AI based chatbots vs.traditional search: A systematic comparison of response quality for dementia management information*

Sourav Maiti[1,*,†], Syeda Mah-e-Fatima[2], Qurratal Ain Fatimah[2] and Ali Hasnain[1,*,†]

[1]School of Pharmacy and Biomedical Sciences, Royal College of Surgeons in Ireland, Dublin
[2]University Hospital Galway, Ireland

Abstract

Artificial intelligence (AI) has emerged as a promising technology in healthcare to answer questions and aid in clinical decision-making processes. This research presents a comparative analysis of three technologies: Google(search engine), ChatGPT-3.5 (AI bot) and ThinkAny (AI search engine), used in the context of asking questions about dementia management while focusing on the results generated by the aforementioned and comparing with a benchmark (standard clinical guidelines). The approach relies on asking questions to these technologies and systematically analyzing the generated responses. The methodology follows performing a series of statistical tests including the Friedman, Wilcoxon signed rank, and Mann-Whitney U test(s) to evaluate the responses generated by the aforementioned and comparing against the standard clinical guidelines (SCG).

Our initial findings reveal significant variations in the responses generated by each technology when comparing with the standard clinical guidelines. The response generated by Google exhibited the greatest deviation from the clinical guidelines across all questions indicating potential limitations in providing accurate information. ChatGPT-3.5 and ThinkAny demonstrated closer adherence towards clinical guideline for some questions but significant dispenses were also observed, indicating they lack the accuracy, credibility and relevance provided by clinical guidelines. ChatGPT-3.5 and ThinkAny performed well in understanding the context when provided clear, direct and relevant information, but also compromises the accuracy. In contrast, Google's response lacks relevance and precision, which shows limitations in its ability to produce reliable response in comparison with the standard guidelines.

In essence, this research evaluates the performance of AI tools while asking questions related to Dementia Management. It is evident that the full potential of using these tools have not yet been exploited but careful consideration(s) of their limitations and biases could ensure their effective and meaningful use for researcher(s), medical professionals and caregivers (friends and/or family).

Keywords

Dementia Management, Standard Clinical Guidelines (SCG), AI Chatbots, ChatGPT-3.5, ThinkAny, Healthcare and Life Sciences

1. Introduction

Dementia, one of the most common disorders characterized by cognitive decline, affects millions of people and poses significant challenges for patients, healthcare professionals, formal and informal caregivers worldwide. According to Nichols et al.[1], the number of people estimated to have dementia is expected to increase from 57 million cases globally in the year 2019 to an estimated 153 million cases in the year 2050. The increase in the number of dementia patients poses a significant challenge for the global healthcare system to effectively deal with the medical management issues in dementia[2].

Effective management of dementia involves addressing the cognitive, emotional, and physical needs of individuals living with dementia so that they can experience improved quality of life. It also involves providing support for their caregivers so they can provide optimal care to people with dementia. It requires adherence, knowledge, and understanding [3] of established clinical guidelines like the one

SeWebMeDA-2025: 8th International Workshop on Semantic Web Solutions for Large-scale Biomedical Data Analytics, June 1, 2025, Portorož, Slovenia

*Corresponding author.
†These authors contributed equally.
souravmaiti@rcsi.ie (S. Maiti); alihasnain@rcsi.com (A. Hasnain)
https://www.rcsi.com/people/profile/alihasnain (A. Hasnain)
0000-0003-4014-4394 (A. Hasnain)

from the European Academy of Neurology [1]. These guidelines [2] provide recommendations on how to address various management issues related to dementia, including follow-up, vascular risk factors in dementia, and pain management in dementia [4].

Information retrieval systems (IRS) are tools that help in finding information within a large collection of data[5]. Search engines like Google or Bing, were and still are a popular IRS tool for accessing/searching medical information[6]. The search engine helps the users to find relevant information and additional resources on different topics. However, the accuracy, credibility, relevance of the retrieved information from the different search engines vary. For results obtained from a search engine, the user has to identify the accurate and relevant information from a credible and trustworthy source[6]. Currently, search engines have sponsored posts at the top for certain searches, which means search engines get paid to highly rank the content at the top of the search results, even though the result might not be accurate, relevant, or from a credible source. These factors raise questions on the reliability of readily available information to patients, healthcare professionals, formal and informal caregivers looking for responses to questions related to managing dementia.

The importance of IRS in healthcare[5] [7] [8] is growing significantly and, to the best of our knowledge, there is a lack of credible research comparing the performance of different systems for answering the questions related to medical management guidelines for dementia patients, researchers, and caregivers. This research aims to address this gap by investigating how technologies like Google, AI Chatbot (ChatGPT-3.5) and AI search engine (ThinkAny) respond to specific dementia management questions taken from the European Academy of Neurology guidelines[4]. Our work uses a Likert scale rating system[9] to evaluate the accuracy, clarity, simplicity, relevance, and usefulness of the retrieved information from each technology for dementia management questions. We apply statistical analysis techniques including the Friedman test[10], Wilcoxon signed-rank test[11] and Mann-Whitney U test[12] with Bonferroni correction, to identify differences in the performance of the technologies across the questions. By comparing the performance of these technologies through statistical analysis, we aim to gain valuable insights into the strengths, limitations, and effectiveness of the upcoming technologies in relation to information retrieval on dementia management issues.

Further sections of this paper are organized as follows. Section 2 summarizes the relevant studies in exploring the accuracy and effectiveness of the technologies providing medical information. Section 3 describes the research methodology and our approach, including the selection of clinical guideline questions, description around the selection of three technologies, the data collection and statistical analysis process. Section 4 presents the findings of the study and provides a detailed discussion of the results generated by using different technologies and compares the performance with clinical guideline. Section 5 addresses the limitations of this paper, and lastly, section 6 presents the conclusion.

