

Original Paper

Assessing the Role of the Generative Pretrained Transformer (GPT) in Alzheimer's Disease Management: Comparative Study of Neurologist- and Artificial Intelligence–Generated Responses

Jiaqi Zeng^{1*}, BMed; Xiaoyi Zou^{2,3*}, MD; Shirong Li⁴, MD; Yao Tang³, MMed; Sisi Teng³, BMed; Huanhuan Li⁵, MSN; Changyu Wang⁶, BSc; Yuxuan Wu⁷, MMed; Luyao Zhang⁸, BSN; Yunheng Zhong⁸, BSN; Jialin Liu^{7,9}, MD; Siru Liu¹⁰, PhD

¹West China Medical School, Sichuan University, Chengdu, China

²Department of Neurology, West China Hospital, Sichuan University, Chengdu, China

³Department of Neurology, Chengdu Shangjin Nanfu Hospital, Chengdu, China

⁴Department of Neurology, Guizhou Provincial People's Hospital, Guiyang, China

⁵Mental Health Center, West China Hospital, Sichuan University, Chengdu, China

⁶West China College of Stomatology, Sichuan University, Chengdu, China

⁷Department of Medical Informatics, West China Medical School, Chengdu, China

⁸West China School of Nursing, Sichuan University, Chengdu, China

⁹Department of Otolaryngology-Head and Neck Surgery, West China Hospital, Sichuan University, Chengdu, China

¹⁰Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, United States

*these authors contributed equally

Corresponding Author:

Jialin Liu, MD

Department of Medical Informatics

West China Medical School

No 37 Guoxue Road

Chengdu, 610041

China

Phone: 86 28 85422306

Email: DLjl8@163.com

Abstract

Background: Alzheimer's disease (AD) is a progressive neurodegenerative disorder posing challenges to patients, caregivers, and society. Accessible and accurate information is crucial for effective AD management.

Objective: This study aimed to evaluate the accuracy, comprehensibility, clarity, and usefulness of the Generative Pretrained Transformer's (GPT) answers concerning the management and caregiving of patients with AD.

Methods: In total, 14 questions related to the prevention, treatment, and care of AD were identified and posed to GPT-3.5 and GPT-4 in Chinese and English, respectively, and 4 respondent neurologists were asked to answer them. We generated 8 sets of responses (total 112) and randomly coded them in answer sheets. Next, 5 evaluator neurologists and 5 family members of patients were asked to rate the 112 responses using separate 5-point Likert scales. We evaluated the quality of the responses using a set of 8 questions rated on a 5-point Likert scale. To gauge comprehensibility and participant satisfaction, we included 3 questions dedicated to each aspect within the same set of 8 questions.

Results: As of April 10, 2023, the 5 evaluator neurologists and 5 family members of patients with AD rated the 112 responses: GPT-3.5: n=28, 25%, responses; GPT-4: n=28, 25%, responses; respondent neurologists: 56 (50%) responses. The top 5 (4.5%) responses rated by evaluator neurologists had 4 (80%) GPT (GPT-3.5+GPT-4) responses and 1 (20%) respondent neurologist's response. For the top 5 (4.5%) responses rated by patients' family members, all but the third response were GPT responses. Based on the evaluation by neurologists, the neurologist-generated responses achieved a mean score of 3.9 (SD 0.7), while the GPT-generated responses scored significantly higher (mean 4.4, SD 0.6; $P<.001$). Language and model analyses revealed no significant differences in response quality between the GPT-3.5 and GPT-4 models (GPT-3.5: mean 4.3, SD 0.7; GPT-4: mean 4.4, SD 0.5; $P=.51$). However, English responses outperformed Chinese responses in terms of comprehensibility (Chinese

responses: mean 4.1, SD 0.7; English responses: mean 4.6, SD 0.5; $P=.005$) and participant satisfaction (Chinese responses: mean 4.2, SD 0.8; English responses: mean 4.5, SD 0.5; $P=.04$). According to the evaluator neurologists' review, Chinese responses had a mean score of 4.4 (SD 0.6), whereas English responses had a mean score of 4.5 (SD 0.5; $P=.002$). As for the family members of patients with AD, no significant differences were observed between GPT and neurologists, GPT-3.5 and GPT-4, or Chinese and English responses.

Conclusions: GPT can provide patient education materials on AD for patients, their families and caregivers, nurses, and neurologists. This capability can contribute to the effective health care management of patients with AD, leading to enhanced patient outcomes.

(*J Med Internet Res* 2024;26:e51095) doi: [10.2196/51095](https://doi.org/10.2196/51095)

KEYWORDS

Alzheimer's disease; artificial intelligence; AI; large language model; LLM; Generative Pretrained Transformer; GPT; ChatGPT; patient information

Introduction

Alzheimer's disease (AD) is a progressive neurodegenerative disorder that has emerged as a significant health challenge worldwide. AD International reports indicate that the US population aged 65 years and older is predicted to surge from 58 million in 2021 to 88 million by 2050 [1,2]. As the most common form of dementia, AD accounts for 60%-80% of cases and significantly affects patients, caregivers, and society [3-5]. Given the disease's progressive nature, efficient management of AD and related cognitive decline is crucial for enhancing the life quality of patients and caregivers [6].

Optimally managing AD often requires a multifaceted approach to address the cognitive, functional, and behavioral symptoms associated with the disorder [7]. Current pharmacological interventions, such as cholinesterase inhibitors and *N*-methyl-D-aspartate receptor antagonists, deliver moderate symptomatic relief but fail to halt disease progression [8-10]. Meanwhile, nonpharmacological interventions, such as cognitive stimulation, physical exercise, and caregiver support programs, have shown potential benefits for patients with AD [11-13]. Patients diagnosed with AD and their caregivers frequently grapple with unmet needs and inadequate awareness about managing the disease and its complications. A lack of health literacy among patients with AD and caregivers has been reported [14,15]. The internet can offer valuable health information, but the complex nature of the primary literature and potential misinformation can often lead to more confusion than clarity [16]. Studies show that online resources provided by dementia-focused organizations are often lengthy and complex, demonstrating a scarcity of easily understandable information for this demographic [17].

Improving health literacy and empowering patients and caregivers about AD require the provision of precise, accessible, and holistic resources that cover both pharmacological and nonpharmacological interventions. Customizing these resources to meet the unique needs of individuals and caregivers at different disease stages is vital, given potential variations in management strategies based on AD severity and progression [18,19]. The increasing prevalence of AD poses a significant challenge for patients, caregivers, and society at large. Effective disease management is paramount to enhance the life quality of those affected. By boosting health literacy and providing

accessible, precise, and comprehensive AD management and treatment information, we can empower patients and caregivers, which will ultimately lead to better outcomes for those living with the disease [20].

The Generative Pretrained Transformer (GPT), developed by OpenAI, is an advanced natural language processing (NLP) model based on the GPT-3.5 architecture. It has been refined through supervised learning, human feedback, and reinforcement learning techniques [21]. Released on November 30, 2022, GPT has demonstrated potential in various medical applications, such as answering United States Medical Licensing Examination (USMLE) questions [22], generating simplified radiology reports for patients [23], optimizing clinical decision support [24], and several other clinical applications [25]. However, concerns exist regarding GPT's comprehension of queries and its restricted capacity to deliver detailed answers [26]. To date, there has been no research evaluating the accuracy and comprehensiveness of GPT in answering specific questions related to AD. Therefore, we collected the perception of participants (evaluator neurologists/family members) regarding the accuracy, comprehensibility, clarity, and usefulness of GPT's responses to frequently asked questions related to the management and caregiving of patients with AD.

The aims of this study were:

- To evaluate the potential of GPT's capability to respond to questions related to AD management
- To compare and assess the difference between GPT's responses to AD's care queries and those provided by neurologists

Methods

Participant Selection

The recruitment of neurologists was carried out through a purposeful sampling approach, with careful consideration given to ensuring the representativeness of the selected neurologists. Factors considered included their affiliation with different hospitals, years of clinical experience, medical specialties, and age. The selection criteria for neurologists were as follows: (1) ≥ 5 years of clinical experience in neurology, (2) a medical doctor degree, (3) grade III hospitals (the highest level of

hospitals in China), and (4) voluntary willingness to participate in this study.

In the selection of family members to participate in the study, our aim was to ensure diversity and representativeness by including individuals of varying ages, sexes, cultural backgrounds, and educational levels who have experience in caring for a person with AD daily. The criteria for selecting these participants were as follows: (1) 18-55 years old, (2) living with or providing daily care of a person with AD for a minimum of 1 year, (3) proficiency in the language skills necessary to comprehend and evaluate the study's content, (4) primary school education or higher, and (5) voluntary willingness to participate in this study.

