Original Paper

Empowering Informal Caregivers of Persons With Early-Stage Dementia by Large Language Models: Mixed Methods Evaluation

Huayu Zhou[1], MEng, MSci; Ziwei Zhu[2], PhD; Kyeung Mi Oh[3], PhD; Sungsoo Ray Hong[1], PhD

[1]Department of Information Sciences and Technology, College of Engineering and Computing, George Mason University, Fairfax, VA, United States
[2]Department of Computer Science, College of Engineering and Computing, George Mason University, Fairfax, VA, United States
[3]School of Nursing, College of Public Health, George Mason University, Fairfax, VA, United States

Corresponding Author:

Sungsoo Ray Hong, PhD
Department of Information Sciences and Technology
College of Engineering and Computing, George Mason University
Research Hall Building, Room 211, 10401 York River Road
Fairfax, VA 22030
United States
Phone: 1 2062291872
Email: shong31@gmu.edu

Abstract

Background: Acquiring relevant knowledge and support is essential for informal caregivers of persons with early-stage dementia, including awareness, access, and use of comprehensive resources for both persons with dementia and caregiver support. With appropriate strategies and early-stage support, informal caregivers can play a vital role in enhancing the well-being of persons with dementia and potentially slowing their progression. While large language models (LLMs) can provide easy access to caregiving knowledge, the risks, perceived challenges, and ways to improve LLM-generated responses in practice remain underexplored.

Objective: In this study, we aim to (1) examine the risks and perceived challenges of using a baseline ChatGPT-4o, an internet-accessible artificial intelligence model, for dementia caregiving support and (2) understand how an enhanced version of ChatGPT-4o, equipped with up-to-date dementia caregiving knowledge, can mitigate these risks and challenges.

Methods: We compiled 32 representative questions from informal caregivers seeking guidance on early-stage dementia. We developed two ChatGPT-4o conditions: C1, the publicly available baseline model, and C2, an experimental version enhanced through prompt engineering and grounded in a conceptual framework—drawn from health science and gerontology literature—to empower caregivers of individuals with early-stage dementia. Using these conditions, we generated 64 responses (32 pairs) to the questions. Twelve experts evaluated them with validated tools assessing accuracy, reasoning, clarity, usefulness, trust, satisfaction, safety, harm, and relevance. A Mann-Whitney *U* test compared the conditions. After the survey, we conducted interviews to explore experts' perceived differences, remaining challenges, and design opportunities. Interviews were transcribed and analyzed using descriptive thematic analysis.

Results: Responses in C2 showed significant improvements in 3 criteria—actionability, relevance, and perceived satisfaction—compared to C1. However, no significant differences were found in the remaining 5 criteria: response accuracy, the model's ability to understand the question, intelligibility, trustworthiness, response safety, and perceived harm. Qualitative analysis of interviews revealed two key insights: (1) differences between baseline and experimental responses and (2) possible reasons for these differences. Twelve experts evaluated wordiness, detail, empathy, satisfaction, accuracy, relevance, and bias. Both models were considered somewhat verbose, but the experimental model's responses were viewed as more detailed, relevant, and actionable. Accuracy appeared similar across models, yet participants reported greater satisfaction with the experimental model's outputs.

Conclusions: Results indicate that both conditions generated responses perceived as reasonable and intelligible. However, the experimental model offered more relevant, practical guidance on caregiving needs, providing specific information aligned with the 32 testing questions and actionable recommendations. This led to higher perceived satisfaction compared to the baseline model.

JMIR Form Res 2026;10:e79975; doi: 10.2196/79975

Keywords: large language model; human-computer interaction; human-AI collaboration; early-stage dementia; informal caregivers; prompt engineering

Introduction

Informal caregivers' dementia literacy plays a critical role in mitigating the progression of dementia; however, obtaining relevant knowledge and effectively applying it in real-life situations can be difficult. These challenges are further compounded by limited access to information and resources related to caregiving. The aging population is increasing rapidly; in 2024, about 6.9 million Americans aged 65 and older are living with Alzheimer's or dementia, a number expected to double by 2060 [1]. Dementia is a leading contributor to older mortality, with one in three individuals over age 85 dying with the disease [2,3]. Informal caregivers, who support persons with dementia, typically family members, face barriers such as limited English proficiency [4], inadequate health insurance [5], low income [6], and social stigma [7], which hinder access to essential dementia knowledge and effective care strategies [8].

The lack of accessible, accurate, and personalized knowledge contributes to both ineffective care and substantial socioeconomic burdens [9,10]. In 2019, the worldwide cost of dementia care was US $1.3 trillion [11], including medical expenses, social care, and informal caregiving, with US $196 billion spent on direct medical costs in the United States alone in 2020 [12]. Beyond finances, caregivers and persons with dementia experience declines in quality of life, as persons with dementia lose autonomy and dignity [13,14] and caregivers face heightened psychological and physical stress [15], yet few tools provide tailored guidance for informal caregivers, highlighting a critical gap in support.

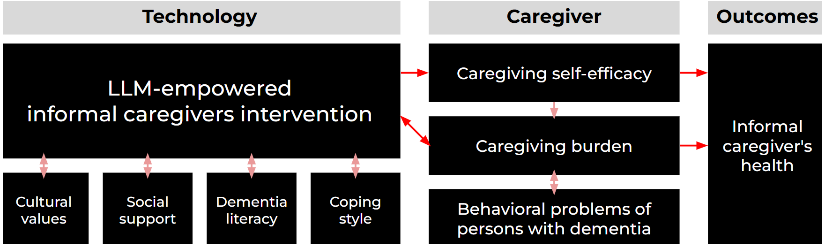
Addressing informal caregivers' limited access can empower them to manage their responsibilities more effectively [16] and strengthen their emotional resilience [17] by recognizing early symptoms [18], managing initial behavioral changes effectively [8], and establishing consistent routines to reduce confusion and anxiety [19]. Additionally, informed informal caregivers can create a stable and safe environment that reduces stress and enhances emotional well-being while also engaging persons with dementia in cognitive and social activities that help maintain mental function, support independence, and foster a sense of purpose [20]. To further support informal caregivers, culturally tailored support can provide personalized guidance respecting caregivers' cultural and language needs [21-23], enhancing the accessibility and relevance of dementia care information.

Although large language models (LLMs) can be an effective technological aid for providing dementia-related knowledge [24,25], there have been concerns about applying these tools to informal caregivers. For instance, using LLMs often fails at providing valid information [26], relevant answers based on personalized context, nonactionable suggestions [27,28], answers that lack genuine contextual understanding and empathy [29], and answers with bias related to age, disability, and race, thereby contributing to disparities in care delivery [30]. While studies started applying ChatGPT in various health care applications, including patient interaction [31], clinical diagnosis support [32], telehealth services [33], health education [34], personalized health advice [35], and broader health promotion efforts [36], several limitations and risks must be addressed to ensure the safe and effective integration of LLMs into dementia caregiving. Finally, it is worth noting that the use of LLMs in caregiving raises privacy and data security concerns; without robust safeguards, sensitive information may be vulnerable to unauthorized access or misuse.

While prior research has extensively documented the burden of dementia and the challenges faced by informal caregivers, fewer studies have examined the potential of artificial intelligence (AI)–driven digital health tools to support caregiving. Recent work has begun to explore the application of LLMs in this domain. For example, Kim et al [27] analyzed digital caregiving strategies for Alzheimer disease and related dementias and highlighted how AI-driven tools can align caregiving tasks with disease stage–specific needs. Pérez-Esteve et al [37] demonstrated that GPT-4o models can provide tailored guidance and reduce caregiver errors, although they do not yet replace professional instruction. Similarly, Shi et al [38] developed Carey, a GPT-4o–based chatbot to provide informational and emotional support to Alzheimer disease and related dementias caregivers, identifying key caregiver needs to inform AI design. Hasan et al [39] introduced ADQueryAid, a conversational AI system leveraging LLMs with retrieval-augmented knowledge to deliver personalized support. Additionally, Chien et al [40] showed that LLMs could accurately identify high-burden caregivers in long-term care settings, highlighting their utility in caregiver assessment. Building on these developments, our study investigates how broadly accessible everyday LLMs can both challenge and benefit informal caregivers, aiming to improve their practical utility in real-world caregiving contexts.

In this study, we used ChatGPT, a publicly available LLM model that is considered the most advanced one at the time of the study, as our baseline condition (C1). In parallel, we designed an experimental condition (C2) using a prompt-engineered ChatGPT-4o model, enhanced through a conceptual framework shown in Figure 1, which is built upon our comprehensive review of literature in geriatrics and gerontology, with a specific focus on dementia caregiving.

Figure 1. Conceptual framework integrating technology, caregiver factors, and outcomes in dementia caregiving. The framework integrates the sociocultural stress and coping model [41] and the self-efficacy model [42] for caregivers to illustrate the dynamic interactions among contextual, psychological, and technological factors. It comprises three main components: (1) technology, referring to LLM-empowered interventions that support informal caregivers; (2) caregiver, encompassing factors such as caregiving self-efficacy, caregiving burden, and behavioral challenges of persons with dementia; and (3) outcomes, focusing on the health and well-being of informal caregivers. LLM: large language model.



Our study uniquely explores how widely LLMs affect informal caregivers, highlighting both the challenges and opportunities they present for enhancing caregiving support in everyday, real-world contexts. Then, this study is guided by the following four research questions: (RQ1) What risks and challenges do informal caregivers encounter when using everyday LLMs (eg, ChatGPT-4o) for dementia care support? (RQ2) How can ChatGPT be enhanced through prompt engineering with dementia-related knowledge, caregiving skills, and culturally relevant supports? (RQ3) To what extent does the enhanced ChatGPT improve caregivers' experiences in terms of (1) actionability, (2) relevance, (3) perceived satisfaction, (4) response accuracy, (5) model's ability to understand the question, (6) intelligibility, (7) trustworthiness, and (8) response safety and perceived harm? (RQ4) What are the implications of these findings for the future design of AI tools tailored to caregiver–person with dementia dyads?