2. Related Work

Over the last decade, the field of Artificial Intelligence (AI) and Large Language Models (LLMs) has been contributing to revolutionize different sectors, especially the healthcare domain[13][5]. LLMs[14][15] are a type of AI systems, based on the transformer model[16], designed to understand and generate human-like text based on the inputs[17]. The introduction of advanced AI models like LLMs and their models of transformer-based architectures[16] such as GPT (Generative Pre-trained Transformer) and BERT (Bidirectional Encoder Representation from Transformers), has fundamentally transformed how information is processed[5]. Compared to traditional methods of processing sequential data using Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN), which process the information in a sequential way, transformers have the capability to process entire sentences at once, focusing on the relationship between words. Transformers operate on the "self-attention"[16] mechanism to get the dependencies between different elements of the input, instead of the recurrence approach, which enables parallel data processing and obtains long-range dependencies in the input.

[1] https://www.ean.org/ l.a 05-03-2025

[2] https://www.ean.org/research/ean-guidelines l.a 05-03-2025

Additionally, transformers can handle short and long input sequences efficiently (parallel processing), compared to RNNs which are not very effective at capturing long-range dependencies[16].

GPTs, relying on the LLMs are trained on large data and can perform a variety of natural language processing tasks [15][18][17][5]. This has led to a rapid increase in the development of chatbots and Generative AI (GenAI) tools across various industries and sectors including the healthcare [8][14]. The AI-powered tools can provide basic medical information, answer patient questions, schedule appointments, translate one language to others, summarization of appointments with doctors and documenting medical records [7][13]. LLMs have been increasingly used in various applications like dementia [2], dentistry [17], medical administrative tasks, improving patient support, speeding drug discovery, analyzing patient data, customer service chatbots [14][15]. The ethical use of AI in healthcare still remains a critical challenge; however, the potential benefits for patients, medical professionals, caregivers is vast and yet to be discovered [14][15].

Sandmann S., et al.[7], evaluate the accuracy of LLMs, GPT-3.5 and GPT-4, in providing initial diagnosis, examination suggestions, and treatment recommendations of 110 medical cases across diverse clinical disciplines. It compares their performance with Google search results, where GPT-4 outperformed GPT-3.5 and Google in most tasks, especially in diagnosis and examining patient tasks. The authors reported cases where LLMs failed to provide accurate diagnoses and reported better performance of GPT-3.5 and GPT-4 on general diseases compared to rare ones, leaving room for improvement. Their findings point towards the potential of LLMs in the clinical decision-making process with essential further development to improve on the accuracy, particularly for rare diseases. The LLMs struggle in cases where less information is present such as for rare diseases and can only generate information which they have been trained on[14][15][7].

Ayoub et al.[13], explore the ability of ChatGPT to provide medical knowledge and recommendations in the healthcare setting. They use the Clinical Practice Guidelines (CPGs) as a reference for generating the patient-oriented questions about various medical conditions and assess the response from both ChatGPT and Google search through the Patient Education Materials Assessment Tool (PEMAT-P). The findings show that ChatGPT performs better than Google search in providing general medical knowledge and patient education with high scores for clarity and usability, while Google search results outperformed ChatGPT in providing accurate medical recommendations [13]. Although ChatGPT shows promise as a supplementary source of medical knowledge, it cannot fully replace professional healthcare knowledge [14]. The authors highlight the importance of healthcare providers understanding both the capabilities and limitations of AI tools like ChatGPT to optimize patient education.

A similar recent study, and more related to our work, has been conducted by Hristidis et al.[19] which provides the analysis of ChatGPT and Google in answering dementia-related and other queries pertaining to cognitive decline. The authors report that Google provided access to a vast array of sources, including up-to-date medical literature, but often presented less contextualized information. The findings presented also indicate that responses from Google were more diverse but often lacked precision, whereas responses generated from ChatGPT were structured and coherent; however, these were susceptible to inaccuracies due to potential biases and outdated training data. While ChatGPT demonstrated a stronger understanding of the topic in question, it compromises accuracy and credibility when compared to the standard clinical guidelines (SCG).

Table 1 presents the list of technologies considered for this work, their date accessed, version, their developer, the year developed, and their coverage statement.

Technology	Version	Year	Developer	Coverage
Google	Search Engine	1997	Larry Page and Sergey Brin	Google Search is a giant library for the internet. It scours websites, news articles, images, and videos for information, when a question is asked or term is searched on any topic. Google digs through this massive collection and presents the most relevant results.

ChatGPT	GPT-3.5	2020	OpenAI	Covers a range of topics, including but not limited to general knowledge, science, technology, literature, history, etc. It is designed to engage in natural language conversations and assist users with various inquiries and tasks.
ThinkAny	Claude 3 Haiku	NA	LMS Solution	ThinkAny AI is a search engine that focuses on understanding the users' questions and providing comprehensive answers. It can access and analyze information from various sources, automate mind maps to visualize connections, and potentially offer solutions or next steps.

Table 1: Technologies being considered.

3. Methodology

This research sought to determine whether ChatGPT-3.5, Google search, and AI search engine responses provide the appropriate information to specific topics on medical management issues in dementia. Our methodology starts with selecting three questions on dementia management taken from the **European Academy of Neurology guideline on medical management issues in dementia** [4]. Three questions were selected as they encompass a diverse range of medical management scenarios while providing care to people with dementia (PwD). They are a good sample among the available questions and cover a range of topics relevant to dementia management; the first question is related to systematic medical follow-up in dementia, whereas the second question is related to the management of vascular risk factors in dementia, and the third question is related to the management of pain in dementia. These three questions were queried across all aforementioned technologies (**Step 1**). At the second stage of our methodology, we used the Likert scale (**Step 2**) to evaluate the accuracy, clarity, relevance, simplicity, and usefulness of the responses provided by each technology. The systematic assessment of responses obtained from each technology was based on a straightforward 5-point Likert scale, for all three questions (details presented in section 3.3). Due to the fact that Likert scale ratings are more suitable for ordinal data (order of response matters), and the intervals between values may not be equal [9], which could lead to misinterpretation of results, two separate non-parametric tests namely 1) Friedman test (**Step 3**) and Wilcoxon signed-rank test (**Step 4**) were performed as presented in section 3.4.1 and 3.4.2 respectively.