Questionnaire Development

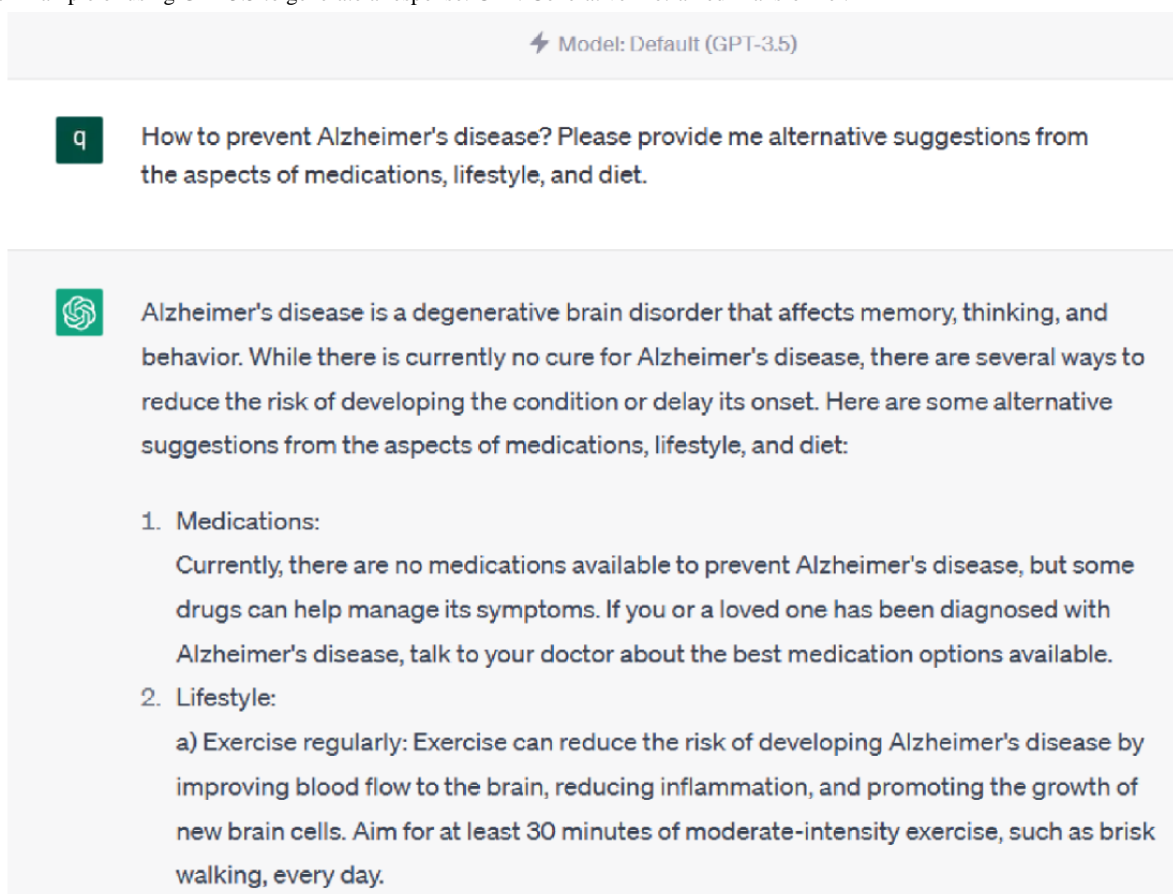
In our initial phase, we systematically compiled frequently encountered inquiries pertaining to AD from authoritative publications issued by esteemed professional associations and organizations [27-35]. A study showed that most of the content of the educational materials provided by these associations is unanimously recognized, with a mean overall score of 32.33 (SD 4.66) for the applicability of their educational content out of a total score of 44 [36]. This indicates that the content is well suited for health education purposes. To foster inclusivity and encompass a broad spectrum of patients and caregivers, we expanded the initial question pool. We meticulously excluded queries that were redundant, had ambiguous connotations, displayed subjectivity susceptible to individual interpretation, and underwent precise linguistic and grammatical refinement to enhance precision. Subsequently, after 3 rounds of expert

discussion, we definitively curated a comprehensive set of 14 questions related to AD. These questions comprehensively address the multifaceted concerns of patients' families and caregivers, encompassing vital subjects, such as general knowledge regarding AD, associated risks, diagnostic processes, treatment modalities, caregiving strategies, prognostic outlook, rehabilitation measures, and nursing practices (Multimedia Appendix 1). All questions used in this study are listed in Table S1 in Multimedia Appendix 2.

GPT and Response Generation

GPT (GPT-3.5, or ChatGPT) leverages reinforcement learning from human feedback or reinforcement learning from human preference approaches to generate coherent and contextually appropriate answers. Building on the foundation of the GPT-3.5 large language model, this research preview prototype demonstrates impressive capabilities in understanding and generating human-like text [36]. It underwent pretraining on an extensive data set comprising text extracted from online sources, such as websites, books, and papers up until 2021.

Questions were input into the GPT May 3 version, using the GPT-3.5 and GPT-4 models. Each question was treated as an individual prompt using the "New Chat" function. Additionally, to explore the variations in GPT's responses, each question was translated into Chinese. Figure 1 shows an example of using GPT-3.5 to generate responses. We also invited 4 respondent neurologists to answer each question to compare the quality of GPT and respondent neurologists' responses. All responses are listed in Table S2 in Multimedia Appendix 2.

Figure 1. Example of using GPT-3.5 to generate a response. GPT: Generative Pretrained Transformer.

Evaluation Response

To minimize bias between the evaluators, comprising neurologists and patients' family members, in their perception of GPT and the responses from respondent neurologists, we eliminated any mentions of artificial intelligence (AI) in the GPT-generated responses. Subsequently, we used a random number generator to generate random identifiers for 8 answers. The coder remained unaware of whether these identifiers were generated by a neurologist or by GPT, ensuring a double-blind approach. The randomly coded answers were independently reviewed and scored by 5 evaluator neurologists and 5 family members of patients for each response. The evaluator neurologists were asked to provide their demographic information, including sex, age, professional title, and years of practice. Each response was assessed using a 5-point Likert scale ranging from 1 (strongly disagree) to 5 (strongly agree), with criteria including accuracy, comprehensiveness, comprehensibility, and overall satisfaction [25,37-40]. Table S3 in [Multimedia Appendix 2](#) shows the questionnaire items for the evaluator neurologists. To further assess the quality of the responses, optional open-ended questions were included, and the evaluator neurologists were asked to explain their scoring choices. The evaluator neurologists were also asked to select the most satisfactory and least satisfactory responses. Patients' family members were asked to provide their demographic information, including sex, age, occupation, marital status, duration of illness, stage of illness, annual household income, and education level. Each response provided was rated on a 5-point Likert scale ranging from 1 (strongly

disagree) to 5 (strongly agree), focusing on criteria such as comprehensibility, usefulness, and overall satisfaction. Table S4 in [Multimedia Appendix 2](#) details the questionnaire items used for the patients' family members. Furthermore, patients' family members were asked to select the most satisfactory and least satisfactory response from each set of responses. In this study, we implemented several measures to ensure that all participants, including patients' family members, could fully comprehend the English responses generated by GPT. All English questionnaires and responses were carefully translated into Chinese by qualified professionals and underwent a rigorous review process to ensure linguistic accuracy and cultural relevance. This translation process was aimed at minimizing any language barriers for the participants ([Multimedia Appendix 3](#)).

Statistical Analysis

Descriptive statistics were calculated for the Likert scale ratings of each response category, including means (SDs). We evaluated the distribution of continuous variables using the Shapiro-Wilk test with IBM SPSS Statistics version 26.0, which indicated nonnormal distribution ([Multimedia Appendices 4 and 5](#)). To compare the ratings of responses produced by GPT and those generated by neurologists, we conducted the Mann-Whitney *U* test. The statistical significance level was set at $P < .05$. Furthermore, we calculated intraclass correlation coefficients (ICCs) to assess the reliability among evaluators. The ICC estimates and 95% CIs were calculated with a 2-way mixed-effects model using both the mean and consistency definitions for the k rater type. The ICC values were interpreted

as follows: values <0.5 indicate low reliability, 0.5-0.74 indicate moderate reliability, 0.75-0.9 indicate good reliability, and >0.9 indicate excellent reliability [41]. All questionnaires were input into IBM SPSS Statistics version 26.0 for processing. We used Microsoft Excel version 2304 to rank the satisfaction of each response; stacked bar charts were generated by Microsoft Excel version 2304.

Ethical Considerations

All participants in this study completed the questionnaire voluntarily to ensure equal participation. The data collected do not contain any personally identifiable information, as all information was anonymized to comply with data protection regulations. According to Article 32 of the Regulations on Ethical Review of Biomedical Research Involving Human Subjects in China [42], this study was exempt from the requirement to obtain approval from an ethical review board.

Results

Study Participants

As of April 10, 2023, we enrolled 9 neurologists, 4 respondent neurologists, and 5 evaluator neurologists in the study. The respondent neurologists included 2 (50%) neurology specialists, 1 (25%) fellow in neurology, and 1 (35%) neurologist, while the evaluator neurologists included 1 (20%) neurology professor,

1 (20%) neurology specialist, 2 (40%) fellows in neurology, and 1 (20%) neurologist. The mean age of the respondent neurologists and evaluator neurologists was 38.50 (SD 10.91) and 43.80 (SD 16.83) years, respectively. The mean clinical experience of the respondent neurologists and evaluator neurologists was 12.75 (SD 8.81) and 19.60 (SD 16.83) years, respectively. The ICC value of the respondent neurologists and evaluator neurologists was 0.95 (95% CI 0.42-0.99) and 0.98 (95% CI 0.85-0.99), respectively, indicating good reliability. Table 1 shows the characteristics of the respondent neurologists and evaluator neurologists in the study.