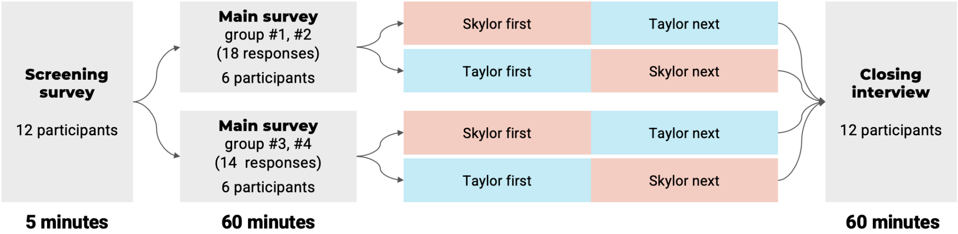
Methods

Study Design

A mixed methods study using surveys and interviews was conducted via Zoom to evaluate a prompt-engineered ChatGPT-4o compared to baseline ChatGPT-4o—an internet-accessible AI tool—based on the evaluation criteria listed above. The study aimed to assess its potential to support informal caregivers of persons with early-stage dementia.

Figure 2 shows a clear study flow diagram applied to summarize the proposed mixed methods. The overall study pipeline proceeded as follows. (1) Each participant first completed a brief 5-minute screening survey to collect professional background information. Participants who passed the screening were assigned a unique ID number used to record their results from both the quantitative (1 hour) and qualitative (1 hour) studies. (2) In the 1-hour quantitative study, participants 1-6 were divided into group 1 and group 2. Each group evaluated 18 testing scenarios containing 18 pairs of LLM-based responses while completing corresponding evaluation surveys. Similarly, participants 7-12 were divided into group 3 and group 4, and each group evaluated 14 testing scenarios with 14 pairs of LLM-based responses, also accompanied by quantitative evaluation surveys. (3) Finally, during the 1-hour qualitative study, all participants completed a semistructured interview consisting of general and specific questions. The corresponding study materials are covered in Multimedia Appendix 1.

Figure 2. Study flow diagram illustrating the mixed methods design. The study began with a 5-minute screening survey. Participants who met eligibility criteria were assigned a unique ID and participated in a 1-hour quantitative session. Participants in groups 1 and 2 (n=6) evaluated 18 pairs of test scenario–based LLM responses, while those in groups 3 and 4 (n=6) evaluated 14 pairs. Under each test scenario, two types of LLM responses were labeled "Skylor" and "Taylor." Finally, all participants completed a 1-hour semistructured qualitative interview. LLM: large language model.



Prompt Engineering Design

We used ChatGPT-4o (September 2024 till November 2024). To prepare C2, we compiled a dataset of 90 question-and-answer pairs to inform the prompt engineering of C1. These pairs were developed across 9 domains. Four of them are key domains identified in the conceptual framework: cultural values, social support, dementia literacy, and coping style. To create the question-and-answer pairs, we collaborated with community outreach lay health workers with over a decade of experience working with persons with dementia and their caregivers. Drawing on their expertise, these professionals generated questions commonly asked by informal caregivers of persons with early-stage dementia within each of the 4 domains. Subsequently, an experienced nurse practitioner with over a decade of clinical practice responded to the generated questions. The answers were grounded in evidence-based guidelines from credible organizations, including the Alzheimer's Association [43], the National Institutes of Health, the National Institute on Aging [44], the Retirement Research Foundation [45], and the Korean Medical Center [46]. Following the initial drafting, one of the authors (KMO) with expertise in nursing and gerontology reviewed the responses to ensure their accuracy, clarity, and relevance. This iterative process resulted in a finalized dataset of 90 question-and-answer pairs attached in Multimedia Appendix 2. Selected examples are provided in Table 1.

Table 1. Selected examples from the 90 question-and-answer pairs across 4 domains: cultural values, social support, coping style, and dementia literacy.

	Questions	Answers	Domains
Q1	I am Hispanic and my spouse was just diagnosed with Alzheimer's disease—how can I be sure I'm getting the care I need for my Hispanic spouse?	• The Alzheimer's Association has partnered with several agencies within different cultures to make sure care and help are provided appropriately. Your first resource would be to reach out to the National Hispanic Council on Aging and you can find them here: [47].	Cultural values
Q11	Are there places I can go to socialize? I feel so isolated since my spouse was diagnosed with Alzheimer's disease. I just need to meet some other people—I don't even have time to go meet my friends.	• Yes, there are places you can go and people you can talk to. You can start with the support groups online that you can find here: [48]. • You can find community resources close to you here: [49]. • You can also call the Alzheimer's Association's hotline and talk to an actual person that can help you 800.272.3900. Or you can click this link for resources specific to the National Capital Area (NCA) [50].	Social support
Q21	What is Alzheimer's disease? What do I tell my family?	• Alzheimer's disease is the most common form of dementia—more information here: [51]. • Regarding what to tell your family, remember Alzheimer disease is nobody's fault and there is help available. For help telling your family you can start here: [52].	Dementia literacy
Q31	My spouse is my best friend and the one I go to with everything, and now I feel like I can't because my spouse is what I'm worried about, my spouse is why I need coping help. I don't know what to do.	• It is hard when your person is the one you have concerns and worries about. There are resources available to help you. There are online support groups, local activities and support groups, and a helpline available 24/7. You can find the online support groups here: [48]. • You can find local activities and supports groups here: [49] and the helpline is 800.272.3900.	Coping style

Prompt engineering is essential for adapting ChatGPT to specific contexts, such as supporting informal caregivers of individuals with early-stage dementia. Studies have shown that prompt design significantly influences the quality, relevance, and accuracy of AI-generated responses. For instance, Li et al [53] found that structured, role-specific prompts improved ChatGPT's performance in dementia care scenarios. Without such guidance, general-purpose models may produce vague or misleading information, particularly in sensitive health-related contexts. Parmanto et al [54] highlighted that unguided prompts often result in irrelevant or unreliable outputs. Prompt engineering helps mitigate these issues by narrowing the model's focus, reducing hallucinations, and encouraging context-specific, empathetic responses. A study by Cheng et al [55] further demonstrated that prompt-engineered AI can provide personalized, culturally sensitive support tailored to caregiver needs and the stages of dementia. Ultimately, prompt engineering enhances clarity, aligns output with user intent, and bridges the gap between general AI capabilities and the specialized demands of caregiving.

In our work, after collecting the 90 question-and-answer pairs, we proceeded to generate the corresponding 90 prompts for C2 by applying prompt engineering techniques to C1. The 90 prompts for C2 are provided in Multimedia Appendix 3. These examples show the details of how the prompts were generated with context-related information. In designing these prompts, we drew on principles of prompt engineering grounded in five fundamental factors [56]: (1) clarity and specificity—ensuring that prompts are unambiguous and clearly stated; (2) user intent—aligning the prompt with the user's needs and the desired output; (3) model understanding—considering the strengths and limitations of ChatGPT to craft prompts that maximize its capabilities while mitigating its weaknesses; (4) domain specificity—incorporating relevant terminology and context to guide the model toward more accurate and relevant responses; and (5) constraints—explicitly defining requirements such as response length or format.

Generating Responses From C1 and C2 for Evaluation

To select a natural and realistic set of questions that informal caregivers may ask frequently in practice, we collaborated with two experts: a community outreach lay health worker and an academic researcher, both with over a decade of experience working with persons with dementia and their caregivers. They were independent of the community caregiving lay health workers involved in the prompt engineering design, minimizing potential bias during the evaluation. They were asked to generate 32 test questions commonly posed by caregivers of persons with early-stage dementia across 4 domains: cultural values, social support, coping style, and dementia literacy. The dementia literacy domain includes 4 subcategories: causes and characteristics, health risks and promotion, communication and behavior, and care considerations. The 32 testing questions are provided in Multimedia Appendix 4.

Using the 32 questions, we generated 64 LLM responses: 32 from C1 and another 32 from C2. Specifically, for each question, we manually entered it into the LLM interface to initiate response generation. Once the LLM's textual output was displayed, we collected it directly as a single response. Each response was saved and assigned a unique identifier corresponding to its input question.

Participants

In January 2025, purposive sampling was used to recruit 12 participants through a screening survey assessing eligibility based on predefined inclusion and exclusion criteria. Eligible participants included informal caregivers of persons with dementia, as well as health care professionals or community outreach program leaders with at least 1 year of experience working with persons with dementia or their caregivers. Recruitment continued until qualitative data saturation was reached, that is, we intended to keep recruiting participants as long as each new case enhanced our understanding; recruitment would end once additional participants no longer offered new information. Moreover, Caine [57] conducted the analysis which revealed that 12 participants appeared most frequently among CHI papers—effectively becoming a de facto community norm for small-scale human-computer interaction (HCI) user studies. This was not presented as the optimal number, but rather as the most common practice (a cultural standard in CHI research). Furthermore, these participants are considered content experts and invited to evaluate LLM-generated responses using validated tools measuring accuracy, reasoning, clarity, usefulness, trust, satisfaction, safety, harm, and relevance.

Data Collection Procedures

From the 64 LLM responses generated by C1 and C2, content experts conducted evaluations using two methods: a quantitative survey and a qualitative follow-up semistructured interview conducted via Zoom. Each session lasted about 2 hours on average—1 hour for the survey and 1 hour for the follow-up interview. Participants received a US $50 Amazon gift card in appreciation of their time.

The 12 experts were divided into 4 groups of 3 to ensure deeper and more focused discussion, as reviewing all 32 questions together could limit individual focus and engagement. Of the 32 testing questions or scenarios, 6 are from the cultural values domain, 7 from social support, 5 from coping style, and 14 from dementia literacy. For the evaluation surveys, groups 1 and 2 assessed 18 scenarios drawn from the cultural values, social support, and coping style domains, while groups 3 and 4 evaluated the remaining 14 from the dementia literacy domain.

Selected examples are provided in Table 2.

Table 2. Selected examples from the 32 test questions assigned to groups across 4 domains: cultural values, social support, coping style, and dementia literacy.