3.1. Questions selected from the Guideline

For our experiments, following three questions were selected from the **European Academy of Neurology guideline on medical management issues in dementia** [4]:

Q1: *Should home-living (non-institutionalized) patients with dementia be offered systematic medical follow-up in a memory clinic setting?*
Q2: *Does systematic management of vascular risk factors in patients with dementia slow the progression of dementia?*
Q3: *Should behavioural symptoms in patients with dementia be treated with mild analgesics?*

3.2. Technologies used

To compare the responses for clinical guideline questions on medical management issues in dementia, we selected three state-of-the-art technologies: ChatGPT-3.5, Google search engine, and ThinkAny (AI search engine) to assess the performance of each technology on the three questions (**Step 1**). The rationale behind this selection is to choose three diverse systems including first is a LLM-based AI system, second a search engine, and third one an AI-based search engine. ChatGPT and ThinkAny are trained using large-scale datasets, including medical information, clinical guidelines, and healthcare

databases, among other information, whereas Google's vast search, combined with advanced algorithms of natural language processing capabilities, is widely recognized and used across various domains. ChatGPT, the current state-of-the-art in natural language processing and text generation, has been trained on vast amounts of data and has already shown potential in understanding and generating relevant responses. ThinkAny is an advanced and new-era AI search engine that can retrieve and aggregate high-quality content and can efficiently answer user questions. Each technology has its own unique strengths and capabilities, and our experiment offers valuable insights into their effectiveness in supporting medical management issues in dementia.

3.2.1. Querying Google

The three clinical guideline questions were sent to Google's search engine one at a time without any modifications to ensure a fair comparison between the standard response from the guideline and the search results. The first Google search link has the highest click-through rate of all search results, so the first result was selected in our case [13]. The information on these websites was evaluated for relevant responses to the question.

3.2.2. Querying ChatGPT-3.5 and ThinkAny

We used the original three clinical guideline questions directly from the guideline aiming to mirror normal usage and ensure a fair comparison with the Google search engine. We are running these questions a single time in this research, and the prompts were not revised after that. We are evaluating the results based on the first run. It was out of the scope of the paper to see the incremental improvement of prompts or how more accurate information can be obtained with references. It is widely recognized that prompt engineering or prompt refinements can increase the performance of the LLMs significantly[20]. In our prompt strategy, we have not provided any general context or added any extra details to help direct the technologies towards an answer.

All prompts were systematically executed between 25 April 2024 and 12 May 2024 through the website https://chatgpt.com/ and https://thinkany.ai/en-UK for ChatGPT-3·5 and ThinkAny, respectively.

Detailed overview of the workflow is pictorially illustrated in Figure 1.

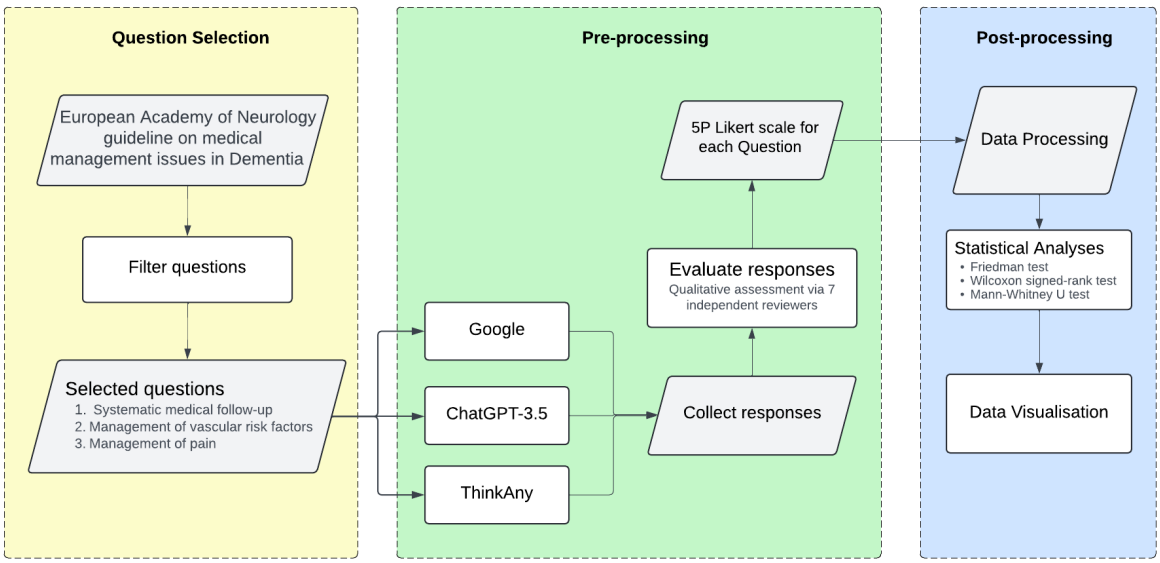

Figure 1: A flow chart depicting the methodology

3.3. Likert scale

The Likert scale is widely used to measure preferences or degrees of agreement with a statement or set of statements. We used the Likert scale to evaluate the accuracy, clarity, relevance, simplicity, and usefulness of the response provided by each technology. The systematic assessment of responses obtained from each technology was based on a straightforward 5-point Likert scale, ranging from 1 (Strongly Disagree) to 5 (Strongly Agree), for all three questions [9]. While this scale is used as a standard assessment scale by different clinicians and physicians, it is still a subjective instrument.

