Issues in Information Systems
Volume 25, Issue 2, pp. 190-202, 2024

DOI: https://doi.org/10.48009/2_iis_2024_115

An avatar-based framework using large language models for assessing patient cognitive state

George Stefanek, *Purdue University Northwest, stefanek@purdue.edu*
Jacob DeMuth, *Purdue University Northwest, demuthj@pnw.edu*

Abstract

The goal of this project was to create an avatar-based framework that uses a large language model to communicate with patients in a natural conversational format for the purpose of doing a cognitive assessment. The framework includes a webpage featuring an avatar with computer-generated voice and a large language model to navigate a conversation through a series of standardized cognitive assessment questions while having the conversation feel natural to the patient. The two cognitive assessments include a mini-cog and mini-mental state exam. While generalized chat bots exist, determining if currently available large language models are adequate to perform these assessments and can redirect questions to complete an exam for patients that may have dementia would be useful for doing computer automated assessments on a more frequent basis without human interaction. Six large language models were compared and scored to determine which did best at performing this type of cognitive assessment. It was found that Chat GPT4 was best at performing cognitive assessments.

Keywords: machine learning, GPT, LLM, AI, avatar

Introduction

Modern day technology has allowed for innovations in providing patients with dementia assessments. The interaction between a patient and caregiver has begun to change from a traditional caregiver-patient conversation where the caregiver records results on paper to an electronic health record system where information is recorded electronically. More automated techniques for performing dementia assessments at home using mobile devices have been discussed by (Öhman, 2021). These techniques require the patient to answer questions using an electronic device but can be confusing to use especially as a person begins to develop early stages of dementia. As the population ages performing these types of assessments regularly and more frequently by caregivers will increase.

In 2017 researchers at Google and the University of Toronto developed what would become the modern-day transformer (Vaswani, 2017). The original idea with this model would be that it would serve to improve language translation. However, in modern times it has become adapted for large language models (LLMs) such as OpenAI's Chat GPT or Meta's Llama 2. The fast and successful development of large language models for chat applications has opened the use of these for various purposes such as medical applications where natural language conversations would be useful during an interaction.

The goal of this research was to determine the out of the box performance of six large language models for the purpose of performing an initial cognitive assessment exam (mini-cog) and an ongoing cognitive assessment (mini-mental state). A web-based framework was put together to facilitate the use of this technology by integrating a web-based avatar with computer-generated voice and a large language model.

Statement of the Problem

The need for assessing the elderly for their cognitive state and performing regular on-going cognitive assessments will increase as the population ages. Providing automated techniques that will be effective at assessing the elderly and cognitively impaired persons for their current cognitive state will be important at being able to do such assessments frequently without the need for caregiver interaction.

There is currently very limited research that uses large language models to communicate with patients for the purpose of doing a cognitive assessment in patients that show early stages of dementia. Additionally, using an avatar to communicate with a patient to do a cognitive assessment is novel. Using an avatar to communicate with dementia patients can have further applications. According to research papers done on the decay of the brain due to dementia, two of the big factors that cause the decay are social isolation and cognitive inactivity (Livingston, 2020). This research will address how effective current large language models are at performing cognitive assessments but can also lead to further work on a way to address social isolation and cognitive inactivity among those suffering from dementia.

An additional benefit of creating a chat bot that can perform a cognitive assessment of the elderly and dementia patients can reduce the burden on caretakers to perform these tasks. Creating an avatar-based framework that integrates with a large language model can also be helpful in providing a starting point for further research into the use of AI in the medical field with an emphasis on cognitive issues and the use of machine learning in conversational settings. By providing a way for dementia patients to communicate with others in this case an AI, they may be able to fulfill the need for interaction with others. This need for others is shown as a risk factor for making dementia more severe (Livingston, 2020).

Literature Review

The prediction of dementia using machine learning techniques has been ongoing for at least a half a decade. Literature regarding machine learning dementia prediction has been ongoing since 2017 with the work of So, Hooshyar, Park, and Lim with their paper exploring techniques for early dementia prediction utilizing machine learning (So, et al. 2017). In their paper they use data from dementia screenings dating from 2008-2013 and trained several models which included MLP, Random Forest, Logistic Regression, Naïve Bayes, and a few others. Their work helped to set the basis for future research into machine learning techniques regarding dementia prediction.