	Group ID	Test question example	Domains
Q5	G1&G2	I live in Virginia and am a caregiver for my family member with dementia. What local resources or organizations in Spanish can provide me with support, whether for caregiving or guidance on managing the situation?	Cultural values
Q10	G1&G2	As a caregiver in Virginia, I'm looking for local resources or organizations that can provide support for caregivers like me. What options are available to help me navigate this journey?	Social support
Q11	G1&G2	I'm looking to connect with other caregivers in my community for support. Are there any peer support groups or opportunities to meet others who are going through similar experiences? Can you suggest any online support groups or forums for caregivers?	Social support
Q14	G1&G2	I would like to connect with someone who is experiencing similar situations. I feel overwhelmed as I am responsible for taking care of both my children and my father at home. How can I ensure I take care of myself as well?	Coping style

	Group ID	Test question example	Domains
Q2	G3&G4	My husband’s behavior has been changing lately, and I’m not sure if it’s just because he’s getting older or if it could be something like dementia. He forgets things, gets confused, and sometimes doesn’t act like himself. What are the early signs of dementia, and how can I tell if this is just aging or something more serious?	Dementia literacy
Q8	G3&G4	I live in Fairfax County and I’m looking for activities or therapies in Korean that might help slow down my husband’s dementia. Are there any programs or services available that have been proven to help with this?	Dementia literacy
Q10	G3&G4	My husband has been acting differently, and sometimes he gets upset or does things that are hard to understand. How can I handle these behaviors, especially if it’s because of dementia?	Dementia literacy
Q14	G3&G4	My husband has been showing signs of mild memory problems, like forgetting names or where he put things, but I’m not sure if this is just part of getting older or something more serious. I’ve heard about mild cognitive impairment (MCI), and I’m worried that it might get worse. What are the risks if his MCI progresses to full dementia, and how can I prepare for what might come next?	Dementia literacy

In the survey, experts reviewed responses in a within-subjects design with a counterbalanced order to eliminate order effects. To avoid bias, we used gender-neutral labels—“Skylor” for C1 and “Taylor” for C2—and withheld information about how these responses were generated. The two responses to test question no. 18 from the coping style domain are presented as an example in Figure S1 in Multimedia Appendix 5. More examples are shown in Table 2. After the survey, we conducted semistructured interviews to gain a deeper understanding of (1) when and why experts felt that Taylor’s or Skylor’s responses were not fully satisfactory; (2) when and why experts felt that responses from Taylor or Skylor were superior to others; (3) what general challenges experts encountered during the interviews; and (4) given the rapid progress in LLMs, which aspects future technology should prioritize to deliver higher-quality caregiving knowledge to informal caregiver–person with dementia dyads. The interview questions are presented in Multimedia Appendix 6.

Evaluation Measures for Quantitative Data

We adapted our evaluation measures from a literature-based framework for human evaluation of LLMs in health care [58]. This multidimensional framework assessed accuracy, understanding and reasoning, clarity and coherence, usefulness and supportiveness, trust and confidence, overall satisfaction, safety and harm, and relevance. Each dimensional measure was rigorously validated in previous studies [34,35,59-68] and assessed using a single item, except for safety and harm, which was evaluated with two separate items. Building on this foundation, we developed our own expanded set of evaluation measures, as shown in Table 3.

Table 3. The 9 evaluation items and their corresponding dimensions adapted from a literature-based framework for large language model assessment in health care.

	Evaluation items	Dimensions
Q1	The response is accurate.	Accuracy of responses [61,63,64]
Q2	The response demonstrates correct reading comprehension, showing that it understands the question.	Understanding and reasoning [65]
Q3	The language used in the response is clear and easy to understand.	Clarity and coherence [61,68]
Q4	The response includes actionable advice or resources.	Usefulness and supportiveness [59,60]
Q5	The response is reliable and based on credible medical and scientific sources.	Trust and confidence [66]
Q6	I am satisfied with the chatbot's response overall.	Satisfaction [34]
Q7	The response provided is balanced and free from bias.	Safety [67]
Q8	The response contains no potentially harmful information.	Harm [35]
Q9	The response provides relevant answers.	Relevance [60-62]

In total, 9 items were rated on a 7-point Likert scale, where 1=strongly disagree and 7=strongly agree; higher scores indicate greater perceived quality.

Data Analysis

Quantitative Analysis

Histograms were generated for each 7-point evaluation item. To compare C1 and C2, Mann-Whitney *U* tests were conducted for each criterion. Moreover, the comparisons of the mean and SD values were reported for each evaluation criterion as well. Additionally, all analyses were performed in Python using pandas, matplotlib, seaborn, and scipy.

Qualitative Analysis

The interviews were audio-recorded and transcribed. We then conducted an iterative qualitative coding process [69,70] to analyze the data. One of the authors (HZ) independently read the transcription and conducted coding by tagging specific text in the transcriptions. As the author coded multiple transcriptions, she started finding commonalities and discrepancies between transcriptions and wrote her findings in the analytic memos. As the author started finding no new topics for codes and memos, she started building themes and categories together. Finally, after reviewing all her coded text snippets, memos, and insightful quotes, she refined the final structure and details of the findings. Every analysis has been done without the help of AI, other than transcription.

Ethical Considerations

This study was reviewed and approved by the institutional review board at George Mason University (STUDY00000288). All participants provided informed consent prior to participation. To protect privacy and confidentiality, all personal identifiers were removed, and audio recordings were securely stored on password-protected servers accessible only to the research team. Participation was entirely voluntary, and participants could withdraw from the study at any time without penalty. Each participant received a US $50 honorarium as compensation for their time and contributions. All study procedures adhered to the ethical principles outlined in the Declaration of Helsinki and complied with JMIR and COPE (Committee on Publication Ethics) standards for research involving human participants.

Results

Study Participants

Twelve content experts participated, each with over a decade of professional experience and substantial involvement with older adults with dementia and their caregivers, including informal caregivers. Their professions included registered nurses, nurse practitioners (psychiatric-mental health, geriatric, and family health), physicians, social workers, and a community outreach program director. Participant details are provided in Table 4.

Table 4. Demographic and professional characteristics of the 12 content experts in this study.

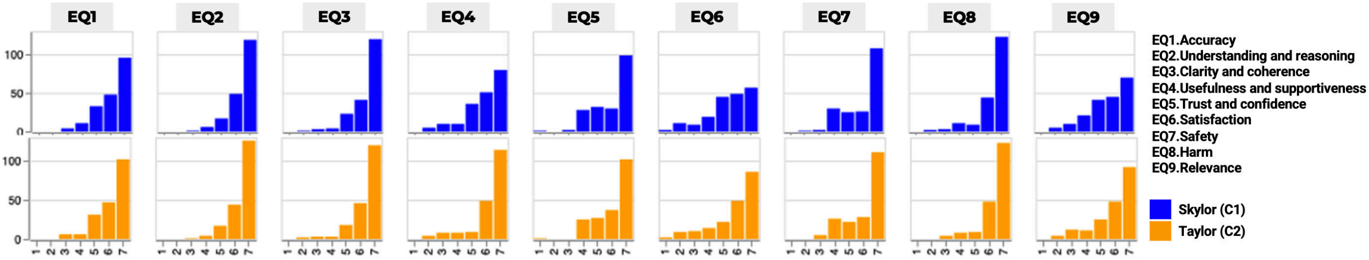
Participant ID	Gender	Age (years), range	Profession	Primary work settings
P1	Female	45-54	Registered nurse	Academic institution
P2	Female	45-54	Physician	Academic institution
P3	Female	55-64	Physician	Academic institution
P4	Female	35-44	Registered nurse	Academic institution
P5	Female	45-54	Social worker	Academic institution
P6	Female	55-64	Registered nurse	Academic institution
P7	Female	35-44	Family nurse practitioner	Academic institution
P8	Female	65 or older	Informal caregiver	Retired
P9	Female	65 or older	Informal caregiver	Retired
P10	Female	55-64	Social worker	Academic institution
P11	Male	35-44	Psychiatric-mental health nurse practitioner	Community-based organization
P12	Female	35-44	Geriatric nurse practitioner, research nurse specialist	National Institutes of Health Clinical Center

Survey Results

Our evaluation comprised 9 items (Q1-Q9) listed in Table 3, each rated on a 7-point Likert scale (1=strongly disagree, 7=strongly agree). We collected 192 responses for each evaluation item: 36 for cultural values, 42 for social support, 30 for coping style, and 84 for dementia literacy domains. Next, we used Mann-Whitney *U* tests to compare the baseline and experimental models on each evaluation item. For each comparison, the *U* statistic and *P* value were calculated to determine if there was a statistically significant difference between the 7-point Likert scale distributions, indicating whether the evaluation responses under the two conditions differed.

The key research findings across all 4 domains—cultural value, social support, coping style, and dementia literacy—are presented in Figure 3. In the conceptual framework, that is, Figure 1, we can tell that the data used for prompt engineering are from all the domains. Specifically, the blue histograms represent the distribution of answers under the baseline condition, while the orange histograms represent those under the experimental condition. Notably, the Mann-Whitney *U* test identified statistically significant differences between the two conditions for evaluation items Q4, Q6, and Q9 (*P*<.05). The statistical details such as *U* statistics and exact *P* values are reported in Table S1 in Multimedia Appendix 7.

Figure 3. Comparison of answer score distributions for evaluation items Q1-Q9 across 4 domains (192 responses): cultural values, social support, coping style, and dementia literacy. Blue histograms represent responses under the baseline condition (C1), and orange histograms represent responses under the experimental condition (C2). Statistically significant differences were observed for Q4 (usefulness and supportiveness), Q6 (overall satisfaction), and Q9 (response relevance), indicating that C2 outperformed C1 in these three aspects (Mann-Whitney *U* test; *P*<.05).

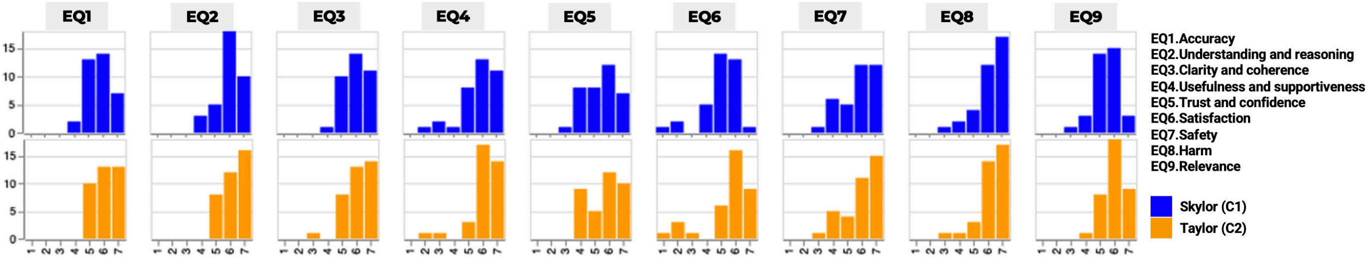


Overall, among the 9 evaluation items, responses generated by the prompt-engineered ChatGPT (ie, the experimental model, C2) outperformed those from the general ChatGPT (ie, the baseline model, C1) on items Q4 (providing actionable advice or resources), Q6 (overall satisfaction with the response), and Q9 (relevance of the answers). In the multidimensional evaluation framework, Q4 is a measure that decides if the response includes actionable advice or resources; Q6 is a measure that determines if the participants are overall satisfied with the given response; additionally, Q9 measures if the given response provides relevant answers. Therefore, this comparison result across all 4 domains says that C2 is better than C1 in terms of usefulness, supportiveness, and satisfaction, as well as relevance.