The responses gathered were evaluated by seven domain experts working in Dementia Research. Two of them were mid-career researchers, three were established researchers, and two entry-level researchers. These experts rated the three questions on the Likert scale, and the final score was calculated as the mean of the seven individual scores.

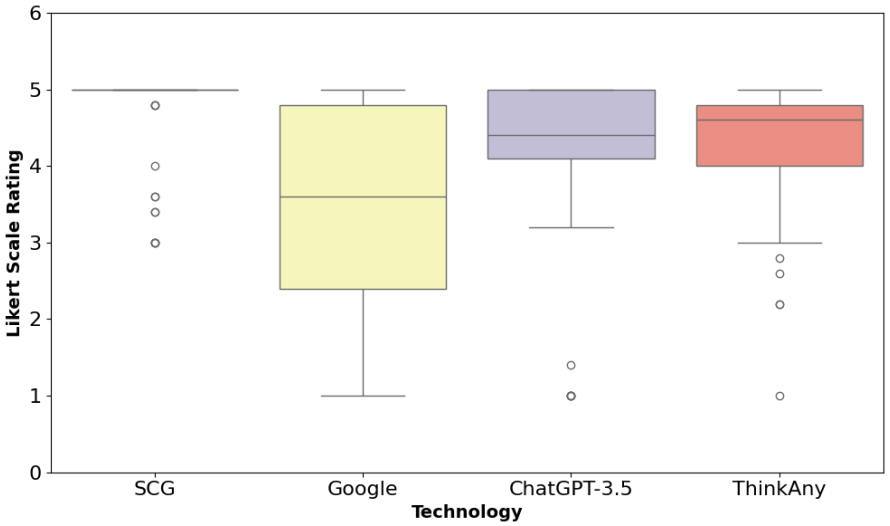

Figure 2: Comparison of Likert scale ratings for clinical guideline questions on medical management issues in dementia across different technologies: Guideline, Google, ChatGPT-3.5 and ThinkAny

Figure2 illustrates the distribution of Likert scale ratings assigned to the response of the three clinical guideline questions on dementia for Guideline, Google, ChatGPT-3.5, and ThinkAny. Each box in the plot represents the inter-quartile range (IQR) of ratings for the specific resource. The median is marked by the horizontal line inside each box plot. The flat line at level 5 for Guideline shows most reviewers rated it highly for all three questions, whereas Google received the whole spread of values from 1 to 5, with most values in the 2 to 5 range. Most scores for ChatGPT and ThinkAny by reviewers ranged between 4 to 5, with few outliers.

Figure3 displays the average Likert scale ratings for each question for each technology. It was observed that Google's ratings by reviewers for ***Q3*** were the lowest ratings received. The guideline received the highest ratings at an average of 4.5 ratings for all three questions. ThinkAny received better ratings compared to ChatGPT for the first two questions; however, ChatGPT's response obtained better scores for ***Q3*** compared to ThinkAny.

3.4. Statistical Analysis

To test if there is a significant difference in the responses provided by the clinical guideline, Google, and the AI tools for each question, we performed statistical analyses to compare the responses from the three technologies for the three questions. The Likert scale ratings are ordinal data (the order of response matters), and the intervals between values may not be equal [9], which could lead to misinterpretation of results. In order to deal with the ordinal data and potential violations of normality or homogeneity of variances assumptions, two non-parametric tests (Friedman test and Wilcoxon signed-rank test) were

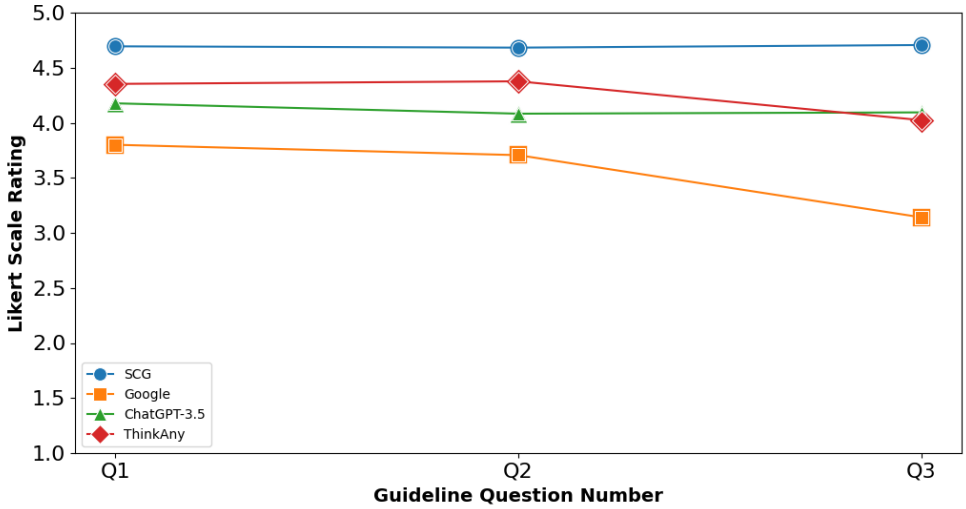

Figure 3: Comparative analysis of average Likert scale ratings for each clinical guideline questions on medical management issues in dementia across technologies: Guideline, Google, ChatGPT-3.5 and ThinkAny

performed as presented in section 3.4.1 and 3.4.2 respectively.

3.4.1. Friedman test

The Friedman test, a non-parametric analogue of repeated measures ANOVA, was conducted to evaluate the differences in mean ratings of the three technologies across the three clinical guideline questions [10]. This test is ideal for Likert scale data as we have multiple dependent groups (Clinical Guideline, Google, ChatGPT-3.5, ThinkAny) and a single independent variable (dementia clinical guideline question). It gives insight into whether there were statistically significant differences in the performance of each technology across the three questions considered.