As of April 2023, we invited 5 family members of patients with AD to participate in this study. Most participants were female (n=4, 80%), with a mean age of 31.20 (SD 14.87) years. In terms of occupation, 3 (60%) participants were students, 1 (20%) was employed, and the remaining 1 (20%) was self-employed. Their annual family income was concentrated (n=4, 80%) in the range of Chinese Yuan (CNY) 10,000-20,000 (~US \$1406-\$2813), and the majority of them (n=4, 80%) lived in urban areas. In addition, their educational background ranged from primary school to a master's degree. Regarding the stage of AD, 4 (80%) were in the middle stage and 1 (20%) in the late stage. The mean duration of AD was 7.0 (SD 4.9) years. The ICC value was 0.89 (95% CI 0.48-0.99), also showing good reliability. Table 2 shows the characteristics of the patients' family members who participated in the survey.

Table 1. Characteristics of neurologists participating in the study.

Characteristics	Respondent neurologists (n=4)	Evaluator neurologists (n=5)
Sex, n (%)		
Male	2 (50)	3 (60)
Female	2 (50)	2 (40)
Age (years), mean (SD)	38.50 (10.91)	43.80 (16.83)
Title of public health technician, n (%)		
Neurology professor	0	1 (20)
Neurology specialist	2 (50)	1 (20)
Fellow in neurology	1 (25)	2 (40)
Neurologist	1 (25)	1 (20)
Clinical experience (years), mean (SD)	12.75 (8.81)	19.60 (16.83)
ICC^a (95% CI)	0.95 (0.42-0.99)	0.98 (0.85-0.99)

^aICC: intraclass correlation coefficient.

Table 2. Characteristics of patients' family members (n=5).

Characteristics	Value
Sex, n (%)	
Male	1 (20)
Female	4 (80)
Age group (years), n (%)	
18-29	3 (60)
30-50	1 (20)
51-52	1 (20)
Occupation, n (%)	
Student	3 (60)
Customer service	1 (20)
Other	1 (20)
Marital status, n (%)	
Married	3 (60)
Unmarried	2 (40)
Living place, n (%)	
Urban	4 (80)
Rural	1 (20)
Education, n (%)	
Primary school diploma	1 (20)
High school diploma	1 (20)
Bachelor's degree	1 (20)
Master's degree	2 (40)
Household income (CNY^a per annum), n (%)	
10,000-20,000 (~US \$1406-\$2813)	4 (80)
20,000-50,000 (~US \$2813-\$7032)	1 (20)
Stage of AD^b, n (%)	
Middle stage	4 (80)
Late stage	1 (20)
Duration of AD (years), mean (SD)	7.0 (4.9)
ICC^c value (95% CI)	0.89 (0.48-0.99)

^aCNY: Chinese Yuan. An exchange rate of CNY 1= US \$0.14 was applied.

^bAD: Alzheimer's disease.

^cICC: intraclass correlation coefficient.

Comparison of GPT Responses with Neurologist Responses

A total of 112 questions were answered by GPT and the respondent neurologists. Specifically, GPT-3.5 and GPT-4 each provided responses to 14 questions, both in English and in Chinese. Additionally, the 4 respondent neurologists individually answered 14 questions in a manner consistent with GPT's responses. All responses are included in the questionnaire ([Multimedia Appendix 2](#)). The average length of GPT responses to each question was 668.9 (SD 218.9) characters, of which

GPT-3.5 responses had an average length of 630.9 (SD 212.7) characters, GPT-4 responses had an average length of 708.2 (SD 222.2) characters, and the average length of the respondent neurologists' responses to each question was 577.6 (SD 717.3) characters. In the evaluation of the most satisfactory responses for each questionnaire, the top 4 (80%) of the top 5 responses rated by the respondent neurologists were GPT responses and the fifth was a respondent neurologist's response ([Multimedia Appendix 1](#)). The top 5 responses with the highest scores rated by the patients' family members were GPT responses (n=4, 80%), except for the third response to a respondent neurologist's

answer (Multimedia Appendix 2). Overall, GPT’s responses received higher satisfaction scores from both evaluator neurologists and patients’ families compared to those from respondent neurologists.

Results of Evaluator Neurologists’ Review of GPT- and Neurologist-Generated Responses

In the evaluator neurologists’ evaluations, on average, both GPT- and neurologist-generated responses were rated as “agree” in comprehensiveness and satisfaction. In terms of accuracy and comprehensibility, GPT-generated responses were rated as “agree,” while neurologist-generated responses were rated as “neither agree nor disagree.” Figure 2 shows the stacked bar graphs representing the scores of each item of the GPT- and

neurologist-generated responses. The detailed information for each item is listed in Table S3 in Multimedia Appendix 2.

Compared to GPT-generated responses, neurologist-generated responses scored lower in accuracy (GPT: mean 4.3, SD 0.6; neurologists: mean 3.7, SD 0.8; $P=.04$), comprehensiveness (GPT: mean 4.4, SD 0.6; neurologists: mean 4.0, SD 0.7; $P=.11$), comprehensibility (GPT: mean 4.4, SD 0.6; neurologists: mean 3.9, SD 0.8; $P<.001$), and satisfaction (GPT: mean 4.4, SD 0.6; neurologists: mean 4.0, SD 0.6; $P<.001$). The mean overall score for neurologist-generated responses was 3.9 (SD 0.7), while that for GPT-generated responses was 4.4 (SD 0.6; $P<.001$). Table 3 shows the means (SDs) for each aspect, and Table S5 in Multimedia Appendix 2 shows the means (SDs) for each item.

Figure 2. Stacked bar charts of neurologists’ scores of each item for GPT (GPT-3.5+GPT-4)-generated responses and neurologist-generated responses, GPT-3.5- and GPT-4-generated responses, Chinese responses, and English responses. GPT: Generative Pretrained Transformer; S1: accuracy; S2: comprehensiveness; S3, S4, S7: comprehensibility; S5, S6, S8: satisfaction.

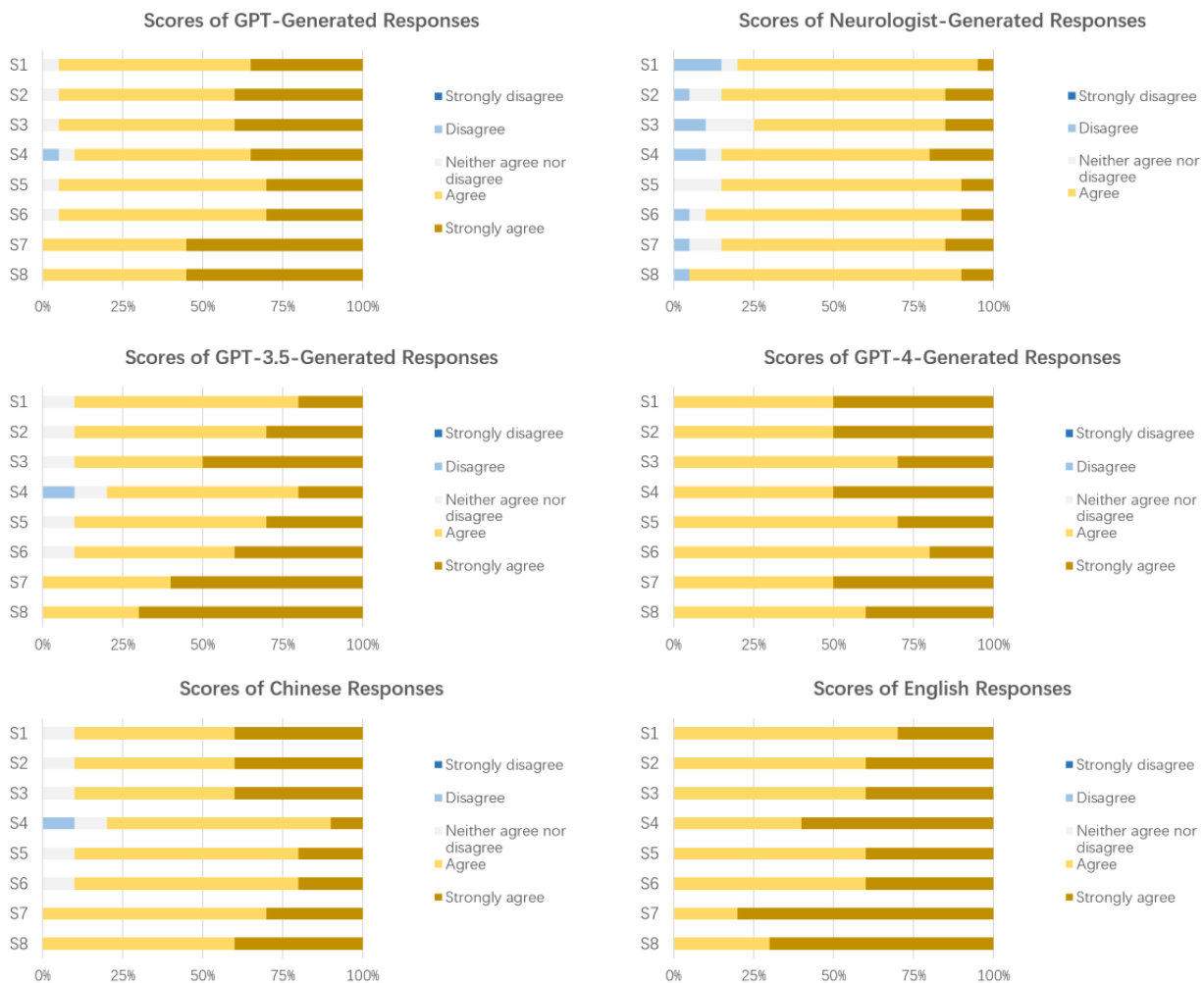


Table 3. Evaluator neurologists' ratings.