Comparisons of answer-score distribution in terms of mean and SD values for each evaluation criterion are reported in Tables S1 and S2 in Multimedia Appendix 8. After analyzing Table S1, we observed that, overall, C2 achieved slightly higher mean scores than C1 across most evaluation criteria (EQ1-EQ5, EQ7-EQ9), indicating a modest advantage in performance. However, C1 outperformed C2 on EQ6 (5.44 vs 4.58), representing the largest negative difference (–0.86), while the largest positive difference favoring C2 was observed for EQ4 (+0.40). These results suggest that although C2 generally provides stronger responses, C1 demonstrates notable strength on specific criteria. After analyzing Table S2, we observed that, overall, the variability of scores was similar across most evaluation criteria. The exception was EQ6, where C2 exhibited substantially higher SD (2.16 vs 1.49), indicating less consistent performance, while differences for all other criteria were minimal (<0.12). In summary, across all evaluation criteria, C2 generally achieved slightly higher mean scores than C1, except for EQ6 where C1 outperformed C2. Overall variability was comparable between conditions, though C2 showed notably higher SD on EQ6, indicating less consistent performance for that criterion. These results suggest that while C2 tends to provide marginally better answers on average, its consistency varies across specific evaluation dimensions.

Moreover, within the cultural values domain, which is one of the four domains in the conceptual framework in Figure 1 and provides 36 responses for evaluation, the Mann-Whitney *U* test indicated statistically significant differences between C1 and C2 for evaluation items Q6 (overall satisfaction with the chatbot’s response) and Q9 (relevance of the response), with both *P* values falling below .05. The key research findings are presented in Figure 4. The statistical details such as *U* statistics and exact *P* values are reported in Table S2 in Multimedia Appendix 7. In the multidimensional evaluation framework, Q6 is a measure that determines if the participants are overall satisfied with the given response; additionally, Q9 measures if the given response provides relevant answers. Therefore, this comparison result across the cultural values domain says that C2 is better than C1 in terms of satisfaction and relevance.

Figure 4. Comparison of answer score distributions for evaluation items Q1-Q9 within the cultural values domain (36 responses). Blue histograms represent responses under the baseline condition (C1), and orange histograms represent responses under the experimental condition (C2). Statistically significant differences were observed for Q6 (overall satisfaction) and Q9 (response relevance), indicating that C2 outperformed C1 in these two aspects (Mann-Whitney *U* test; *P*<.05).

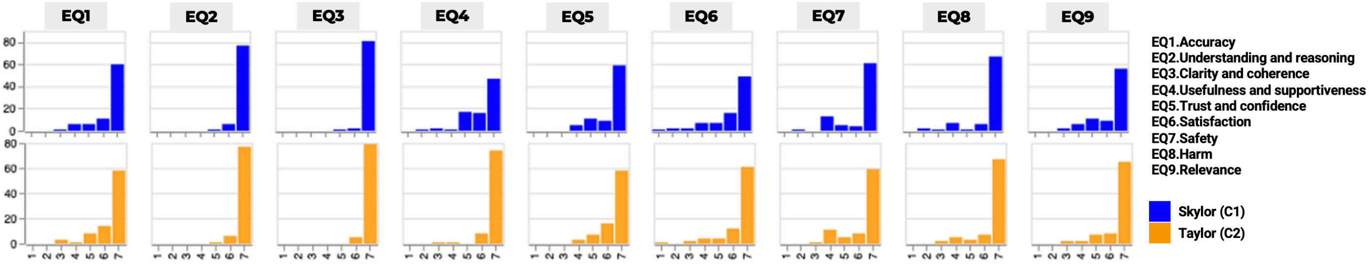


Comparisons of answer-score distribution in terms of mean and SD values for each evaluation criterion are reported in Tables S3 and S4 in Multimedia Appendix 8. After analyzing Table S3, we observed that, for the cultural values domain, C2 consistently achieved higher mean scores than C1 across all evaluation criteria, with differences ranging from modest (+0.16 on EQ7) to more pronounced (+0.53 on EQ6). This indicates that C2 provided generally stronger responses than

C1 in addressing questions related to cultural values. After analyzing Table S4, we observed that, in the cultural values domain, C2 consistently achieved higher mean scores than C1 across all evaluation criteria, indicating generally stronger performance. Variability was comparable for most criteria, though C2 showed slightly higher SD on EQ6 (1.57 vs 1.25), suggesting less consistent responses for that criterion, while differences for other criteria were minimal.

Furthermore, within the dementia literacy domain, which is one of the four domains in the conceptual framework in Figure 1 and provides 84 responses for evaluation, the Mann-Whitney *U* test revealed statistically significant differences between C1 and C2 for evaluation items Q4 (the response includes actionable advice or resources) and Q6 (overall satisfaction with the chatbot's response), with both *P* values below .05. The key research findings are presented in Figure 5. The statistical details such as *U* statistics and exact *P* values are reported in Table S3 in Multimedia Appendix 7. In the multidimensional evaluation framework, Q4 is a measure that decides if the response includes actionable advice or resources; Q6 is a measure that determines if the participants are overall satisfied with the given response. Therefore, this comparison result across the dementia literacy domain says that C2 is better than C1 in terms of usefulness, supportiveness, as well as satisfaction.

Figure 5. Comparison of answer score distributions for evaluation items Q1-Q9 within the dementia literacy domain (84 responses). Blue histograms represent responses under the baseline condition (C1), and orange histograms represent responses under the experimental condition (C2). Statistically significant differences were observed for Q4 (usefulness and supportiveness) and Q6 (overall satisfaction), indicating that C2 outperformed C1 in these two aspects (Mann-Whitney *U* test; *P*<.05).



Comparisons of answer-score distribution in terms of mean and SD values for each evaluation criterion are reported in Tables S5 and S6 in Multimedia Appendix 8. After analyzing Table S5, we observed that in the dementia literacy domain, C2 achieved slightly higher mean scores than C1 for most evaluation criteria, particularly on EQ4 (6.82 vs 6.21) and EQ9 (6.57 vs 6.32). For several criteria, such as EQ1 and EQ2, mean scores were identical or nearly identical, indicating that both conditions performed similarly on certain aspects of dementia literacy. Overall, C2 demonstrated a modest advantage in performance across this domain. After analyzing Table S6, we observed that, in the dementia literacy domain, C2 achieved slightly higher mean scores than C1 for most evaluation criteria, with the largest improvements observed on EQ4 (6.82 vs 6.21) and EQ9 (6.57 vs 6.32). Variability was generally comparable between the two conditions, although C2 exhibited lower SDs on several criteria, including EQ4 (0.60 vs 1.08) and EQ9 (0.92 vs 1.09), indicating more consistent responses. Overall, these results suggest that C2 provided marginally stronger and more consistent performance in addressing dementia literacy questions.

Semistructured Interview Results

Immediately after completing the survey, each participant was invited to a follow-up interview. Before answering the interview questions, they reviewed the test scenarios and the corresponding model responses they had seen during the survey.

Response Differences Between C1 and C2

Qualitative insights from the semistructured interviews provided a deeper understanding of participants' perspectives on the differences between responses from C1 and C2. First, the interviews focused on two central questions: (1) What are the key differences between responses generated under the baseline and experimental conditions? and (2) What factors might explain these differences? Participants discussed aspects such as wordiness, detail, empathy, satisfaction, accuracy, relevance, and potential bias. Overall, both models' responses were viewed as somewhat wordy, but the experimental model's replies were generally considered more detailed, relevant, and actionable. Accuracy was acknowledged in both, yet participants reported higher satisfaction with the experimental model. These findings, organized by overarching themes with related subthemes and illustrative participant quotes, are presented in Table 5.

Table 5. Themes and illustrative quotes summarizing participant feedback on differences between C1 and C2 responses.

Theme and subtheme	Supporting quote	Participant ID
Response delivery difference		
Detailed	"The same observation but Taylor's answers were more detailed, links to resources, information about the location and the transportation."	P7
Empathy	"For the first six questions, Taylor's answers seem to be more empathetic."	P6
Response content difference		
Actionable	"For Taylor, it sorts of provided more actionable ways to get the information that one was desiring."	P5
Satisfied	"Taylor's provided phone numbers, web addresses, links while Skylor's didn't. Taylor provided more links, more useful ways for someone to actually take the action."	P10
Accurate	"Taylor's was kind of like, find a doctor or find a provider. And Skylor's were like, no, go to your general, you know either your Pcp, or your general practitioner.And Taylor's was a little bit more ambiguous. As to, you know. Great to find a doctor that's not terribly helpful. I know I need to do that. But you know which one do I go to see? So, I felt like again, Taylor's was a bit more ambiguous. Skylor's was a little bit more direct."	P2
Relevant	"Taylor seemed to have provided more accurate information and more specific information, or more relevant information based on the scenario in terms of clear and easier to understand."	P6
Biased	"I think there were some points where, like your questions asked about like just bias between, you know, it was this one bias, or that one bias, and just in general, of course, coming from nurse practitioner lens like. I tend to like or favor responses that say, healthcare provider versus doctor, or primary care physician, or whatever, or just include that, as like a little bit more general, and it seems like it wasn't necessarily consistent that, like Skylor, use this language and Taylor use this terminology like it. It kind of was interchangeable. I think like at some points where someone would say doctor and the others would say healthcare provider. Maybe it seemed like in the beginning, Taylor was saying, like healthcare provider, a little bit more. Which I always prefer."	P9

In summary, the qualitative findings map closely to the evaluation framework. Participants' perceptions of greater detail and relevance in the experimental model reflect usefulness and supportiveness and clarity and coherence. Observations of accuracy align with the framework's accuracy dimension, while concerns about potential bias relate to safety and harm and trust and confidence. Comments on wordiness pertain to clarity and coherence, and reports of higher satisfaction and empathy connect to overall satisfaction and supportiveness. Overall, these insights demonstrate how participants' experiences substantiate the framework's multidimensional evaluation of LLM responses.

Implications

Reflections on Interview Design and Future LLM Applications

In addition to exploring participants' perspectives on the differences between responses from C1 and C2, we included two additional questions aimed at identifying limitations of the current interview design and eliciting participants' views on the potential of LLM-based tools to empower informal caregivers of persons with dementia. Specifically, we asked the following questions: (1) What general challenges did participants encounter during the interviews? and (2) What future opportunities do they envision for using LLMs in this area of research? The following sections present key implications derived from the interview findings.