If the Friedman test reveals a significant difference (p-value < 0.05), at least one pair of systems (Guideline vs Google, Guideline vs ChatGPT-3.5 or Guideline vs ThinkAny) has statistically different Likert scores, which the Friedman test does not pinpoint[10]. To pinpoint the technology, post-hoc tests were performed, and since for each question multiple comparisons were performed, we applied Bonferroni correction to adjust the p-value threshold for significance in the post-hoc tests. This was done to reduce Type 1 errors and account for the increased chance of false positives due to making multiple comparisons.

The Table 2 below presents the p-value obtained from the Friedman Test of the three technologies for the three clinical guideline questions on medical management issues in dementia. The three technologies, Google, ChatGPT-3.5, and ThinkAny, are compared with the standard clinical guideline result for each question.

Question	p-value
Q1	0.0511
Q2	0.0018
Q3	0.00047

Table 2: Friedman test p-value results for three clinical guideline questions.

3.4.2. Wilcoxon signed-rank test

Wilcoxon signed-rank test was applied to compare the mean ratings between the clinical guideline and each technology individually for each clinical guideline question. This non-parametric test was chosen

as it is suitable for paired data and is useful when analyzing differences between two technologies[11]. By conducting this test for each question separately, we can possibly identify the specific technology where its mean ratings differ significantly from the clinical guideline. It will provide us with valuable insights into the relative performance of each technology compared to the clinical guideline on a question-by-question basis. The Mann-Whitney U test was also considered to compare the two non-parametric tests. It compares the mean ratings between the clinical guideline and each technology individually for each clinical guideline question. Through this test, we can assess if there are statistically significant differences in mean ratings of each technology and guideline[12]. It will enable us to gather valuable insights into the overall performance of each technology. The three non-parametric tests will provide us with a comprehensive understanding of the effectiveness of these technologies in generating responses to clinical guideline questions on medical management issues in dementia.

All statistical analyses were conducted using Python 3.11.1, using the appropriate Python library (SciPy) and the significance level was set to α=0·05. The Python code to perform data processing, statistical analysis, and generate data visualizations has been provided via Github link: https://github.com/Sourav-rcsi/Clinical-Guidelines

4. Results and Discussion

To compare the responses obtained from three technologies: Google, ChatGPT-3.5, and ThinkAny with the standard clinical guideline, we first conducted a single comparison of one test per question. We used the Friedman test to check if there was any overall difference [10] in Likert ratings and if there was a statistically significant difference in the overall distribution of Likert scores across the technologies for each question. Then we conducted separate statistical tests for each question so we could analyze the performance of each source independently and draw meaningful conclusions. We wanted to test if there is a significant difference in the responses provided by the standard clinical guideline, Google, ChatGPT-3.5, and ThinkAny for each question.

We conduct the non-parametric Wilcoxon signed-rank test[11] and the Mann-Whitney U[12] test for independent samples on each individual question with the overall significance level (α=0.05). Since we are conducting multiple tests, we applied **Bonferroni correction**, and adjusted the significance level by dividing the original significance level (α=0.05) by the number of comparisons being made, in this case 3, which gives us the adjusted significance values of alpha as α=0.0167 for the two tests.

4.1. Statistical Tests

Table 3 presents the *p-value* obtained from the **Wilcoxon Signed-Rank test** of the three technologies for the three clinical guideline questions on medical management issues in dementia. The three technologies Google, ChatGPT-3.5, and ThinkAny are compared with the standard clinical guideline result for each question.

Question|Technology	Google	ChatGPT-3.5	ThinkAny
Q1	0.0107	0.0353	0.0317
Q2	0.0063	0.0166	0.0187
Q3	0.00096	0.0066	0.0053

Table 3: *p*-value of Wilcoxon Signed-Rank test for the three technologies across the three questions.

4.1.1. *Q1* Results

For ***Q1***, the Friedman test resulted in *p*-value of 0.051, close to the original alpha value but not significant, indicating that there is statistically no evidence that a significant difference in the mean responses among the three technologies exists. However, the borderline *p*-value indicates that the difference

could be worth examining further. The post-hoc analysis using the Wilcoxon signed-rank test revealed that individual comparisons between the clinical guideline and each technology: Google, ChatGPT-3.5 and ThinkAny performed statistically significantly differently from the clinical guideline response. Statistically significant differences were observed in the comparison between the clinical guideline and Google search results which resulted in a significant p-value of 0.0107 even after Bonferroni correction (adjusted alpha=0.0167), indicating that Google's response differed significantly from the clinical guideline. The comparison between the clinical guideline and ChatGPT-3.5, and clinical guideline and ThinkAny, were initially significant but failed to remain so after *Bonferroni correction.* The Mann-Whitney U test supports these findings, indicating significant differences before Bonferroni correction between the clinical guideline and each technology except for ThinkAny (p=0.0908).

4.1.2. *Q2* Results

For ***Q2***, Friedman test shows statistically significant differences among the responses of the three technologies (*p-value* = 0.0018), indicating a difference in performance across the technologies. Wilcoxon signed-rank test results showed that all the technologies differed significantly from the clinical guideline, with all p-values falling below the *Bonferroni*-adjusted significance level. The *Mann-Whitney U test* confirmed these differences with highly significant results, indicating differences in median scores and their distributions. Pairwise comparison using the Wilcoxon signed-rank test and Mann-Whitney U test showed significant differences between the clinical guideline and Google search results with p=0.0062 and p=0.0015 respectively, indicating that Google's response deviated significantly from the clinical guideline. The comparison between the clinical guideline and ChatGPT-3.5 as well as the clinical guideline and ThinkAny were initially significant but failed to remain so after Bonferroni correction.