Group and response category	Accuracy	Comprehensiveness	Comprehensibility	Satisfaction	Overall
Group 1					
GPT ^{a,b} -generated responses, mean (SD)	4.3 (0.6)	4.4 (0.6)	4.4 (0.6)	4.4 (0.6)	4.4 (0.6)
Neurologist-generated responses, mean (SD)	3.7 (0.8)	4.0 (0.7)	3.9 (0.8)	4.0 (0.6)	3.9 (0.7)
<i>P</i> value	.04	.11	<.001	<.001	<.001
Group 2					
GPT-3.5-generated responses, mean (SD)	4.1 (0.6)	4.2 (0.6)	4.3 (0.7)	4.4 (0.9)	4.3 (0.7)
GPT-4-generated responses, mean (SD)	4.5 (0.5)	4.5 (0.5)	4.4 (0.5)	4.3 (0.5)	4.4 (0.5)
<i>P</i> value	.19	.35	.67	.36	.51
Group 3					
GPT ^b Chinese responses, mean (SD)	4.3 (0.7)	4.3 (0.7)	4.1 (0.7)	4.2 (0.8)	4.4 (0.6)
GPT ^b English responses, mean (SD)	4.3 (0.5)	4.4 (0.5)	4.6 (0.5)	4.5 (0.5)	4.5 (0.5)
<i>P</i> value	.91	.85	.005	.04	.002

^aGPT: Generative Pretrained Transformer.

^bGPT-3.5+GPT-4.

To further analyze the responses generated by GPT, we divided the responses into 2 groups according to the model that generated each response and the language of the response. We further compared the scores of each aspect (Table 3) and the scores of the items of the 2 groups of responses generated by GPT. Figure 2 shows stacked bar charts for the scores of each item. The results indicated that there were no statistically significant differences between GPT-3.5 and GPT-4 in terms of accuracy, comprehensiveness, comprehensibility, and satisfaction. However, although there were no significant differences in accuracy and comprehensiveness scores between Chinese and English responses, there were significant differences in comprehensibility (Chinese: mean 4.1, SD 0.7; English: mean 4.6, SD 0.5; $P=.005$) and satisfaction (Chinese: mean 4.2, SD 0.8; English: mean 4.5, SD 0.5; $P=.04$) scores. The mean total score for Chinese responses was 4.4 (SD 0.6), while that for English responses was 4.5 (SD 0.5; $P=.002$). This suggests that GPT performs differently in different languages, with responses in the English context being more comprehensible and satisfying.

Results of Patient Families' Review of GPT- and Neurologist-Generated Responses

In patient family ratings, on average, the comprehensibility of GPT-generated responses was rated as "agree," while practicality and satisfaction were rated as "neither agree nor disagree." Figure 3 shows stacked bar graphs representing the scores of each item of GPT- and neurologist-generated responses. The neurologist-generated responses scored similarly to the GPT-generated responses in terms of comprehensibility, usefulness, and satisfaction. The total scores for both were the same, 3.9. The means (SDs) for each aspect are shown in Table 4, the detailed information for each item is listed in Table S4 in Multimedia Appendix 2, and the means (SDs) for each item are shown in Table S8 in Multimedia Appendix 2.

As mentioned before, we divided the responses into 2 groups and compared the scores of each aspect (Table 4) and the scores of the items of the 2 groups of responses generated by GPT. We found that there were no significant differences in the comprehensibility, usefulness, and satisfaction scores between GPT-3.5 and GPT-4 and between Chinese and English responses.

Figure 3. Stacked bar charts of patient family scores for each item for GPT- and neurologist-generated responses, GPT-3.5- and GPT-4-generated responses, Chinese responses, and English responses. GPT: Generative Pretrained Transformer; I1-I3: comprehensibility; I4, I6: practicality; I5, I7, I8: satisfaction.

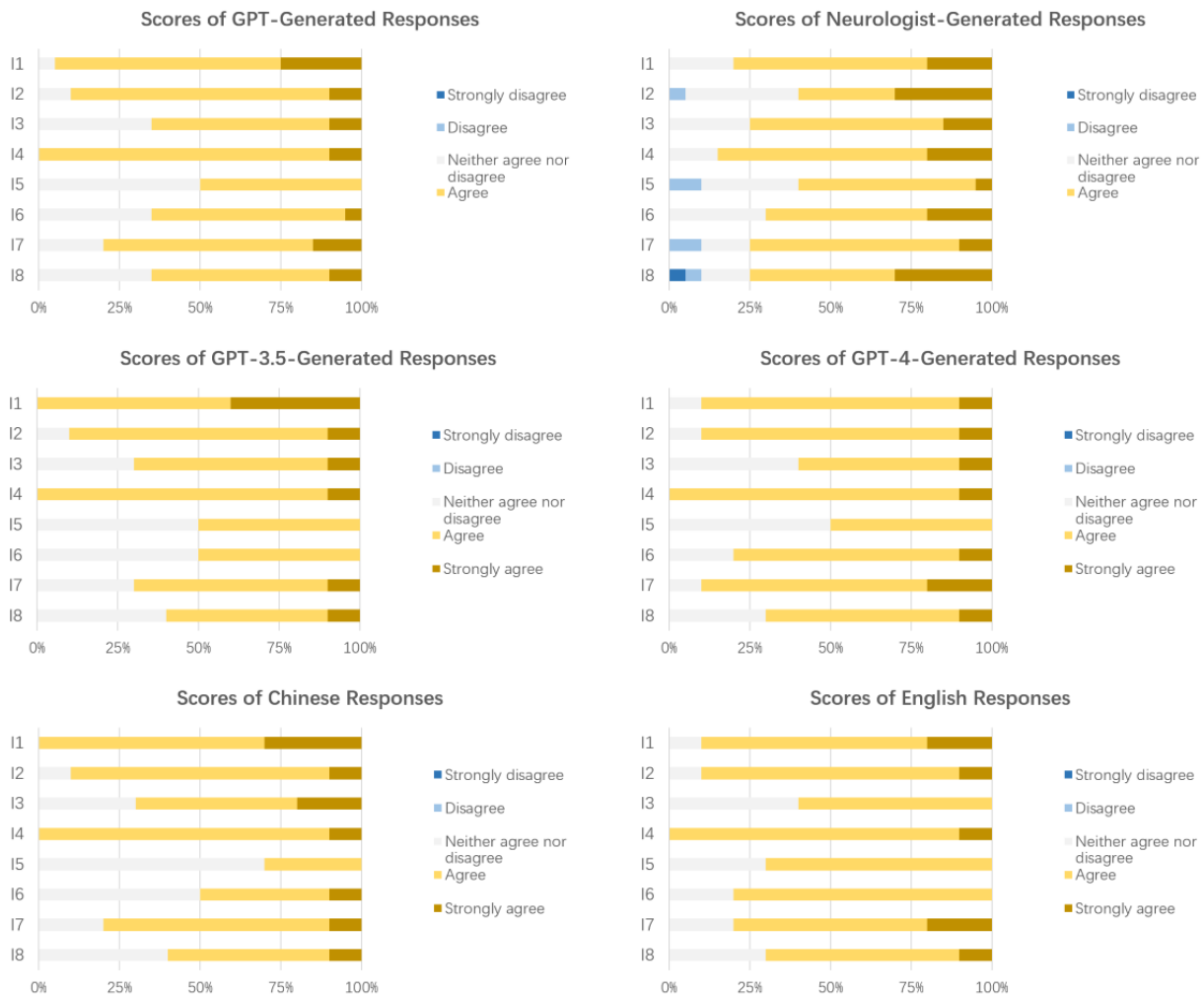


Table 4. Patients’ family members’ ratings.

Group and response category	Comprehensibility	Usefulness	Satisfaction	Overall
Group 1				
GPT ^{a,b} -generated responses, mean (SD)	4.0 (0.5)	3.9 (0.5)	3.7 (0.6)	3.9 (0.6)
Neurologist-generated responses, mean (SD)	3.9 (0.7)	4.0 (0.7)	3.7 (0.9)	3.9 (0.8)
<i>P</i> value	.62	.60	.57	.75
Group 2				
GPT-3.5-generated responses, mean (SD)	4.1 (0.6)	3.8 (0.5)	3.7 (0.6)	3.9 (0.6)
GPT-4-generated responses, mean (SD)	3.9 (0.5)	4.0 (0.5)	3.8 (0.6)	3.9 (0.6)
<i>P</i> value	.26	.34	.39	.65
Group 3				
GPT ^b Chinese responses, mean (SD)	4.1 (0.6)	3.9 (0.6)	3.6 (0.6)	3.9 (0.6)
GPT ^b English responses, mean (SD)	3.9 (0.5)	4.0 (0.4)	3.8 (0.6)	3.9 (0.5)
<i>P</i> value	.26	.60	.19	.62

^aGPT: Generative Pretrained Transformer.