Implication 1: Future Interview Design

Two main themes emerged regarding the limitations of the current interview design, as shown in Table 6. The first theme, response content, includes 5 subthemes. The second theme, testing scenario design, reflected concerns that the interview scenarios were not tailored specifically for dementia care. Therefore, interview results revealed that participants valued empathy and actionable advice but found responses overly verbose and lacking perspectives from social workers or community-based resources. Quantitative ratings confirmed that concise, emotionally attuned answers were associated with higher satisfaction, linking linguistic style to usability outcomes. These insights point to the need for future interviews to employ dementia-specific scenarios and include interdisciplinary expertise when shaping prompts and evaluation criteria. Incorporating diverse professional viewpoints and assessing dimensions such as emotional tone, cognitive load, and decision support can refine both data collection and model evaluation, guiding the design of more realistic and user-centered caregiving dialogues.

Table 6. Themes and illustrative quotes summarizing participant feedback on limitations of the current interview design.

Theme and subtheme	Supporting quotes	Participant ID
Response content limitation		
Wordy	"The volume of words needs to be cut down."	P8
Lacking social worker contact information	"So, what I thought lacked from both of them were. There was very little information regarding talk to a social worker, which is something that they both really missed out. On neither of them indicated that these people should get in contact with a social worker, and how to do that."	P2

Theme and subtheme	Supporting quotes	Participant ID
Lacking coping strategy	"And then none of them talked about. Really seeing your own provider for coping strategies, as well as oftentimes the caregivers become anxious or depressed, and so treatment for that, either by, you know, medicinally or with therapy."	P2
Lacking care team information	"Usually, people who have dementia have a care team made up of a primary care provider, a social worker, and potentially a pharmacist … So, the social worker and the primary care, or the neurologist were glaringly missing from almost every answer."	P11
Bias on guidance provided	"In the responses, there are suggestions such as 'this center' you can visit for help, or 'that organization' you can reach out for assistance, or 'that medical provider' we can contact, so, how could I trust this, how could I believe that there is no preference chosen by the system?"	P3
Test scenario design limitation		
Not specifically designed for dementia	"Taylor already assumes that it is true that they have dementia, or they have a cognitive decline. Did they tell the system (ChatGPT) already? They know that it was a cognitive decline which I'm assuming. That was the basis of the conversation with the system."	P5

Implication 2: LLM-Based Applications

The results highlight the potential for LLMs to provide personalized, empathetic, and context-aware support for informal caregivers. Building on these findings, future systems could integrate multimodal data from wearables and caregiver inputs to tailor recommendations dynamically—such as suggesting rest during stress or prompting social connection during isolation. Embedding such intelligence within existing caregiver platforms would facilitate coordination among families and professionals. To ensure responsible deployment, systems must incorporate transparent data practices, bias mitigation, and ethical oversight aligned with clinical and privacy standards. These principles can guide the evolution of LLM-driven caregiver tools from research prototypes to trustworthy, scalable digital companions.

Discussion

Principal Findings

The study addressed the overarching aim of exploring how LLMs such as ChatGPT can support informal dementia caregivers by systematically evaluating both their limitations and potential benefits. In relation to RQ1, our findings indicated that while baseline LLM responses were generally clear and reasonable, they often lacked sufficient relevance and contextual depth for caregiving situations, reflecting current challenges and risks in everyday use. Addressing RQ2 and RQ3, the experimental model—enhanced through prompt engineering with dementia-specific and culturally informed caregiving knowledge—significantly improved the actionability, relevance, and practical value of responses, leading to higher informal caregiver satisfaction. These results demonstrate that tailored prompt strategies can meaningfully enhance LLM usability and reliability in informal caregiving contexts. In connection to RQ4, the findings underscore the importance of integrating domain-specific knowledge, cultural sensitivity, and safety considerations into future AI tool design to better support informal caregiver–person with dementia dyads in real-world settings.

Interpretation and Contextualization of Findings

Overall, the findings suggest that both the baseline and experimental models produced responses that were clear and reasonable. However, the experimental model delivered more relevant and practical guidance for informal caregiving, offering detailed information aligned with the 32 test questions and actionable suggestions. As a result, participants reported greater satisfaction with the experimental model than with the baseline one. Additionally, our study found that the qualitative insights complemented the quantitative results, revealing nuanced user perceptions that provided a deeper understanding of how caregivers evaluated response clarity, relevance, and emotional tone. These perceptions further illuminate the emotional and cognitive contexts in which caregivers engage with AI-generated content, offering insight into the types of communication that foster trust and reassurance.

Caregiving in the context of cognitive impairment often involves significant emotional burden [71], uncertainty, and a need for reassurance [72,73]. Participants valued responses that provided accurate information with empathy and compassion, resembling communication within supportive caregiving networks. Consistent with prior work showing that emotionally attuned messaging enhances engagement and trust [74,75], our findings underscore that tone and empathy are integral to digital health communication. Nonetheless, recent studies caution that AI-generated empathy can appear inconsistent or superficial, potentially affecting user trust. Our results extend this discussion by demonstrating that even subtle tone variations in LLM responses can shape caregivers' perceptions of credibility and emotional support, emphasizing both the promise and responsibility of emotionally intelligent design.

Comparison With Prior Work

Recent research has increasingly explored LLMs as tools to support dementia caregiving. Parmanto et al [54], for instance, introduced CaLM, a retrieval-augmented system that delivers context-specific guidance to family caregivers of persons with dementia. Similarly, the DEMENTIA PLAN framework [76] integrates LLMs with knowledge graphs to combine practical care recommendations and emotional support for persons with mild dementia. Other work has used LLMs to analyze clinical documentation for assessing caregiver burden, showing improved accuracy over traditional text-analysis approaches [40]. However, findings across these studies remain mixed—while many report promise in enhancing personalization and empathy, others caution about biases, limited generalizability, and potential overreliance on model-generated information. Addressing such variability is essential for understanding how LLMs can be responsibly deployed in real caregiving contexts.

Our study extends this emerging literature by providing a quantitative, within-subject comparison between an experimental LLM-based caregiving support model and a baseline condition across multiple evaluation domains. Although we did not benchmark C2 directly against retrieval-augmented or knowledge-graph–based systems such as CaLM or DEMENTIA PLAN, our approach differs in focusing on interaction-level validation—empirically assessing caregivers' perceived actionability, relevance, and satisfaction rather than optimizing information retrieval or structured knowledge integration. This distinction underscores the model's emphasis on practical impact and caregiver-centered outcomes. By linking user feedback to specific caregiving needs, our findings highlight how conversational tuning and prompt design can translate advanced language capabilities into tangible, emotionally attuned support for informal dementia caregivers, while also delineating future pathways for integrating retrieval or knowledge-graph components in more comprehensive systems.

Prior studies indicate that LLMs can deliver emotionally attuned responses when guided by tailored prompting or fine-tuning. Wang et al [77] developed a fine-tuned GPT-2 trained on dementia caregiver therapy transcripts that emulated natural, emotionally resonant dialogue. Lee et al [78] found that multiple LLMs, including GPT-4, produced responses rated as more empathic than human-written ones in sensitive scenarios. A systematic review by Sorin et al [29] similarly confirmed that LLMs can demonstrate cognitive empathy toward users' emotional cues. Nonetheless, recent work cautions that empathy quality varies with training data, prompt design, and context. Our qualitative results extend this literature by showing that empathy, tone, and contextual relevance are perceived by caregivers as essential features of supportive AI communication in dementia care.

Implications for Future AI Tool Design

This study highlights the potential of LLMs, specifically ChatGPT-4o, to support informal caregivers of individuals in the early stages of dementia. By combining prompt engineering with a carefully curated dataset, we tailored AI-generated responses to better meet caregivers' informational and emotional needs—an area often overlooked in traditional digital health tools.

A key strength of this work lies in its interdisciplinary innovation. In the medical domain, our approach demonstrates how LLMs can be adapted to deliver context-sensitive support, offering caregivers timely guidance and emotional reassurance. This is particularly valuable in early-stage dementia care, where decision-making is complex and emotionally charged. LLMs can also serve as accessible knowledge hubs, helping caregivers navigate challenges and risks while identifying areas for improvement in AI-generated communication.

From an HCI perspective, our use of prompt engineering contributes to the development of more intuitive, adaptive, and inclusive interfaces. By aligning AI responses with user expectations and emotional tone, we enhance usability and trust, two critical factors in the adoption of health technologies. This interdisciplinary application bridges technical innovation with human-centered design, advancing both fields toward more intelligent and empathetic digital health systems.

While these contributions highlight the study's methodological and practical strengths, several limitations should also be acknowledged.

Limitations

First, the prompt generation methodology was manually designed, which may limit scalability and reproducibility. Future research should explore automated or data-driven prompt optimization techniques to improve consistency and efficiency. Pathways toward scalable prompt engineering include methods that move beyond manual prompt crafting toward data-driven and adaptive optimization. Automated prompt optimization can iteratively refine prompt structures using feedback signals or performance metrics, reducing reliance on human trial-and-error. Fine-tuning approaches, in which LLMs are adapted using domain-specific conversational data, can further align system responses with the contextual and emotional nuances of caregiving interactions. Additionally, retrieval-augmented generation techniques can dynamically incorporate up-to-date, evidence-based caregiving information from external databases or knowledge repositories. Together, these strategies enable more consistent, interpretable, and scalable LLM performance across diverse caregiving scenarios, supporting sustainable deployment in real-world digital health environments.

Second, the use of the public ChatGPT interface limited our ability to control key model parameters such as memory and maximum output length, which may affect consistency and reproducibility across sessions. This constraint also restricted opportunities for systematic tuning and integration within the experimental setup. In future work, transitioning to an application programming interface (API) or a self-hosted model will enable more precise control over system configurations, logging, and parameter settings. Such an approach would enhance reproducibility, support iterative optimization, and facilitate seamless integration into real-world caregiving platforms where reliability and personalization are critical.

Third, the sample size in our study (12 content expert participants) was small, which might raise concerns about statistical power for Mann-Whitney *U* tests. However, Caine [57] reviewed 5 years of CHI proceedings, tallied participant numbers, and found that 12 participants was the modal (most common) sample size, reflecting a community norm rather than an evidence-based standard. In addition, the expert sample was predominantly composed of nurses and physicians, with only a small number of informal caregivers represented. This professional weighting may limit the extent to which the evaluation captures the lived experiences, daily challenges, and nuanced perspectives of caregivers who provide continuous, hands-on support to individuals with cognitive impairment. Consequently, the findings may reflect a more clinical than experiential understanding of user needs.