4.1.3. *Q3* Results

For ***Q3***, the Friedman test indicated the most substantial differences among the responses of the three technologies (*p-value* = 0.0004), indicating significant differences in performance across the technologies. All technologies based on the Wilcoxon signed-rank test differed significantly from the clinical guideline and their p-values were exceptionally low, indicating strong evidence against the null hypothesis of no difference. The Mann-Whitney U test had similar findings, indicating significant differences in median scores and their distributions between the clinical guideline and each technology. There were significant differences between the clinical guideline and Google search results with p= 0.00097 and p=0.000083 for the two tests, indicating that Google's response deviated significantly compared to the clinical guideline. Google's performance was the worst (smallest p-value) among all the technologies across both tests for this question.

Table 4 presents the p-value obtained from the *Mann-Whitney U test* of the three technologies for the three clinical guideline questions on medical management issues in dementia. The three technologies Google, ChatGPT-3.5, and ThinkAny are compared with the standard clinical guideline result for each question.

Question|Technology	Google	ChatGPT-3.5	ThinkAny
Q1	0.0213	0.0198	0.0908
Q2	0.0015	0.0088	0.006
Q3	0.000083	0.0169	0.0011

Table 4: p-value of Mann-Whitney U test for the three technologies across the three questions.

Overall, the findings indicate that while there are significant differences in how each technology fares against the clinical guideline across the various questions, the extent of these difference varies. While the *Friedman tests* showed varying degrees of performance among the three technologies across the three clinical guideline questions. The *Wilcoxon signed-rank test* and *Mann-Whitney U test* showed significant

differences between Google and the clinical guideline answer for all three questions, indicating Google's response diverged significantly from the clinical guideline answer. Google showed the largest deviation from the clinical guideline, especially in ***Q3***. AI chatbot ChatGPT-3.5 and AI search engine, ThinkAny also showed significant differences but have closer performance metrics to the clinical guideline, suggesting they might be more reliable in contexts where adherence to clinical guidelines is critical. However, for ***Q2*** and ***Q3***, they showed statistically significant differences across the tests, suggesting that while these resources can provide useful information, they lack the accuracy, credibility, and depth provided by specialized clinical guidelines. These findings highlight the importance of resource selection in information gathering and decision-making processes, especially in situations where resource selection is essential.

4.2. AI tools

Figure4 and Figure5 illustrate the comparative analysis of *Wilcoxon Signed-Rank test* and *Mann-Whitney U test* for the three clinical guideline questions across technologies: Guideline, Google, ChatGPT-3.5, and ThinkAny

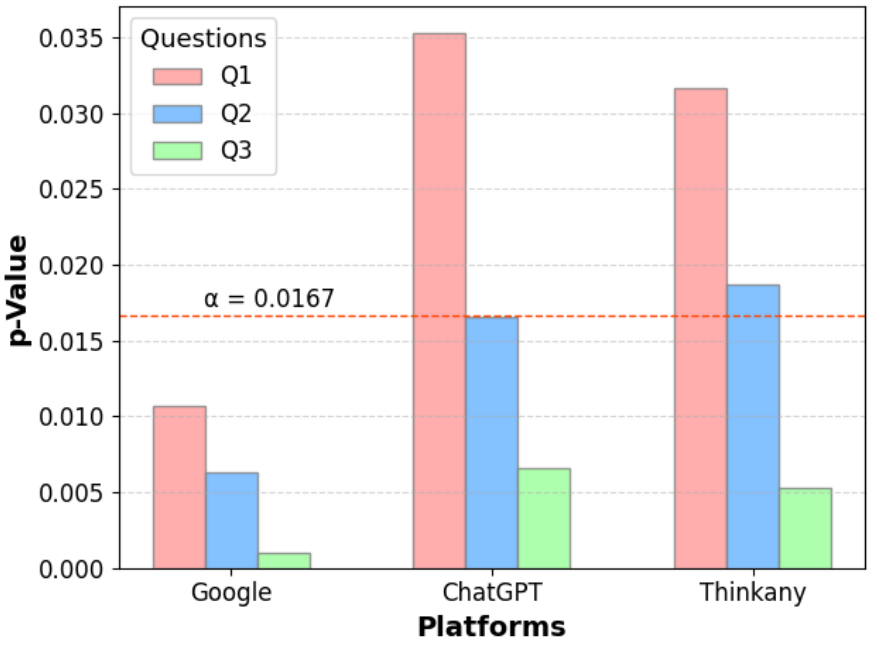

Figure 4: Comparative analysis of Wilcoxon Signed-Rank test with significance level (α=0.0167) for the three clinical guideline questions across the three technologies of Google, ChatGPT-3.5 and ThinkAny

4.2.1. ChatGPT-3.5 response to *Q1*

For ***Q1***, ChatGPT exhibits a thorough understanding of the question's context and recommendations even though it was not presented with a context before asking the question. The response presented uses simple language and avoids any complex medical jargon, is coherent, well-structured, and provides a summary at the end with clear explanations of the benefits associated with memory clinic settings for dementia care. It also provides a detailed analysis of the advantages of memory clinic settings like specialized care, care coordination, education, support and environment. The response recognizes the importance of personalized treatment and multidisciplinary team in memory clinics, which shows the ability of ChatGPT to understand the need for tailored plans and recognizes the importance of collaboration among various healthcare professions. The response also talks about the environment of memory clinics and telemedicine, indicating the importance of accessible healthcare for people with

