargo-Moorhead, Alzheimer's Association Minnesota-North Dakota), and Nextdoor Fargo. While internet-based recruitment may initially seem informal, it is increasingly recognized as a valid and effective method for reaching geographically dispersed or difficult-to-reach populations like ADRD caregivers who are often active in online communities seeking support and information.

To ensure a systematic and targeted recruitment process, we implemented the following steps:

a. Recruitment notices: Carefully crafted recruitment messages explaining the study's purpose, scope, and potential benefits were displayed in prominent places within the online groups, with the permission of community moderators.
b. Sampling methodology: We employed both convenience sampling, targeting caregivers already participating in these support groups, and purposive sampling, selecting specific online communities catering to ADRD caregivers. This ensured our sample included individuals with direct caregiving experience, which was a key criterion for our study.
c. Informed consent and inclusion/exclusion criteria: We obtained informed consent from all participants through an online consent form detailing the study's purpose, procedures, potential risks, and their rights. Participants had to be 18 years or older, currently serving (or have served) as caregivers for individuals with ADRD, and have reliable internet access. Individuals under 18 or those who were not caregivers were excluded.

2. Blind, randomized interaction: Participants engaged in conversations with both ADQueryAid and the base GPT-3.5 model in a blind, randomized fashion. They completed specific tasks using both chatbots without knowing which system they were interacting with at any given time.
3. Data collection and analysis: We recorded their interactions, including chat history and feedback on usability and effectiveness using the Chatbot Usability Questionnaire (CUQ), a validated tool designed to assess various aspects of chatbot usability. These data were analyzed to compare ADQueryAid's performance against ChatGPT 3.5.

Comprehensive assessment

By combining scenario-based use cases and the user study with real caregivers, we aimed to comprehensively assess ADQueryAid's capabilities and usability. This methodology provided valuable insights into the system's strengths and areas for improvement, directly informed by the experiences and needs of the target user group. These findings will guide future development and refinement efforts, ensuring ADQueryAid continues to evolve as a practical and user-friendly tool for ADRD caregivers.

Data security and privacy

We recognize the importance of data security and privacy, particularly in the context of healthcare applications like ADQueryAid. While our current

prototype incorporates measures to protect user data, we also acknowledge the limitations of relying on a third-party service like ChatGPT, especially concerning HIPAA compliance.

In the present implementation, we have taken steps to safeguard user privacy, including:

1. Data minimization: We strictly limit the collection of personal data to only what is essential for providing personalized and contextually relevant support.
2. Secure storage: All user data is encrypted both at rest and in transit using industry-standard encryption protocols. Access to data is restricted to authorized personnel only.
3. De-identification: We employ techniques to de-identify user data wherever feasible, replacing personally identifiable information (PII) with unique identifiers to minimize the risk of re-identification.

While these measures represent a good-faith effort to safeguard user privacy, we acknowledge the limitations of the current implementation, particularly the reliance on ChatGPT, which is not inherently HIPAA-compliant.

In future work, we plan to address this limitation by:

1. Developing or adopting a HIPAA-compliant LLM: We will actively explore the development or adoption of a large language model that is specifically designed to meet HIPAA requirements, ensuring full compliance with healthcare data privacy regulations.
2. Enhanced security measures**:** We will continue to strengthen our overall security posture by implementing stricter access controls, conducting regular security assessments, and providing comprehensive user education on data privacy and security.

We are committed to upholding the highest standards of data security and privacy. While our current prototype has limitations in terms of HIPAA compliance, we are actively working towards a fully compliant solution in future iterations of ADQueryAid.

Data availability

User demographic data is included in the manuscript, with more detailed demographic information omitted to protect user identities. Responses from all 20 participants to the questionnaire are provided in the supplementary information files, which were used to derive the results presented in our findings. Chat history is not provided to safeguard user privacy.