Future research should aim to balance professional and informal perspectives by including a greater number of family caregivers, as well as allied care professionals such as social workers, occupational therapists, and community health aides. These participants can provide valuable insights into the social, emotional, and logistical dimensions of caregiving, thereby enriching system design and ensuring that interventions are grounded in both professional expertise and authentic caregiver experience. Additionally, except for ChatGPT, at least one more LLM model, such as Gemini, could be applied as another baseline for more robust comparative evaluation.

Taken together, these limitations highlight areas for methodological refinement, broader participant inclusion, and technical enhancement in future work. Addressing these considerations will strengthen the reliability, generalizability, and practical impact of LLM-based interventions for informal dementia caregivers, laying the groundwork for more robust, scalable, and caregiver-centered digital health solutions.

Conclusions

Future Directions for LLM-Supported Dementia Caregiving

This study demonstrates the potential of LLMs to support informal caregivers of individuals with dementia through emotionally attuned, context-specific communication. By applying prompt engineering within a medical HCI framework, we tailored AI responses to better address caregivers' informational and emotional needs. Findings from 12 expert participants provide preliminary insights into how such systems might enhance caregiver support, though these results should not be generalized beyond this initial evaluation.

Beyond this exploratory evaluation, the study offers a framework for integrating AI, HCI, and health care, highlighting how intelligent digital tools can be designed to meet the complex cognitive, emotional, and informational needs of informal caregivers. Rather than presenting a scalable solution, this work lays the groundwork for more rigorous, large-scale validation of empathetic, context-aware digital health interventions. Future research should examine real-world impact on caregiver knowledge, well-being, and decision-making, while leveraging technical advances—such as transitioning to the ChatGPT API—to overcome current limitations and enable deeper integration with diverse datasets.

Considering the broader implications of this work alongside insights from the literature [79-81], we propose 5 directions through which future design can enhance LLM-driven empowerment for informal caregivers and persons with dementia.

Direction 1: Considering Social Dynamics and Resources

Designing future LLM-based agents to support informal caregivers of individuals with early-stage dementia requires moving beyond a dyadic caregiving model and considering a broader network of care [82-84]. These agents should facilitate coordination among primary care physicians, neurologists, social workers, and local community organizations, recognizing that dementia care is inherently collaborative. Designing with this extended care team in mind can help ensure that the agent's outputs align with multidisciplinary care goals and reduce the informational fragmentation caregivers often face.

Direction 2: Advancing Actionability Through Biometrics

It is essential to incorporate diverse behavioral and perceptual data about the person with dementia to contextualize LLM-generated guidance. Informal caregivers make decisions based not only on static health records but also on evolving patterns, such as changes in mobility, sleep cycles, or daily stress levels. Future systems should be designed to integrate such multimodal signals, enabling LLMs to generate more personalized and situationally aware recommendations that align with real-time conditions.

Direction 3: Conversational Tone and Manner Matter

The tone and structure of LLM-generated content must be carefully refined to support caregivers' emotional and cognitive load. Informal caregivers, who often operate under high stress, may find verbose or abstract language overwhelming. LLMs should adopt a succinct, step-by-step, and conversational manner of delivery that facilitates comprehension and action. Designing for interactional clarity can help transform generic advice into practically actionable microinterventions.

Direction 4: Integrating Up-to-Date Dementia-Related Resources With Personal Context

LLM-based agents must be grounded not only in a person's medical history but also in the latest clinical and community-based dementia care practices. As dementia care evolves, outdated or generalized information can lead caregivers astray. Future systems should incorporate mechanisms for continuously updating their knowledge base with validated sources, ensuring that recommendations [85-89] remain aligned with current standards of care.

Direction 5: Privacy-Minded Conversation

Finally, as LLMs are positioned to support caregiving teams, privacy must be treated as a first-class design concern. Stakeholders may have different sensitivities regarding what information is shared or withheld—for instance, a caregiver may wish to conceal emotional burnout from a physician, or a person with dementia may prefer to limit what their adult children can access. Designing privacy-aware groupware requires mechanisms for nuanced control over information disclosure, aligned with the preferences and boundaries of each stakeholder.

Collectively, these directions underscore the broader implications of LLM-based interventions: when thoughtfully designed and ethically implemented, such systems can strengthen LLM-driven empowerment by enhancing informal caregivers' decision-making, emotional resilience, and access to reliable knowledge—ultimately fostering a more responsive, connected, and compassionate dementia care ecosystem.

Acknowledgments

We thank all study participants for sharing their experiences and perspectives on applying large language models to support informal caregivers of persons with dementia.

The authors declare the use of generative artificial intelligence (GAI) in the research and writing process. According to the GAIDeT taxonomy (2025) [90], the tasks of proofreading, editing, and research design were delegated to GAI tools under full human supervision. The GAI tools used were GPT-5 (for proofreading and editing) and ChatGPT-4o (for research design). Responsibility for the final manuscript lies entirely with the authors. GAI tools are not listed as authors and do not bear responsibility for the final outcomes.

In a methodological note, ChatGPT was used within a controlled, evaluation-based design to generate two types of responses: one from the standard ChatGPT-4o model and one from a prompt-engineered experimental version grounded in a health-science and gerontology framework. In total, 32 matched pairs (64 responses) were produced for expert evaluation. All the 64 responses are made available as Multimedia Appendix 9.

Funding

This work was supported in part by the National Science Foundation Future of WorkProgram (award 2310210). The funder had no involvement in the study design,data collection, analysis, interpretation of results, or the writing of the manuscript.

Authors' Contributions

HZ was responsible for designing the two experimental conditions, developing the survey instrument for the study, collecting data from medical experts during the study, conducting the data analysis after the study, drafting the initial paper, and contributing to its review and editing. ZZ provided technical guidance during the development of the two experimental conditions and reviewed and guided the editing of the paper draft. SRH served as a corresponding author, worked with HZ and KMO to define the research questions and study pipeline, provided direction for the development of the experimental condition, supported the preparation of study materials and user interface, coconducted the study, led the analysis of results, identified the key findings, authored the implications for design, and reviewed and guided the editing of the paper draft. KMO, in collaboration with SRH, contributed to framing the research pipeline, preparing preliminary data for prompt engineering, coordinating with local experts to collect frequently asked questions from informal caregivers, recruiting medical experts, analyzing the resulting data, and reviewing and guiding the editing of the paper draft. All authors had full access to this study's data, including multimedia appendices, and approved the final version of this paper for submission.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Study materials.

[PDF File (Adobe File), 730 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Dataset of 90 question-and-answer pairs.

[DOCX File (Microsoft Word File), 128 KB-Multimedia Appendix 2]

Multimedia Appendix 3

Prompt examples.

[PDF File (Adobe File), 403 KB-Multimedia Appendix 3]

Multimedia Appendix 4

Dataset of 32 test questions or scenarios.

[DOCX File (Microsoft Word File), 26 KB-Multimedia Appendix 4]

Multimedia Appendix 5

Interface for test question no. 18 with two types of responses.
[PDF File (Adobe File), 1245 KB-Multimedia Appendix 5]

Multimedia Appendix 6

Interview questions.
[DOCX File (Microsoft Word File), 39 KB-Multimedia Appendix 6]

Multimedia Appendix 7

Table reports on the Mann-Whitney *U* test results.
[PDF File (Adobe File), 530 KB-Multimedia Appendix 7]

Multimedia Appendix 8

Table reports on the comparisons in terms of mean and SD values for each evaluation criterion under two conditions.
[DOCX File (Microsoft Word File), 22 KB-Multimedia Appendix 8]

Multimedia Appendix 9

Study materials appendix: 32 testing scenarios and the corresponding 32 pairs of large language model–based responses.
[PDF File (Adobe File), 811 KB-Multimedia Appendix 9]

References

1. Alzheimer's Association. 2024 Alzheimer's disease facts and figures. Alzheimers Dement. May 2024;20(5):3708-3821. [doi: 10.1002/alz.13809] [Medline: 38689398]
2. Kim HJ, Kehoe P, Gibbs LM, Lee JA. Caregiving experience of dementia among Korean American family caregivers. Issues Ment Health Nurs. Feb 2019;40(2):158-165. [doi: 10.1080/01612840.2018.1534909] [Medline: 30620625]
3. Meyer OL, Nguyen KH, Dao TN, Vu P, Arean P, Hinton L. The sociocultural context of caregiving experiences for Vietnamese dementia family caregivers. Asian Am J Psychol. Sep 2015;6(3):263-272. [doi: 10.1037/aap0000024] [Medline: 26617956]
4. Semere W, Nápoles AM, Gregorich S, Livaudais-Toman J, Karliner L. Caregiving for older adults with limited English proficiency: transitioning from hospital to home. J Gen Intern Med. Sep 2019;34(9):1744-1750. [doi: 10.1007/s11606-019-05119-y] [Medline: 31236893]
5. Skaria AP. The economic and societal burden of Alzheimer disease: managed care considerations. Am J Manag Care. Sep 2022;28(10 Suppl):S188-S196. [doi: 10.37765/ajmc.2022.89236] [Medline: 36197132]
6. Ho HT, Jia R, Habibi N, et al. Experiences of informal caregivers of people with dementia in low- and middle-income countries: a systematic review protocol. JBI Evid Synth. Jul 1, 2024;22(7):1362-1370. [doi: 10.11124/JBIES-23-00470] [Medline: 38477072]
7. Van den Bossche P, Schoenmakers B. The impact of dementia's affiliate stigma on the mental health of relatives: a cross section survey. Front Psychol. 2021;12:789105. [doi: 10.3389/fpsyg.2021.789105] [Medline: 35126240]
8. Zwaanswijk M, Peeters JM, van Beek AP, Meerveld JH, Francke AL. Informal caregivers of people with dementia: problems, needs and support in the initial stage and in subsequent stages of dementia: a questionnaire survey. Open Nurs J. 2013;7:6-13. [doi: 10.2174/1874434601307010006] [Medline: 23346266]
9. Gris F, D'Amen B, Lamura G, Paciaroni L, Socci M, Melchiorre MG. Personalized technological support for informal caregivers of older people with dementia: a co-design approach involving potential end users and healthcare professionals in three focus groups in Italy. Healthcare (Basel). Sep 28, 2023;11(19):2640. [doi: 10.3390/healthcare11192640] [Medline: 37830677]
10. DuBose L, Fan Q, Fisher L, et al. Development and evaluation of a web-based platform for personalized educational and professional assistance for dementia caregivers: proposal for a mixed methods study. JMIR Res Protoc. Aug 7, 2024;13:e64127. [doi: 10.2196/64127] [Medline: 39110962]
11. Waymouth M, Siconolfi D, Friedman EM, Saliba D, Ahluwalia SC, Shih RA. Barriers and facilitators to home- and community-based services access for persons with dementia and their caregivers. J Gerontol B Psychol Sci Soc Sci. May 26, 2023;78(6):1085-1097. [doi: 10.1093/geronb/gbad039] [Medline: 36896936]
12. Wimo A, Seeher K, Cataldi R, et al. The worldwide costs of dementia in 2019. Alzheimers Dement. Jul 2023;19(7):2865-2873. [doi: 10.1002/alz.12901] [Medline: 36617519]
13. Martis CS, Bhandary RP, Chandrababu R, et al. Caring burden and quality of life among the caregivers of people living with dementia – a cross-sectional study in Udupi district of Karnataka. Home Health Care Serv Q. 2024;43(3):191-204. [doi: 10.1080/01621424.2023.2301417] [Medline: 38190733]