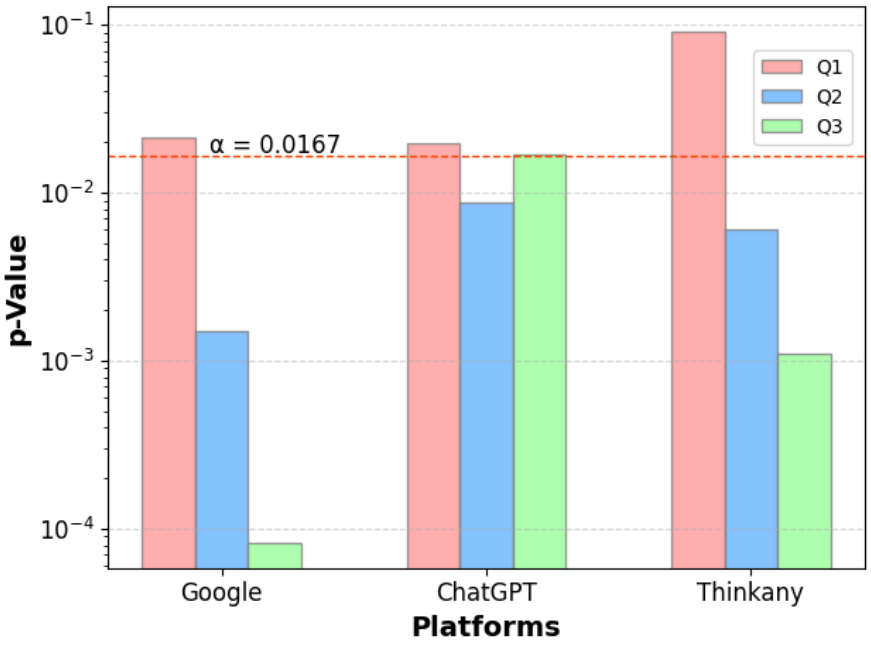

Figure 5: Comparative analysis of Mann-Whitney U test with significance level (α=0.0167) for three clinical guideline questions across technologies: Guideline, Google, ChatGPT-3.5 and ThinkAny

mobility issues or geographical constraints. All these factors show the comprehensive understanding of ChatGPT in providing dementia care, however, the reference or sources of the obtained information was not mentioned in its response.

4.2.2. ChatGPT-3.5 response to *Q2*

For ***Q2***, ChatGPT demonstrates understanding of the topic. The response talks about the significance of addressing vascular risk factors particularly in case of vascular dementia and how managing the risk factors such as hypertension, diabetes, high cholesterol, obesity, smoking, and lack of physical activity can help mitigate cognitive decline and promote brain health. It shows that ChatGPT understands the context of the question and provides a relevant response. ChatGPT acknowledges that dementia is a complex condition influenced by multiple factors and advocates for a comprehensive care strategy including medical management, cognitive stimulation, social engagement and support for caregivers. The response is accurate, clear, coherent, relevant and useful to the question, however, again the reference or sources of the obtained information was not mentioned in its response.

4.2.3. ChatGPT-3.5 response to *Q3*

For ***Q3***, the response generated by ChatGPT shows a comprehensive but cautious response compared to the other responses. The response emphasizes the assessment of pain, identification of underlying causes, non-pharmacological interventions, risk-benefit assessment, monitoring and reassessment and interdisciplinary collaboration. It highlights the importance of minimizing medication use while prioritizing individualized care and ongoing monitoring to optimize safety and effectiveness. This reflects a good understanding of the complexities involved in managing behavioural symptoms in dementia and the need for a patient-centered approach for treatment. Again, the reference or sources of the obtained information was not mentioned in its response.

4.2.4. ThinkAny response to *Q1*

For ***Q1***, ThinkAny provides an accurate, clear, coherent, relevant, well-structured and useful response. It provides a direct answer in comparison to ChatGPT, which provides slightly indirect answer. It provides clear and direct arguments like ChatGPT but are supported by credible and trustworthy references, however some of the references mentioned in its response is out-of-date and is not based on current dementia research. The response highlights the progressive nature of dementia and the importance of addressing complex medical needs of patients through regular follow-ups. It emphasizes the role of memory clinic settings in monitoring patient conditions, adjusting treatments and providing support to both patients and caregivers. The response acknowledges the significance of community-based services and care coordination in enabling patients to remain safely at home. ThinkAny's response effectively addresses the importance of systematic follow-up in memory clinic settings for dementia care.

4.2.5. ThinkAny response to *Q2*

For ***Q2***, ThinkAny effectively integrates information from various with credible and trustworthy sources and provides the references to them. It provides a direct answer in comparison to ChatGPT, which provides a slightly indirect answer. The response directly presents a logical argument which is accurate, clear, coherent well-structured, relevant and useful. It acknowledges the absence of a cure for dementia and presents the available interventions to manage symptoms and improve quality of life. The response presents a well-supported assessment of the potential impact of systematic management of vascular risk factors on slowing the progression of dementia and suggests that controlling vascular risk factors has the potential to slow the progression of dementia.

4.2.6. ThinkAny response to *Q3*

For ***Q3***, ThinkAny presents a balanced, accurate, clear, coherent, relevant and useful response, however, does not take a cautious approach like ChatGPT. It provides a direct answer in comparison to ChatGPT, which provides a slightly indirect answer. It again acknowledges the absence of a cure for dementia and uses a credible source to highlight the availability of interventions to manage the symptoms and improve quality of life, suggesting mild analgesics could potentially be used for this purpose. It mentions the use of mild analgesics as a recommended approach in caregiver strategies for addressing behavioral issues in dementia patients. It also emphasizes careful evaluation and monitoring is needed by healthcare professionals considering factors like the patient's overall condition, other medications and potential side effects. ThinkAny's response provides a well-rounded assessment on the use of mild analgesics for treating behavioral symptoms in dementia.

5. Limitations

There are limitations to this work. The assessment relies on subjective Likert scale ratings, which may introduce bias and variability in the results. It focuses on specific clinical guideline questions and may overlook the broader aspects of dementia management. Only the first Google search link was used for assessment, the first link does not represent the content in every link. Google search results vary based on the location, search history and other criteria, so each search may result in a different first link. The LLMs are constantly updated by their providers and are trained on current data, this can lead to different results if the questions are re-entered, which increases the variability in the response received and limits the reproducibility of the exact performance results. LLMs are trained on large datasets including medical and non-medical content, which may not be able to fully capture the medical context and guidelines, leading to gaps in their knowledge. The inner workings of GPTs often lack transparency, making it difficult to assess the reliability and accuracy of their responses to medical clinical guideline questions.