^bGPT-3.5+ GPT-4.

Comments on GPT- and Neurologist-Generated Responses

In this study, the comments of the 5 evaluator neurologists on all responses were qualitatively analyzed. According to the evaluator neurologists' comments, the GPT responses had some shortcomings:

- Deviation in understanding the question. For example, 1 of the GPT responses to the question "What are the precautions for medication in AD, and how can medication-related side effects be minimized?" was "diet and lifestyle", and the advice on care and minimizing medication-related side effects was confusing.
- Omission of information. Some evaluator neurologists commented that the responses generated by GPT were incomplete. For example, in response to the question "How can I prevent AD? Please give me advice on alternative medications, lifestyle, and diet," the list of medications was incomplete.

There were more problems with the responses generated by the 4 respondent neurologists than with those generated by GPT. Problems found with respondent neurologists' responses included:

- Missing information. For example, for the question "How can I prevent AD? Please give me alternative advice on medication, lifestyle, and diet," 1 (25%) respondent neurologist did not provide a comprehensive list of risk factors.
- Template responses. For example, 2 (50%) respondent neurologists provided overly structured answers to the questions "What is the role of regular physical examinations and health assessments in preventing AD?" and "What are the causes or risk factors for AD?"
- Excessive use of medical jargon, for example, in responses to the questions "What is the role of regular health checks and health assessments in preventing AD?" and "What symptoms should a person with AD look for and go to the emergency room for?"
- Lack of relevance. For example, for the question "What are the precautions for medication in AD, and how can medication-related side effects be minimized?" 1 (25%) respondent neurologist did not mention ways to avoid medication-related side effects.
- Lack of practicality. For example, the response of 1 (25%) respondent neurologist to the question "What precautions should I take when treating people with AD with medication, and how can I minimize adverse effects related to medication?" was not practical.
- Potential for ambiguity. For example, when asked, "How do you determine the stage of progression of AD in a patient?", 1 (25%) respondent neurologist responded that there are 3 stages, with a total of 12 years, which could lead to confusion about the patient's life expectancy.

These results suggest that although GPT may suffer from misunderstanding, a lack of information, and a lack of relevance, its overall rating is still better than the individual responses of respondent neurologists.

Discussion

Principal Findings

The main finding of this research highlights the exceptional performance of GPT responses compared to those provided by evaluator neurologists, showing a significant advantage across 4 key dimensions: accuracy, comprehensiveness, comprehensibility, and overall satisfaction (Table 3). Although neurologists' responses still performed well in all aspects, they were outperformed by those of GPT. Additionally, a qualitative analysis revealed certain limitations in both GPT's and neurologists' responses, with more shortcomings identified in the neurologists' responses. Qualitative analysis can be influenced by factors such as the specific questions asked, the expertise of the neurologist, and the criteria used to score the responses. These findings align with prior research indicating the superior performance of GPT compared to human professionals in various medical specialties, particularly in cardiology, interventional radiology, and ophthalmology [43-45]. However, the application of GPT in a clinical setting must be approached with care, emphasizing ethics and transparency to safeguard patients' rights and the quality of health care [25,46].

The superior performance of GPT responses compared to those of neurologists can be attributed to GPT's remarkable potential for comprehensive data coverage [47]. GPT excels in effectively covering a wide spectrum of medical literature, allowing it to provide information that spans different medical domains [48]. In contrast, neurologists tend to have expertise focused on specific medical specialties or in dealing with individual patient histories, which tends to limit their scope of response [49]. The discrepancies observed between the responses of GPT and those of neurologists can be attributed to various factors affecting the latter. These include time constraints, the depth of expertise, and the variability inherent in individual knowledge and experience. Conversely, the limitations in GPT's performance may be due to deficiencies in training data, which may include inaccuracies and omissions [43,50]. Moreover, as an AI language model, GPT may not fully capture and communicate the contextual nuances and emotional subtleties that are often essential in medical consultations [51,52]. A remarkable limitation is its interaction style: When faced with ambiguous queries, GPT lacks the capability to request clarification through follow-up questions, such as "Did you mean home care services?" [50]. Furthermore, several studies have indicated that the responses generated by GPT may present readability challenges for a general audience, potentially impacting their comprehensibility and accessibility [50,53-56]. Therefore, it is important to maintain a balanced perspective when evaluating the effectiveness of AI models like GPT in clinical settings. This involves acknowledging their strengths in terms of data coverage and information retrieval, while also being aware of their limitations in areas where human practitioners outperform them.

In the comparative analysis, GPT-4 (score range 4.2-4.5) outperformed GPT-3.5 (score range 3.9-4.7). However, this difference did not have statistical significance ($P>.05$),

indicating that the improvements in GPT-4 may not be significant enough to substantially improve the quality of responses to AD-related queries. Nevertheless, GPT-4 has displayed superior performance in various domains, especially in the USMLE and StatPearls questions related to ophthalmology, epilepsy, and patient education. GPT-4 has statistically significantly outperformed GPT-3.5 ($P < .001$ to $< .005$) [45,57,58]. Additionally, it has proven to be effective in delivering accurate and comprehensible information to patients about medical procedures, associated risks, benefits, and recovery periods, thereby supporting informed decision-making [59]. OpenAI reports that GPT-4 is “82% less likely to respond to requests for disallowed content and 40% more likely to produce factual responses than GPT-3.5 on our internal evaluations” [60]. Moreover, GPT-4 benefits from training on more recent data, extending up to September 2021, including up-to-date information, unlike GPT-3.5, which was limited to data available before June 2021. This enhancement allows GPT-4 to generate more up-to-date responses [61]. There has been a notable increase in the consistency of GPT-4’s responses over time, possibly due to its robust training and advanced sampling mechanisms that promote response stability [62]. With continued advances in AI, the GPT-4 is poised to become a valuable tool for patient management and health care delivery.

The comparison analysis showed that the responses generated by the GPT model in English (score range 4.3-4.7) significantly outperformed those in Chinese (score range 3.8-4.5; $P < .05$). This difference in performance suggests more advanced training and proficiency of the GPT model in English language contexts, indicating that responses to AD-related consultations may be more effective in English. Additionally, Hristidis et al’s [50] research supported the limited capability of the GPT model in languages other than English, attributing this to a relatively lower level of training and development in non-English languages. Conversely, Takagi et al [57] provided evidence of GPT-4’s effectiveness in the areas of clinical reasoning and medical knowledge within the Japanese language context. This finding provides an important perspective on the potential of GPT models in different linguistic environments, emphasizing the necessity for more extensive training and development in languages other than English to achieve optimal performance in a wider range of linguistic settings.

To better evaluate the usefulness of the GPT model’s responses, we compared the responses of the GPT model to patients’ family members with those provided by the neurologists. The focus was on comprehensibility, usefulness, and overall satisfaction with the patient’s actual care needs. The results showed that the responses of the GPT model were similar to those of the neurologists, with a score of 3.9, and the difference was not statistically significant ($P > .05$). This suggests that the similarity in scores may be due to the limited medical expertise of the

patients’ families. Their ability to recognize the professional depth of the responses may be limited, leading them to prioritize the usefulness and understandability of the information over its medical accuracy and complexity [58,63,64]. This observation emphasizes a crucial aspect of using AI in patient care communication: the need to balance professional medical advice with lay understandability and applicability.

Limitations

There are several limitations of our study that need to be discussed. First, the sample size used for assessment was relatively small, consisting of only 5 neurologists and 5 family members of patients with AD as raters. Although we carefully considered the representativeness of both neurologists and family members when selecting them, the limited sample size could introduce potential biases into the results. Factors such as the clinical experience of the selected neurologists, the time constraints they faced, and the educational background and medical knowledge of the patients’ family members may have influenced the outcomes. Consequently, the results of this study may not be fully reflective of the broader population of neurologists and patients’ families. Second, it is important to note that GPT was trained on text data only up to 2021 and does not include information on drugs, clinical guidelines, or research developments beyond that year. Consequently, GPT’s responses may contain outdated clinical recommendations, which could impact the accuracy and relevance of its suggestions. Furthermore, GPT has the capability to generate responses that sound confident but may not necessarily be accurate, including the possibility of providing incorrect answers or fabricated references [65-67]. Finally, it is essential to acknowledge that the assessment process relies on the perceptions of the participants, specifically the assessors who are neuroscientists and family members. Although we have developed normative assessment criteria, these evaluations inherently involve subjective judgments, introducing a degree of subjectivity to the assessment process. These limitations emphasize the need for caution in interpreting our findings and the importance of further research to address these limitations. Such research will contribute to a more comprehensive understanding of the potential and limitations of GPT in health care settings.

Conclusion

The results of this study suggest that GPT is a promising tool in the management of AD. The responses provided by GPT can be a valuable resource, providing useful information and support to patients and their families. However, it is important to emphasize that GPT should be used judiciously and should not be relied upon as the sole source of knowledge and expertise. Neurologists should continue to work with GPT, using their clinical expertise and judgment to ensure that patients receive accurate and tailored treatment recommendations.