14. Bentwich ME, Dickman N, Oberman A. Autonomy and dignity of patients with dementia: perceptions of multicultural caretakers. Nurs Ethics. Feb 2018;25(1):37-53. [doi: 10.1177/0969733016642625] [Medline: 27113259]
15. Boumans J, van Boekel LC, Baan CA, Luijkx KG. How can autonomy be maintained and informal care improved for people with dementia living in residential care facilities: a systematic literature review. Gerontologist. Nov 16, 2019;59(6):e709-e730. [doi: 10.1093/geront/gny096] [Medline: 30239712]
16. Bayly M, Morgan D, Kosteniuk J, et al. Protocol for a systematic review on interventions for caregivers of persons with mild cognitive impairment and early dementia: does early stage intervention improve caregiver well-being and ability to provide care? BMJ Open. May 10, 2019;9(5):e028441. [doi: 10.1136/bmjopen-2018-028441] [Medline: 31079088]
17. Signe A, Elmståhl S. Psychosocial intervention for family caregivers of people with dementia reduces caregiver's burden: development and effect after 6 and 12 months. Scand J Caring Sci. Mar 2008;22(1):98-109. [doi: 10.1111/j.1471-6712.2007.00498.x] [Medline: 18269429]
18. Frias CE, Cabrera E, Zabalegui A. Informal caregivers' roles in dementia: the impact on their quality of life. Life (Basel). Oct 23, 2020;10(11):251. [doi: 10.3390/life10110251] [Medline: 33113995]
19. Roland KP, Chappell NL. Meaningful activity for persons with dementia: family caregiver perspectives. Am J Alzheimers Dis Other Demen. Sep 2015;30(6):559-568. [doi: 10.1177/1533317515576389] [Medline: 25788432]
20. Teichmann B, Gkioka M, Kruse A, Tsolaki M. Informal caregivers' attitude toward dementia: the impact of dementia knowledge, confidence in dementia care, and the behavioral and psychological symptoms of the person with dementia. a cross-sectional study. J Alzheimers Dis. 2022;88(3):971-984. [doi: 10.3233/JAD-215731] [Medline: 35723101]
21. Balbim GM, Marques IG, Cortez C, Magallanes M, Rocha J, Marquez DX. Coping strategies utilized by middle-aged and older Latino caregivers of loved ones with alzheimer's disease and related dementia. J Cross Cult Gerontol. Dec 2019;34(4):355-371. [doi: 10.1007/s10823-019-09390-8] [Medline: 31705279]
22. Caceres BA, Perez A. Implications of the CARE Act for Latino caregivers. J Gerontol Nurs. Mar 1, 2018;44(3):9-14. [doi: 10.3928/00989134-20180213-04] [Medline: 29470586]
23. Rote S, Angel J, Hinton L. Characteristics and consequences of family support in Latino dementia care. J Cross Cult Gerontol. Dec 2019;34(4):337-354. [doi: 10.1007/s10823-019-09378-4] [Medline: 31346891]
24. Busch F, Hoffmann L, Rueger C, et al. Current applications and challenges in large language models for patient care: a systematic review. Commun Med (Lond). Jan 21, 2025;5(1):26. [doi: 10.1038/s43856-024-00717-2] [Medline: 39838160]
25. Lin C, Kuo CF. Roles and potential of large language models in healthcare: a comprehensive review. Biomed J. Oct 2025;48(5):100868. [doi: 10.1016/j.bj.2025.100868] [Medline: 40311872]
26. Huang L, Yu W, Ma W, et al. A survey on hallucination in large language models: principles, taxonomy, challenges, and open questions. ACM Trans Inf Syst. Mar 31, 2025;43(2):1-55. [doi: 10.1145/3703155]
27. Kim S, Han DY, Bae J. Transforming Alzheimer's digital caregiving through large language models. Curr Alzheimer Res. 2024;21(7):503-516. [doi: 10.2174/0115672050301740241118044604] [Medline: 39592896]
28. Kim T, Kim H, Roberts AC, Jacobs M, Kay M. Opportunities in mental health support for informal dementia caregivers suffering from verbal agitation. Proc ACM Hum-Comput Interact. Apr 17, 2024;8(CSCW1):1-26. [doi: 10.1145/3637381]
29. Sorin V, Brin D, Barash Y, et al. Large language models and empathy: systematic review. J Med Internet Res. Dec 11, 2024;26:e52597. [doi: 10.2196/52597] [Medline: 39661968]
30. Tierney AA, Reed ME, Grant RW, Doo FX, Payán DD, Liu VX. Health equity in the era of large language models. Am J Manag Care. Mar 1, 2025;31(3):112-117. [doi: 10.37765/ajmc.2025.89695]
31. Croxford E, Gao Y, Pellegrino N, et al. Current and future state of evaluation of large language models for medical summarization tasks. Npj Health Syst. 2025;2:6. [doi: 10.1038/s44401-024-00011-2] [Medline: 40124388]
32. Anbumani S, Ahunbay E. Toward the clinically effective evaluation of artificial intelligence-generated responses. JCO Clin Cancer Inform. Dec 2024;8:e2400258. [doi: 10.1200/CCI-24-00258] [Medline: 39661915]
33. Alanzi TM. Impact of ChatGPT on teleconsultants in healthcare: perceptions of healthcare experts in Saudi Arabia. J Multidiscip Healthc. 2023;16:2309-2321. [doi: 10.2147/JMDH.S419847] [Medline: 37601325]
34. Baglivo F, De Angelis L, Casigliani V, Arzilli G, Privitera GP, Rizzo C. Exploring the possible use of AI chatbots in public health education: feasibility study. JMIR Med Educ. Nov 1, 2023;9:e51421. [doi: 10.2196/51421] [Medline: 37910155]
35. Wilhelm TI, Roos J, Kaczmarczyk R. Large language models for therapy recommendations across 3 clinical specialties: comparative study. J Med Internet Res. Oct 30, 2023;25:e49324. [doi: 10.2196/49324] [Medline: 37902826]
36. Morita PP, Abhari S, Kaur J, Lotto M, Miranda P, Oetomo A. Applying ChatGPT in public health: a SWOT and PESTLE analysis. Front Public Health. 2023;11:1225861. [doi: 10.3389/fpubh.2023.1225861] [Medline: 37465170]

37. Pérez-Esteve C, Guilabert M, Matarredona V, et al. AI in home care-evaluation of large language models for future training of informal caregivers: observational comparative case study. J Med Internet Res. Apr 28, 2025;27:e70703. [doi: 10.2196/70703] [Medline: 40294407]
38. Shi JM, Yoo DW, Wang K, Rodriguez VJ, Karkar R, Saha K. Mapping caregiver needs to AI chatbot design: strengths and gaps in mental health support for Alzheimer's and dementia caregivers. arXiv. Preprint posted online on Jun 18, 2025. [doi: 10.48550/arXiv.2506.15047]
39. Hasan WU, Zaman KT, Wang X, Li J, Xie B, Tao C. Empowering Alzheimer's caregivers with conversational AI: a novel approach for enhanced communication and personalized support. npj Biomed Innov. 2024;1(1):1-9. [doi: 10.1038/s44385-024-00004-8]
40. Chien SC, Yen CM, Chang YH, et al. Using large language model (LLM) to identify high-burden informal caregivers in long-term care. Comput Methods Programs Biomed. Oct 2024;255:108329. [doi: 10.1016/j.cmpb.2024.108329] [Medline: 39029418]
41. Knight BG, Sayegh P. Cultural values and caregiving: the updated sociocultural stress and coping model. J Gerontol B Psychol Sci Soc Sci. Jan 2010;65B(1):5-13. [doi: 10.1093/geronb/gbp096] [Medline: 19934166]
42. Khan TS, Hirschman KB, McHugh MD, Naylor MD. Self-efficacy of family caregivers of older adults with cognitive impairment: a concept analysis. Nurs Forum. Jan 2021;56(1):112-126. [doi: 10.1111/nuf.12499] [Medline: 32888197]
43. Alzheimer's Association. URL: https://www.alz.org/ [Accessed 2026-2-10]
44. National Institute on Aging. URL: https://www.nia.nih.gov/ [Accessed 2026-2-10]
45. RRF Foundation for Aging. URL: https://www.rrf.org/ [Accessed 2026-02-10]
46. Functions and roles. National Medical Center. URL: https://www.nmc.or.kr/nmc/contents/nid_roles [Accessed 2026-02-10]
47. Our commitment to diversity, equity and inclusion. Alzheimer's Association. URL: https://www.alz.org/about/our-commitment-to-diversity-and-inclusion [Accessed 2026-02-10]
48. Support groups. Alzheimer's Association. URL: https://www.alz.org/help-support/community/support-groups [Accessed 2026-02-10]
49. Carelike – provider search. Community Resource Finder. URL: https://www.communityresourcefinder.org/ [Accessed 2026-02-10]
50. National Capital Area Chapter. Alzheimer's Association. URL: https://www.alz.org/nca [Accessed 2026-02-10]
51. What is Alzheimer's? Alzheimer's Association. URL: https://www.alz.org/alzheimers-dementia/what-is-alzheimers [Accessed 2026-02-10]
52. Sharing your diagnosis. Alzheimer's Association. URL: https://www.alz.org/help-support/i-have-alz/know-what-to-expect/sharing-your-diagnosis [Accessed 2026-02-10]
53. Li Z, Xie B, Hilsabeck R, et al. Effects of different prompts on the quality of GPT-4 responses to dementia care questions. In: 2024 IEEE 12th International Conference on Healthcare Informatics (ICHI). IEEE; 2024:412-417. [doi: 10.1109/ICHI61247.2024.00059]
54. Parmanto B, Aryoyudanta B, Soekinto TW, et al. A reliable and accessible caregiving language model (CaLM) to support tools for caregivers: development and evaluation study. JMIR Form Res. Jul 31, 2024;8:e54633. [doi: 10.2196/54633] [Medline: 39083337]
55. Cheng ST, Ng PHF. The PDC30 chatbot—development of a psychoeducational resource on dementia caregiving among family caregivers: mixed methods acceptability study. JMIR Aging. Jan 6, 2025;8:e63715. [doi: 10.2196/63715] [Medline: 39761547]
56. Ekin S. Prompt engineering for ChatGPT: a quick guide to techniques, tips, and best practices. TechRxiv. Preprint posted online on May 4, 2023. [doi: 10.36227/techrxiv.22683919.v2]
57. Caine K. Local standards for sample size at CHI. In: CHI '16: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery; 2016:981-992. [doi: 10.1145/2858036.2858498]
58. Tam TYC, Sivarajkumar S, Kapoor S, et al. A framework for human evaluation of large language models in healthcare derived from literature review. NPJ Digit Med. Sep 28, 2024;7(1):258. [doi: 10.1038/s41746-024-01258-7] [Medline: 39333376]
59. Yun JY, Kim DJ, Lee N, Kim EK. A comprehensive evaluation of ChatGPT consultation quality for augmentation mammoplasty: a comparative analysis between plastic surgeons and laypersons. Int J Med Inform. Nov 2023;179:105219. [doi: 10.1016/j.ijmedinf.2023.105219] [Medline: 37776670]
60. Rogasch JMM, Metzger G, Preisler M, et al. ChatGPT: can you prepare my patients for [^{18}F]FDG PET/CT and explain my reports? J Nucl Med. Dec 1, 2023;64(12):1876-1879. [doi: 10.2967/jnumed.123.266114] [Medline: 37709536]
61. Bazzari FH, Bazzari AH. Utilizing ChatGPT in telepharmacy. Cureus. Jan 2024;16(1):e52365. [doi: 10.7759/cureus.52365] [Medline: 38230387]