6. Conclusion

AI-powered chatbots and search engines can serve as an important resource in addressing medical management issues in dementia patients and meet the needs of healthcare professionals, patients and caregivers (formal and informal). Factors beyond accuracy, relevance such as those helpful in capturing the evolving nature of search and retrieval algorithms are needed. AI chatbots ChatGPT-3.5 and ThinkAny showed significant differences, but have closer performance metrics to the clinical guideline than Google, suggesting that while these resources can provide useful information, they lack the accuracy, credibility and depth provided by specialized clinical guidelines. However, comparison of AI tools with Google search engine suggests AI tools offer more contextualized and personalized responses, which might be more reliable in contexts where obtaining general, but personalized information is required. Our paper finds researcher, medical professional (like physician, nurse), formal and especially informal caregivers can benefit significantly from leveraging AI tools and AI search engines as companion tool to generate immediate answers for their day-to-day problems. This paper serves as a stepping stone towards understanding the role of AI in answering guideline questions on medical management issues in dementia. It paves the way for future research aimed at harnessing the full potential of AI tools and search engines in addressing the complex challenges posed by dementia and other medical conditions.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] E. Nichols, J. D. Steinmetz, S. E. Vollset, K. Fukutaki, J. Chalek, F. Abd-Allah, A. Abdoli, A. Abualhasan, E. Abu-Gharbieh, T. T. Akram, et al., Estimation of the global prevalence of dementia in 2019 and forecasted prevalence in 2050: an analysis for the global burden of disease study 2019, The Lancet Public Health 7 (2022) e105–e125.

[2] B. BT, J.-M. Chen, Performance assessment of chatgpt versus bard in detecting alzheimer's dementia, Diagnostics 14 (2024) 817.

[3] Z. Arvanitakis, R. C. Shah, D. A. Bennett, Diagnosis and management of dementia, Jama 322 (2019) 1589–1599.

[4] K. Frederiksen, C. Cooper, G. Frisoni, L. Frölich, J. Georges, M. Kramberger, C. Nilsson, P. Passmore, L. Mantoan Ritter, D. Religa, et al., A european academy of neurology guideline on medical management issues in dementia, European journal of neurology 27 (2020) 1805–1820.

[5] S. Panja, Information retrieval systems in healthcare: Understanding medical data through text analysis, in: Transformative Approaches to Patient Literacy and Healthcare Innovation, IGI Global, 2024, pp. 180–200.

[6] S. Gul, S. Ali, A. Hussain, Retrieval performance of google, yahoo and bing for navigational queries in the field of "life science and biomedicine", Data technologies and applications 54 (2020) 133–150.

[7] S. Sandmann, S. Riepenhausen, L. Plagwitz, J. Varghese, Systematic analysis of chatgpt, google search and llama 2 for clinical decision support tasks, Nature Communications 15 (2024) 2050.

[8] J. Li, A. Dada, B. Puladi, J. Kleesiek, J. Egger, Chatgpt in healthcare: a taxonomy and systematic review, Computer Methods and Programs in Biomedicine (2024) 108013.

[9] A. T. Jebb, V. Ng, L. Tay, A review of key likert scale development advances: 1995–2019, Frontiers in psychology 12 (2021) 637547.

[10] T. J. Cleophas, A. H. Zwinderman, T. J. Cleophas, A. H. Zwinderman, Non-parametric tests for three or more samples (friedman and kruskal-wallis), Clinical data analysis on a pocket calculator: understanding the scientific methods of statistical reasoning and hypothesis testing (2016) 193–197.

[11] R. F. Woolson, Wilcoxon signed-rank test, Wiley encyclopedia of clinical trials (2007) 1–3.

[12] P. E. McKnight, J. Najab, Mann-whitney u test, The Corsini encyclopedia of psychology (2010) 1–1.

[13] N. F. Ayoub, Y.-J. Lee, D. Grimm, V. Divi, Head-to-head comparison of chatgpt versus google search for medical knowledge acquisition, Otolaryngology–Head and Neck Surgery (2023).
[14] A. J. Thirunavukarasu, D. S. J. Ting, K. Elangovan, L. Gutierrez, T. F. Tan, D. S. W. Ting, Large language models in medicine, Nature medicine 29 (2023) 1930–1940.
[15] C. Peng, X. Yang, A. Chen, K. E. Smith, N. PourNejatian, A. B. Costa, C. Martin, M. G. Flores, Y. Zhang, T. Magoc, et al., A study of generative large language model for medical research and healthcare, NPJ Digital Medicine 6 (2023) 210.
[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).
[17] K. Giannakopoulos, A. Kavadella, V. Stamatopoulos, E. Kaklamanos, et al., Evaluation of generative artificial intelligence large language models chatgpt, google bard, and microsoft bing chat in supporting evidence-based dentistry: A comparative mixed-methods study., Journal of Medical Internet Research (2023).
[18] K. Chowdhary, K. Chowdhary, Natural language processing, Fundamentals of artificial intelligence (2020) 603–649.
[19] V. Hristidis, N. Ruggiano, E. L. Brown, S. R. R. Ganta, S. Stewart, Chatgpt vs google for queries related to dementia and other cognitive decline: comparison of results, Journal of Medical Internet Research 25 (2023) e48966.
[20] J. White, Q. Fu, S. Hays, M. Sandborn, C. Olea, H. Gilbert, A. Elnashar, J. Spencer-Smith, D. C. Schmidt, A prompt pattern catalog to enhance prompt engineering with chatgpt, arXiv preprint arXiv:2302.11382 (2023).