Acknowledgments

We used OpenAI’s generative artificial intelligence tool Generative Pretrained Transformer (GPT) to answer questions for the research survey. The original GPT transcripts are provided in [Multimedia Appendix 6](#).

Data Availability

The input prompts and artificial intelligence outputs are available in [Multimedia Appendix 2](#).

Authors' Contributions

JL and S Liu contributed to conceptualization; JZ, JL, and CW to data curation; JZ, XZ, JL, and S Liu to writing—original draft; and JZ, XZ, CW, JL, S Liu, YT, ST, HL, YW, LZ, JL, and S Li to formal analysis, investigation, supervision, and writing—review and editing.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Questionnaire development.

[\[DOCX File , 17 KB-Multimedia Appendix 1\]](#)

Multimedia Appendix 2

Input prompts and AI outputs. AI: artificial intelligence.

[\[DOCX File , 173 KB-Multimedia Appendix 2\]](#)

Multimedia Appendix 3

Translation development process.

[\[DOCX File , 16 KB-Multimedia Appendix 3\]](#)

Multimedia Appendix 4

Normal distribution test of evaluator neurologists' reviews.

[\[DOC File , 69532 KB-Multimedia Appendix 4\]](#)

Multimedia Appendix 5

Normal distribution test of patients' family members' reviews.

[\[DOC File , 53451 KB-Multimedia Appendix 5\]](#)

Multimedia Appendix 6

Original GPT transcripts. GPT: Generative Pretrained Transformer.

[\[TXT File , 80 KB-Multimedia Appendix 6\]](#)

References

1. Kivipelto M, Mangialasche F, Ngandu T. Lifestyle interventions to prevent cognitive impairment, dementia and Alzheimer disease. *Nat Rev Neurol*. Nov 2018;14(11):653-666. [doi: [10.1038/s41582-018-0070-3](https://doi.org/10.1038/s41582-018-0070-3)] [Medline: [30291317](https://pubmed.ncbi.nlm.nih.gov/30291317/)]
2. Patterson C. World Alzheimer report 2018. The state of the art of dementia research: new frontiers. Alzheimer's Disease International. 2018. URL: <https://www.alzint.org/resource/world-alzheimer-report-2018/> [accessed 2023-04-10]
3. Alzheimer's Association. 2021 Alzheimer's disease facts and figures. *Alzheimers Dement*. Mar 23, 2021;17(3):327-406. [doi: [10.1002/alz.12328](https://doi.org/10.1002/alz.12328)] [Medline: [33756057](https://pubmed.ncbi.nlm.nih.gov/33756057/)]
4. Wimo A, Seeher K, Cataldi R, Cyhlarova E, Dielemann JL, Frisell O, et al. The worldwide costs of dementia in 2019. *Alzheimers Dement*. Jan 08, 2023;19(7):2865-2873. [FREE Full text] [doi: [10.1002/alz.12901](https://doi.org/10.1002/alz.12901)] [Medline: [36617519](https://pubmed.ncbi.nlm.nih.gov/36617519/)]
5. Jia J, Wei C, Chen S, Li F, Tang Y, Qin W, et al. The cost of Alzheimer's disease in China and re - estimation of costs worldwide. *Alzheimers Dement*. Feb 09, 2018;14(4):483-491. [doi: [10.1016/j.jalz.2017.12.006](https://doi.org/10.1016/j.jalz.2017.12.006)] [Medline: [29433981](https://pubmed.ncbi.nlm.nih.gov/29433981/)]
6. Kales HC, Gitlin LN, Lyketsos CG. Assessment and management of behavioral and psychological symptoms of dementia. *BMJ*. Mar 02, 2015;350(mar02 7):h369-h369. [FREE Full text] [doi: [10.1136/bmj.h369](https://doi.org/10.1136/bmj.h369)] [Medline: [25731881](https://pubmed.ncbi.nlm.nih.gov/25731881/)]
7. Livingston G, Sommerlad A, Orgeta V, Costafreda SG, Huntley J, Ames D, et al. Dementia prevention, intervention, and care. *Lancet*. Dec 16, 2017;390(10113):2673-2734. [doi: [10.1016/S0140-6736\(17\)31363-6](https://doi.org/10.1016/S0140-6736(17)31363-6)] [Medline: [28735855](https://pubmed.ncbi.nlm.nih.gov/28735855/)]
8. Cummings JL, Tong G, Ballard C. Treatment combinations for Alzheimer's disease: current and future pharmacotherapy options. *J Alzheimers Dis*. Feb 12, 2019;67(3):779-794. [doi: [10.3233/jad-180766](https://doi.org/10.3233/jad-180766)] [Medline: [30689575](https://pubmed.ncbi.nlm.nih.gov/30689575/)]
9. Yu T, Lane H, Lin C. Novel therapeutic approaches for Alzheimer's disease: an updated review. *Int J Mol Sci*. Jul 30, 2021;22(15):8208. [doi: [10.3390/ijms22158208](https://doi.org/10.3390/ijms22158208)] [Medline: [34360973](https://pubmed.ncbi.nlm.nih.gov/34360973/)]