62. Khlaif ZN, Mousa A, Hattab MK, et al. The potential and concerns of using AI in scientific research: ChatGPT performance evaluation. JMIR Med Educ. Sep 14, 2023;9:e47049. [doi: 10.2196/47049] [Medline: 37707884]
63. Gilson A, Safranek CW, Huang T, 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. Feb 8, 2023;9:e45312. [doi: 10.2196/45312] [Medline: 36753318]
64. Lechien JR, Georgescu BM, Hans S, Chiesa-Estomba CM. ChatGPT performance in laryngology and head and neck surgery: a clinical case-series. Eur Arch Otorhinolaryngol. Jan 2024;281(1):319-333. [doi: 10.1007/s00405-023-08282-5]
65. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. Nature New Biol. Aug 2023;620(7972):172-180. [doi: 10.1038/s41586-023-06291-2] [Medline: 37438534]
66. Cankurtaran RE, Polat YH, Aydemir NG, Umay E, Yurekli OT. Reliability and usefulness of ChatGPT for inflammatory bowel diseases: an analysis for patients and healthcare professionals. Cureus. Oct 2023;15(10):e46736. [doi: 10.7759/cureus.46736] [Medline: 38022227]
67. An Y, Fang Q, Wang L. Enhancing patient education in cancer care: intelligent cancer patient education model for effective communication. Comput Biol Med. Feb 2024;169:107874. [doi: 10.1016/j.compbiomed.2023.107874] [Medline: 38157775]
68. Deiana G, Dettori M, Arghittu A, Azara A, Gabutti G, Castiglia P. Artificial intelligence and public health: evaluating ChatGPT responses to vaccination myths and misconceptions. Vaccines (Basel). Jul 7, 2023;11(7):1217. [doi: 10.3390/vaccines11071217] [Medline: 37515033]
69. Saldana JM. The Coding Manual for Qualitative Researchers. 3rd ed. SAGE Publications; 2025. ISBN: 9781473902497
70. Layder D. Sociological Practice. SAGE Publications Ltd; 1998. [doi: 10.4135/9781849209946] ISBN: 0761954295
71. Corrêa L, Ottaviani AC, Bregola AG, de Oliveira NA, Bento SR, Pavarini SCI. Cognitive performance, burden and stress in aged caregivers of older adults with and without cognitive impairment. Dement Neuropsychol. 2023;17:e20220073. [doi: 10.1590/1980-5764-DN-2022-0073] [Medline: 37396717]
72. Backhouse T, Jeon YH, Killett A, Green J, Khondoker M, Mioshi E. Nurturing attentiveness: a naturalistic observation study of personal care interactions between people with advanced dementia and their caregivers. Gerontologist. Jun 1, 2024;64(6):gnae004. [doi: 10.1093/geront/gnae004] [Medline: 38267689]
73. Hedman R, Hillerås P, Tyrrell M. Living in uncertainty while a spouse is undergoing a cognitive assessment: voices of women care partners. Dementia (London). Nov 2022;21(8):2631-2646. [doi: 10.1177/14713012221128448] [Medline: 36131252]
74. Luetke Lanfer H, Reifegerste D, Weber W, et al. Digital clinical empathy in a live chat: multiple findings from a formative qualitative study and usability tests. BMC Health Serv Res. Mar 8, 2024;24(1):314. [doi: 10.1186/s12913-024-10785-8] [Medline: 38459522]
75. Abou Hashish EA. Compassion through technology: digital empathy concept analysis and implications in nursing. Digit Health. 2025;11. [doi: 10.1177/20552076251326221] [Medline: 40093701]
76. Song Y, Lyu C, Zhang P, Brunswicker S, Dutt N, Rahmani A. DEMENTIA-PLAN: an agent-based framework for multi-knowledge graph retrieval-augmented generation in DEMENTIA care. arXiv. Preprint posted online on Mar 26, 2025. [doi: 10.48550/arXiv.2503.20950]
77. Wang L, Mujib MI, Williams J, Demiris G, Huh-Yoo J. An evaluation of generative pre-training model-based therapy chatbot for caregivers. arXiv. Preprint posted online on Jul 28, 2021. [doi: 10.48550/arXiv.2107.13115]
78. Lee YK, Suh J, Zhan H, Li JJ, Ong DC. Large language models produce responses perceived to be empathic. arXiv. Preprint posted online on Mar 26, 2024. [doi: 10.48550/arXiv.2403.18148]
79. Yan C, Chung JJY, Kiheon Y, Gingold Y, Adar E, Hong SR. FlatMagic: improving flat colorization through AI-driven design for digital comic professionals. In: CHI '22: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery; 2022:1-17. [doi: 10.1145/3491102.3502075]
80. Kwon N, Sun TS, Gao Y, et al. 3DPFIX: improving remote novices' 3D printing troubleshooting through human-AI collaboration design. Proc ACM Hum-Comput Interact. Apr 17, 2024;8(CSCW1):1-33. [doi: 10.1145/3637288]
81. Ganguly A, Yan C, Chung JJY, et al. ShadowMagic: designing human-AI collaborative support for comic professionals' shadowing. In: UIST '24: Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology. Association for Computing Machinery; 2024:1-15. [doi: 10.1145/3654777.3676332]
82. Hong S, Suh M, Kim TS, et al. Design for collaborative information-seeking: understanding user challenges and deploying collaborative dynamic queries. Proc ACM Hum-Comput Interact. 2019;3(CSCW):1-24. [doi: 10.1145/3359208]

83. Ara Z, Ganguly A, Peppard D, et al. Collaborative job seeking for people with autism: challenges and design opportunities. In: CHI '24: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery; 2024:1-17. [doi: 10.1145/3613904.3642197] [Medline: 40538597]

84. Hong S, Suh M, Henry Riche N, Lee J, Kim J, Zachry M. Collaborative dynamic queries: supporting distributed small group decision-making. In: CHI '18: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery; 2018:1-12. [doi: 10.1145/3173574.3173640]

85. Hong SR, Hullman J, Bertini E. Human factors in model interpretability: industry practices, challenges, and needs. Proc ACM Hum-Comput Interact. May 28, 2020;4(CSCW1):1-26. [doi: 10.1145/3392878]

86. Lee JH, Hong S, Cho H, Kim YS. VIZMO game browser: accessing video games by visual style and mood. In: CHI '15: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems. Association for Computing Machinery; 2015:149-152. [doi: 10.1145/2702123.2702264]

87. Gao Y, Gu S, Jiang J, Hong SR, Yu D, Zhao L. Going beyond XAI: a systematic survey for explanation-guided learning. ACM Comput Surv. Jul 31, 2024;56(7):1-39. [doi: 10.1145/3644073]

88. Choi M, Park C, Yang S, Kim Y, Choo J, Hong SR. AILA: attentive interactive labeling assistant for document classification through attention-based deep neural networks. In: CHI '19: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery; 2019:1-12. [doi: 10.1145/3290605.3300460]

89. Gao Y, Sun TS, Zhao L, Hong SR. Aligning eyes between humans and deep neural network through interactive attention alignment. Proc ACM Hum-Comput Interact. Nov 7, 2022;6(CSCW2):1-28. [doi: 10.1145/3555590]

90. Suchikova Y, Tsybuliak N, Teixeira da Silva JA, Nazarovets S. GAIDeT (Generative AI Delegation Taxonomy): a taxonomy for humans to delegate tasks to generative artificial intelligence in scientific research and publishing. Account Res. Aug 8, 2025:1-27. [doi: 10.1080/08989621.2025.2544331] [Medline: 40781729]

Abbreviations

AI: artificial intelligence
API: application programming interface
C1: baseline condition
C2: experimental condition
COPE: Committee on Publication Ethics
EQ: evaluation criterion
HCI: human-computer interaction
LLM: large language model
RQ: research question

Edited by Amaryllis Mavragani, Javad Sarvestan; peer-reviewed by Balu Bhasuran, Sadhasivam Mohanadas; submitted 01.Jul.2025; final revised version received 19.Nov.2025; accepted 24.Nov.2025; published 05.Mar.2026

Please cite as:
Zhou H, Zhu Z, Oh KM, Hong SR
Empowering Informal Caregivers of Persons With Early-Stage Dementia by Large Language Models: Mixed Methods Evaluation
JMIR Form Res 2026;10:e79975
URL: https://formative.jmir.org/2026/1/e79975
doi: 10.2196/79975