Received: 5 July 2024; Accepted: 2 October 2024;
Published online: 04 December 2024

References

1. Xie, B., Tao, C., Li, J., Hilsabeck, R. C. & Aguirre, A. Artificial intelligence for caregivers of persons with Alzheimer's disease and related dementias: systematic literature review. *JMIR Med. Inform.* **8**, e18189 (2020).
2. Li, J., Maharjan, B., Xie, B. & Tao, C. A personalized voice-based diet assistant for caregivers of Alzheimer disease and related dementias: system development and validation. *J. Med. Internet Res.* **22**, e19897 (2020).
3. Alloatti, F. et al. "Can you help me measure my blood sugar?" Co-design of a voice interface to assist patients and caregivers at home. In *2021 IEEE Symposium on Computers and Communications (ISCC)* 1–4 (IEEE, 2021).
4. Kocaballi, A. B. Conversational AI-Powered Design: ChatGPT as Designer, User, and Product. In *Woodstock '18: ACM Symposium on Neural Gaze Detection*, 2018, Woodstock, NY (ACM, 2023).
5. Wang, B., Li, G. & Li, Y. Enabling conversational interaction with mobile UI using large language models. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* 1–17 (ACM, 2023).
6. Kung, T. H. et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLoS Digit. Health* **2**, e0000198 (2023).
7. Gilson, A. et al. How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? The implications of large language models for medical education and knowledge assessment. *JMIR Med. Educ.* **9**, e45312 (2023).
8. Han, Z., Battaglia, F., Udaiyar, A., Fooks, A. & Terlecky, S. R. An explorative assessment of ChatGPT as an aid in medical education: use it with caution. *Med. Teach.* **46**, 657–664 (2024).
9. Nastasi, A. J., Courtright, K. R., Halpern, S. D. & Weissman, G. E. Does ChatGPT Provide Appropriate and Equitable Medical Advice? A Vignette-Based, Clinical Evaluation Across Care Contexts. *medRxiv*. https://doi.org/10.1101/2023.02.25.23286451 (2023).
10. Hopkins, A. M., Logan, J. M., Kichenadasse, G. & Sorich, M. J. Artificial intelligence chatbots will revolutionize how cancer patients access information: ChatGPT represents a paradigm-shift. *JNCI Cancer Spectr.* **7**, pkad010 (2023).
11. Koopman, B. & Zuccon, G. Dr ChatGPT, tell me what I want to hear: How different prompts impact health answer correctness. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Bouamor, H., Pino, J., & Bali, K. (Eds.), Association for Computational Linguistics, Singapore, pp. 15012–15022. https://doi.org/10.18653/v1/2023.emnlp-main.928 (2023).
12. White, J. et al. A prompt pattern catalog to enhance prompt engineering with ChatGPT. Preprint at https://arxiv.org/abs/2302.11382 (2023).
13. Perry, N., Srivastava, M., Kumar, D. & Boneh, D. Do Users Write More Insecure Code with AI Assistants? In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security (CCS '23)*, Association for Computing Machinery, New York, NY, USA, 2785–2799. https://doi.org/10.1145/3576915.3623157 (2023).
14. Dang, H., Mecke, L., Lehmann, F., Goller, S. & Buschek, D. How to Prompt? Opportunities and Challenges of Zero- and Few-Shot Learning for Human-AI Interaction in Creative Applications of Generative Models. *Generative AI and HCI Workshop at CHI 2022*. https://arxiv.org/abs/2209.01390 (2022).
15. Qin, G. & Eisner, J. Learning how to ask: querying LMs with mixtures of soft prompts. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Association for Computational Linguistics, Stroudsburg, 2021).
16. Li, J. et al. Development and evaluation of ADCareOnto—an ontology for personalized home care for persons with Alzheimer's disease. In *2020 IEEE International Conference on E-health Networking, Application & Services (HEALTHCOM)* (IEEE, 2021).

Acknowledgements

The authors would like to thank all the participants for their contributions to the study. This work was supported by the National Science Foundation (NSF) with award number 2218046 and the National Institutes of Health (NIH) U01AG088076.

Author contributions

W.U.H. was involved in the system design and implemented the entire chatbot system. K.T.Z. designed and executed the user study, including data analysis, and managed the IRB protocol. X.W. contributed to the design and development of the knowledge graph and conducted the use case evaluation. J.L. proposed the overall research idea, led the design, and drafted the manuscript. C.T., B.X., and J.L. provided guidance and oversight for the research.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s44385-024-00004-8.

Correspondence and requests for materials should be addressed to Juan Li.