10. Mangialasche F, Solomon A, Winblad B, Mecocci P, Kivipelto M. Alzheimer's disease: clinical trials and drug development. *Lancet Neurol.* Jul 2010;9(7):702-716. [doi: [10.1016/s1474-4422\(10\)70119-8](https://doi.org/10.1016/s1474-4422(10)70119-8)] [Medline: [20610346](https://pubmed.ncbi.nlm.nih.gov/20610346/)]
11. Risk reduction of cognitive decline and dementia: WHO guidelines. World Health Organization. 2019. URL: <https://iris.who.int/bitstream/handle/10665/312180/9789241550543-eng.pdf> [accessed 2023-04-10]
12. Scheltens P, De Strooper B, Kivipelto M, Holstege H, Chételat G, Teunissen CE, et al. Alzheimer's disease. *Lancet.* Apr 2021;397(10284):1577-1590. [doi: [10.1016/s0140-6736\(20\)32205-4](https://doi.org/10.1016/s0140-6736(20)32205-4)] [Medline: [33667416](https://pubmed.ncbi.nlm.nih.gov/33667416/)]
13. Forbes D, Forbes SC, Blake CM, Thiessen EJ, Forbes S. Exercise programs for people with dementia. *Cochrane Database Syst Rev.* Apr 15, 2015;2015(4):CD006489. [doi: [10.1002/14651858.CD006489.pub4](https://doi.org/10.1002/14651858.CD006489.pub4)] [Medline: [25874613](https://pubmed.ncbi.nlm.nih.gov/25874613/)]
14. Yu L, Wilson RS, Schneider JA, Bennett DA, Boyle PA. Financial and health literacy predict incident Alzheimer's disease dementia and pathology. *J Alzheimers Dis.* Feb 20, 2017;56(4):1485-1493. [FREE Full text] [doi: [10.3233/JAD-161132](https://doi.org/10.3233/JAD-161132)] [Medline: [28157101](https://pubmed.ncbi.nlm.nih.gov/28157101/)]
15. Rostamzadeh A, Stapels J, Genske A, Haidl T, Jünger S, Seves M, et al. Health literacy in individuals at risk for Alzheimer's dementia: a systematic review. *J Prev Alzheimers Dis.* 2020;7(1):47-55. [doi: [10.14283/jpad.2019.34](https://doi.org/10.14283/jpad.2019.34)] [Medline: [32010926](https://pubmed.ncbi.nlm.nih.gov/32010926/)]
16. Palotti J, Zuccon G, Hanbury A. Consumer health search on the web: study of web page understandability and its integration in ranking algorithms. *J Med Internet Res.* Jan 30, 2019;21(1):e10986. [FREE Full text] [doi: [10.2196/10986](https://doi.org/10.2196/10986)] [Medline: [30698536](https://pubmed.ncbi.nlm.nih.gov/30698536/)]
17. Werner P, Karnieli - Miller O, Eidelman S. Current knowledge and future directions about the disclosure of dementia: a systematic review of the first decade of the 21st century. *Alzheimers Dement.* Oct 24, 2012;9(2):e74-e88. [doi: [10.1016/j.jalz.2012.02.006](https://doi.org/10.1016/j.jalz.2012.02.006)] [Medline: [23098912](https://pubmed.ncbi.nlm.nih.gov/23098912/)]
18. Clare L, Wu Y, Teale JC, MacLeod C, Matthews F, Brayne C, et al. Potentially modifiable lifestyle factors, cognitive reserve, and cognitive function in later life: a cross-sectional study. *PLoS Med.* Mar 21, 2017;14(3):e1002259. [FREE Full text] [doi: [10.1371/journal.pmed.1002259](https://doi.org/10.1371/journal.pmed.1002259)] [Medline: [28323829](https://pubmed.ncbi.nlm.nih.gov/28323829/)]
19. Gitlin LN, Winter L, Dennis MP, Hodgson N, Hauck WW. A biobehavioral home-based intervention and the well-being of patients with dementia and their caregivers: the COPE randomized trial. *JAMA.* Sep 01, 2010;304(9):983-991. [FREE Full text] [doi: [10.1001/jama.2010.1253](https://doi.org/10.1001/jama.2010.1253)] [Medline: [20810376](https://pubmed.ncbi.nlm.nih.gov/20810376/)]
20. Chesser AK, Keene Woods N, Smothers K, Rogers N. Health literacy and older adults: a systematic review. *Gerontol Geriatr Med.* Mar 15, 2016;2:2333721416630492. [doi: [10.1177/2333721416630492](https://doi.org/10.1177/2333721416630492)] [Medline: [28138488](https://pubmed.ncbi.nlm.nih.gov/28138488/)]
21. Lee H. The rise of ChatGPT: exploring its potential in medical education. *Anat Sci Educ.* Mar 28, 2023;17(5):926-931. [doi: [10.1002/ase.2270](https://doi.org/10.1002/ase.2270)] [Medline: [36916887](https://pubmed.ncbi.nlm.nih.gov/36916887/)]
22. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLoS Digit Health.* Feb 9, 2023;2(2):e0000198. [FREE Full text] [doi: [10.1371/journal.pdig.0000198](https://doi.org/10.1371/journal.pdig.0000198)] [Medline: [36812645](https://pubmed.ncbi.nlm.nih.gov/36812645/)]
23. Jeblick K, Schachtner B, Dextl J, Mittermeier A, Stüber AT, Topalis J, et al. ChatGPT makes medicine easy to swallow: an exploratory case study on simplified radiology reports. *Eur Radiol.* Oct 05, 2023;34(5):2817-2825. [doi: [10.1007/s00330-023-10213-1](https://doi.org/10.1007/s00330-023-10213-1)] [Medline: [37794249](https://pubmed.ncbi.nlm.nih.gov/37794249/)]
24. Liu S, Wright AP, Patterson BL, Wanderer JP, Turer RW, Nelson SD, et al. Using AI-generated suggestions from ChatGPT to optimize clinical decision support. *J Am Med Inform Assoc.* Jun 20, 2023;30(7):1237-1245. [FREE Full text] [doi: [10.1093/jamia/ocad072](https://doi.org/10.1093/jamia/ocad072)] [Medline: [37087108](https://pubmed.ncbi.nlm.nih.gov/37087108/)]
25. Liu J, Wang C, Liu S. Utility of ChatGPT in clinical practice. *J Med Internet Res.* Jun 28, 2023;25:e48568. [FREE Full text] [doi: [10.2196/48568](https://doi.org/10.2196/48568)] [Medline: [37379067](https://pubmed.ncbi.nlm.nih.gov/37379067/)]
26. Bogost I. ChatGPT is dumber than you think: treat it like a toy, not a tool. *The Atlantic.* 2022. URL: <https://www.theatlantic.com/technology/archive/2022/12/chatgpt-openai-artificial-intelligence-writing-ethics/672386/> [accessed 2023-04-10]
27. Nursing and rehabilitation. China Association for Alzheimer's Disease. URL: <http://www.caad.org.cn/col/col50/index.html> [accessed 2023-04-10]
28. The global voice on dementia. Alzheimer's Disease International. URL: <https://www.alzint.org/> [accessed 2023-04-10]
29. Dementia Australia is here for you. Dementia Australia. URL: <https://www.dementia.org.au/> [accessed 2023-04-10]
30. The moment is now. Walk with us. Alzheimer's Association. URL: <https://www.alz.org/> [accessed 2023-04-10]
31. Anna Richardson: love, loss and dementia. Alzheimer's Society. URL: <https://www.alzheimers.org.uk/> [accessed 2023-04-10]
32. Latest news. Hong Kong Alzheimer's Association. URL: <https://www.hkada.org.hk/> [accessed 2023-04-10]
33. Kia ora. Alzheimer New Zealand. URL: <https://alzheimers.org.nz/> [accessed 2023-04-10]
34. Association for Dementia and Alzheimer's of South Africa. URL: <https://www.adasa.org.za/> [accessed 2023-04-10]
35. Dementia Singapore. Dementia Singapore. URL: <https://dementia.org.sg/> [accessed 2023-04-10]
36. Xia D, Chen P, Huang M, Li X. Current status and analysis of online patient education materials released Alzheimer's Disease Associations at home and abroad. *Chin J Health Educ.* 2020;36(02):182-185. [doi: [10.16168/j.cnki.issn.1002-9982.2020.02.020](https://doi.org/10.16168/j.cnki.issn.1002-9982.2020.02.020)]
37. Swart ECS, Good CB, Henderson R, Manolis C, Huang Y, Essien UR, et al. Identifying outcome measures for atrial fibrillation value-based contracting using the Delphi method. *Res Soc Admin Pharm.* Aug 2022;18(8):3425-3431. [doi: [10.1016/j.sapharm.2021.10.008](https://doi.org/10.1016/j.sapharm.2021.10.008)] [Medline: [34764046](https://pubmed.ncbi.nlm.nih.gov/34764046/)]

38. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med.* Jun 2016;15(2):155-163. [FREE Full text] [doi: [10.1016/j.jcm.2016.02.012](https://doi.org/10.1016/j.jcm.2016.02.012)] [Medline: [27330520](https://pubmed.ncbi.nlm.nih.gov/27330520/)]
39. Delanian Halsdorfer N, Blasquez J, Bensoussan L, Gentile S, Collado H, Viton J, et al. An assessment of patient satisfaction for a short-stay program in a physical and rehabilitation medicine day hospital. *Ann Phys Rehabil Med.* Jun 2011;54(4):236-247. [FREE Full text] [doi: [10.1016/j.rehab.2011.04.001](https://doi.org/10.1016/j.rehab.2011.04.001)] [Medline: [21570929](https://pubmed.ncbi.nlm.nih.gov/21570929/)]
40. Gavurova B, Dvorsky J, Popesko B. Patient satisfaction determinants of inpatient healthcare. *IJERPH.* Oct 28, 2021;18(21):11337. [FREE Full text] [doi: [10.3390/ijerph182111337](https://doi.org/10.3390/ijerph182111337)] [Medline: [34769856](https://pubmed.ncbi.nlm.nih.gov/34769856/)]
41. Thayaparan AJ, Mahdi E. The Patient Satisfaction Questionnaire Short Form (PSQ-18) as an adaptable, reliable, and validated tool for use in various settings. *Med Educ Online.* Jul 23, 2013;18(1):21747. [FREE Full text] [doi: [10.3402/meo.v18i0.21747](https://doi.org/10.3402/meo.v18i0.21747)] [Medline: [23883565](https://pubmed.ncbi.nlm.nih.gov/23883565/)]
42. National Health Commission of the People's Republic of China, Ministry of Education, Ministry of Science and Technology, National Administration of Traditional Chinese Medicine. Regulations on ethical review of life sciences and medical research involving human subjects, document no. 4. China Government. Feb 18, 2023. URL: https://www.gov.cn/zhengce/zhengceku/2023-02/28/content_5743658.htm [accessed 2023-04-10]
43. Sarraju A, Bruemmer D, Van Iterson E, Cho L, Rodriguez F, Laffin L. Appropriateness of cardiovascular disease prevention recommendations obtained from a popular online chat-based artificial intelligence model. *JAMA.* Mar 14, 2023;329(10):842-844. [doi: [10.1001/jama.2023.1044](https://doi.org/10.1001/jama.2023.1044)] [Medline: [36735264](https://pubmed.ncbi.nlm.nih.gov/36735264/)]
44. Waisberg E, Ong J, Zaman N, Kamran SA, Sarker P, Tavakkoli A, et al. GPT-4 for triaging ophthalmic symptoms. *Eye (Lond.).* May 25, 2023;37(18):3874-3875. [doi: [10.1038/s41433-023-02595-9](https://doi.org/10.1038/s41433-023-02595-9)] [Medline: [37231187](https://pubmed.ncbi.nlm.nih.gov/37231187/)]
45. Scheschenja M, Viniol S, Bastian MB, Wessendorf J, König AM, Mahnken AH. Feasibility of GPT-3 and GPT-4 for in-depth patient education prior to interventional radiological procedures: a comparative analysis. *Cardiovasc Intervent Radiol.* Oct 23, 2023;47(2):245-250. [FREE Full text] [doi: [10.1007/s00270-023-03563-2](https://doi.org/10.1007/s00270-023-03563-2)] [Medline: [37872295](https://pubmed.ncbi.nlm.nih.gov/37872295/)]
46. Wang C, Liu S, Yang H, Guo J, Wu Y, Liu J. Ethical considerations of using ChatGPT in health care. *J Med Internet Res.* Aug 11, 2023;25:e48009. [doi: [10.2196/48009](https://doi.org/10.2196/48009)] [Medline: [37566454](https://pubmed.ncbi.nlm.nih.gov/37566454/)]
47. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. *Healthcare.* Mar 19, 2023;11(6):887. [FREE Full text] [doi: [10.3390/healthcare11060887](https://doi.org/10.3390/healthcare11060887)] [Medline: [36981544](https://pubmed.ncbi.nlm.nih.gov/36981544/)]
48. Huh S. Issues in the 3rd year of the COVID-19 pandemic, including computer-based testing, study design, ChatGPT, journal metrics, and appreciation to reviewers. *J Educ Eval Health Prof.* Jan 31, 2023;20:5. [FREE Full text] [doi: [10.3352/jeehp.2023.20.5](https://doi.org/10.3352/jeehp.2023.20.5)] [Medline: [36718045](https://pubmed.ncbi.nlm.nih.gov/36718045/)]
49. Sallam M, Salim NA, Al-Tammemi AB, Barakat M, Fayyad D, Hallit S, et al. ChatGPT output regarding compulsory vaccination and COVID-19 vaccine conspiracy: a descriptive study at the outset of a paradigm shift in online search for information. *Cureus.* Feb 2023;15(2):e35029. [FREE Full text] [doi: [10.7759/cureus.35029](https://doi.org/10.7759/cureus.35029)] [Medline: [36819954](https://pubmed.ncbi.nlm.nih.gov/36819954/)]
50. Hristidis V, Ruggiano N, Brown EL, Ganta SRR, Stewart S. ChatGPT vs Google for queries related to dementia and other cognitive decline: comparison of results. *J Med Internet Res.* Jul 25, 2023;25:e48966. [FREE Full text] [doi: [10.2196/48966](https://doi.org/10.2196/48966)] [Medline: [37490317](https://pubmed.ncbi.nlm.nih.gov/37490317/)]
51. Zhu L, Mou W, Chen R. Can the ChatGPT and other large language models with internet-connected database solve the questions and concerns of patient with prostate cancer and help democratize medical knowledge? *J Transl Med.* Apr 19, 2023;21(1):269. [FREE Full text] [doi: [10.1186/s12967-023-04123-5](https://doi.org/10.1186/s12967-023-04123-5)] [Medline: [37076876](https://pubmed.ncbi.nlm.nih.gov/37076876/)]
52. Ray P. ChatGPT: a comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet Things Cyber-Phys Syst.* 2023;3:121-154. [FREE Full text] [doi: [10.1016/j.iotcps.2023.04.003](https://doi.org/10.1016/j.iotcps.2023.04.003)]
53. Johnson SB, King AJ, Warner EL, Aneja S, Kann BH, Bylund CL. Using ChatGPT to evaluate cancer myths and misconceptions: artificial intelligence and cancer information. *JNCI Cancer Spectr.* Mar 01, 2023;7(2):pkad015. [FREE Full text] [doi: [10.1093/jncics/pkad015](https://doi.org/10.1093/jncics/pkad015)] [Medline: [36929393](https://pubmed.ncbi.nlm.nih.gov/36929393/)]
54. Momenaei B, Wakabayashi T, Shahlaee A, Durrani AF, Pandit SA, Wang K, et al. Appropriateness and readability of ChatGPT-4-generated responses for surgical treatment of retinal diseases. *Ophthalmol Retina.* Oct 03, 2023;7(10):862-868. [doi: [10.1016/j.oret.2023.05.022](https://doi.org/10.1016/j.oret.2023.05.022)] [Medline: [37277096](https://pubmed.ncbi.nlm.nih.gov/37277096/)]
55. Moons P, Van Bulck L. Using ChatGPT and Google Bard to improve the readability of written patient information: a proof of concept. *Eur J Cardiovasc Nurs.* Mar 12, 2024;23(2):122-126. [doi: [10.1093/eurjcn/zvad087](https://doi.org/10.1093/eurjcn/zvad087)] [Medline: [37603843](https://pubmed.ncbi.nlm.nih.gov/37603843/)]
56. Haver HL, Lin CT, Sirajuddin A, Yi PH, Jeudy J. Use of ChatGPT, GPT-4, and Bard to improve readability of ChatGPT's answers to common questions about lung cancer and lung cancer screening. *Am J Roentgenol.* Nov 20, 2023;221(5):701-704. [doi: [10.2214/ajr.23.29622](https://doi.org/10.2214/ajr.23.29622)] [Medline: [37341179](https://pubmed.ncbi.nlm.nih.gov/37341179/)]
57. Takagi S, Watari T, Erabi A, Sakaguchi K. Performance of GPT-3.5 and GPT-4 on the Japanese Medical Licensing Examination: comparison study. *JMIR Med Educ.* Jun 29, 2023;9:e48002. [FREE Full text] [doi: [10.2196/48002](https://doi.org/10.2196/48002)] [Medline: [37384388](https://pubmed.ncbi.nlm.nih.gov/37384388/)]
58. Moshirfar M, Altaf AW, Stoakes IM, Tuttle JJ, Hoopes PC. Artificial intelligence in ophthalmology: a comparative analysis of GPT-3.5, GPT-4, and human expertise in answering StatPearls questions. *Cureus.* Jun 2023;15(6):e40822. [FREE Full text] [doi: [10.7759/cureus.40822](https://doi.org/10.7759/cureus.40822)] [Medline: [37485215](https://pubmed.ncbi.nlm.nih.gov/37485215/)]

59. Kim H, Shin D, Kim J, Lee G, Cho JW. Assessing the performance of ChatGPT's responses to questions related to epilepsy: a cross-sectional study on natural language processing and medical information retrieval. *Seizure: Eur J Epilepsy*. Jan 2024;114:1-8. [FREE Full text] [doi: [10.1016/j.seizure.2023.11.013](https://doi.org/10.1016/j.seizure.2023.11.013)] [Medline: [38007922](https://pubmed.ncbi.nlm.nih.gov/38007922/)]
60. GPT-4 is OpenAI's most advanced system, producing safer and more useful responses. Open AI. URL: <https://openai.com/product/gpt-4> [accessed 2023-12-01]
61. Martindale J, Bizzaco M. GPT-4 vs. GPT-3.5: how much difference is there? Digital Trends Media Group. Apr 4, 2024. URL: <https://www.digitaltrends.com/computing/gpt-4-vs-gpt-35/> [accessed 2023-12-01]
62. Brin D, Sorin V, Vaid A, Soroush A, Glicksberg BS, Charney AW, et al. Comparing ChatGPT and GPT-4 performance in USMLE soft skill assessments. *Sci Rep*. Oct 01, 2023;13(1):16492. [FREE Full text] [doi: [10.1038/s41598-023-43436-9](https://doi.org/10.1038/s41598-023-43436-9)] [Medline: [37779171](https://pubmed.ncbi.nlm.nih.gov/37779171/)]
63. Jackson DN, Trivedi N, Baur C. Re-prioritizing digital health and health literacy in Healthy People 2030 to affect health equity. *Health Commun*. Apr 30, 2020;36(10):1155-1162. [doi: [10.1080/10410236.2020.1748828](https://doi.org/10.1080/10410236.2020.1748828)] [Medline: [32354233](https://pubmed.ncbi.nlm.nih.gov/32354233/)]
64. Bangerter LR, Griffin J, Harden K, Rutten LJ. Health information-seeking behaviors of family caregivers: analysis of the Health Information National Trends Survey. *JMIR Aging*. Jan 14, 2019;2(1):e11237. [FREE Full text] [doi: [10.2196/11237](https://doi.org/10.2196/11237)] [Medline: [31518309](https://pubmed.ncbi.nlm.nih.gov/31518309/)]
65. ChatGPT: optimizing language models for dialogue. OpenAI. URL: <https://openai.com/blog/chatgpt/> [accessed 2023-04-10]
66. Castelvechi D. Are ChatGPT and AlphaCode going to replace programmers? *Nature*. Dec 08, 2022:Online ahead of print. [doi: [10.1038/d41586-022-04383-z](https://doi.org/10.1038/d41586-022-04383-z)] [Medline: [36481949](https://pubmed.ncbi.nlm.nih.gov/36481949/)]
67. Stokel-Walker C. AI bot ChatGPT writes smart essays — should professors worry? *Nature*. Dec 09, 2022:Online ahead of print. [doi: [10.1038/d41586-022-04397-7](https://doi.org/10.1038/d41586-022-04397-7)] [Medline: [36494443](https://pubmed.ncbi.nlm.nih.gov/36494443/)]

Abbreviations

- AD:** Alzheimer's disease
AI: artificial intelligence
CNY: Chinese Yuan
GPT: Generative Pretrained Transformer
ICC: intraclass correlation coefficients
NLP: natural language processing
USMLE: States Medical Licensing Examination

Edited by G Eysenbach, T Leung; submitted 23.07.23; peer-reviewed by L Zhu, K Giannakopoulos, A Kavadella; comments to author 16.09.23; revised version received 06.10.23; accepted 25.09.24; published 31.10.24

Please cite as:

Zeng J, Zou X, Li S, Tang Y, Teng S, Li H, Wang C, Wu Y, Zhang L, Zhong Y, Liu J, Liu S
Assessing the Role of the Generative Pretrained Transformer (GPT) in Alzheimer's Disease Management: Comparative Study of Neurologist- and Artificial Intelligence-Generated Responses
J Med Internet Res 2024;26:e51095
URL: <https://www.jmir.org/2024/1/e51095>
doi: [10.2196/51095](https://doi.org/10.2196/51095)
PMID:

©Jiaqi Zeng, Xiaoyi Zou, Shirong Li, Yao Tang, Sisi Teng, Huanhuan Li, Changyu Wang, Yuxuan Wu, Luyao Zhang, Yunheng Zhong, Jialin Liu, Siru Liu. Originally published in the Journal of Medical Internet Research (<https://www.jmir.org>), 31.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on <https://www.jmir.org/>, as well as this copyright and license information must be included.