Modern day research into dementia prediction using transformers is still a new concept. Early works date back to 2022 with Agbavor and Liang (Agbabvor, 2022) and Llias and Askounis (Llias & Askounis, 2022). Llias and Askounis focused on taking transcripts from dementia patients and extracting features based on the words they used to train a machine learning model to predict dementia (Llias & Askounis, 2022). Agbavor and Liang focused their work on predicting early dementia using a person's voice (Agbabvor & Liang, 2022). They used a person's voice and utilized ChatGPT 3 to create embeddings. These embeddings were then fed into a machine learning model which would predict if a person had dementia. Agbavor and Liang note in the end of their paper that their work could be used in a "web application or a voice-powered app used at the doctor's office to aid clinicians in AD screening and early diagnosis (Agbavor & Liang, 2022)".

The idea of a chat bot designed for dementia patients has been mostly theoretical. Most papers such as "Care: A chatbot for dementia care" by Muller, Paluch, and Hasanat (Muller, Paluch & Hasanat, 2022) and "Conversational Affective Social Robots for Ageing and Dementia Support" by Lima, Wairagkar, Gupta, Rodriguez, Barnaghi, Sharp, and Vaidyanathan focus on what a chat bot for dementia patients could look like (Lima, et. al., 2021). These papers, while great foundations for future work, are theoretical and do not implement any of their findings.

Öhman's 2021 paper discusses the use of remote technology for providing dementia assessments (Öhman, 2021). The paper discusses the use of smart phones to allow patients to take dementia assessments from home, but the paper also points out several issues with this medium. It discusses the use of Cambridge's CANTAB system which can help provide assessments. This as well as a similar system from the National Health Institute known as NIH-TB are both available on mobile and computer devices.

Bouazizi, et al. (Bouazizi et al., 2022) described a method for detecting dementia from transcribed speech of patients. Their approach is different from the more conventional approach of using an LLM to detect if a patient is making coherent and meaningful sentences. Bouazizi, et al. used an approach using a Long Short-Term Memory (LSTM) neural network to detect whether the center of focus of a patient changes over time as the subject describes the content of the "cookie theft" image - an image used commonly for evaluating a person's cognitive ability. They divided the image into regions of interest, and identified in each sentence spoken the regions were talked about by the patient. They used a Long Short-Term Memory (LSTM) neural network to learn different patterns used by dementia patients to describe the content of the cookie theft image and used it to perform a 10-fold cross validation-based classification. Their experimental results on the Pitt corpus from the DementiaBank resulted in a 82.9% accuracy.

Li et al. (Li, et al., 2023) also used the "cookie theft" image to evaluate dementia in patients. They used a Transformer LLM (GPT-2) pre-trained on general English text and paired with an artificially degraded version of itself (GPT-D), to compute a ratio between the two models' perplexities on language from transcripts between cognitively healthy and impaired subjects. They looked at the extent to which the best-performing degraded model reflected linguistic anomalies known to occur in Alzheimer's dementia such as the use of higher frequency words and repetitiveness. They used three publicly available datasets: DB, ADReSS, and the Carolinas Conversation Collection (CCC). This technique generalized well to spontaneous conversations. GPT-D generated text with characteristics known to be associated with Alzheimer's Disease, demonstrating the induction of dementia-related linguistic anomalies. The best-performing cumulative pattern for the ADReSS training set resulted in an accuracy of 0.85.

Searle, Ibrahim, and Dobson (Searle, Ibrahim & Dobson, 2020) compared performance from numerous AI models for the classification of Alzheimer's Disease. The ADReSS dataset was used which had acoustically pre-processed and balanced datasets for the classification and prediction of Alzheimer's Disease and associated phenotypes through the modelling of spontaneous speech. Supplied textual transcripts were analyzed from the spontaneous speech dataset and performance compared across numerous models for the classification of Alzheimer's Disease versus controls and the prediction of Mental Mini State Exam scores. Transcripts were prepared into a PAR dataset with only participant utterances concatenated together into a single paragraph and a PAR+INV dataset with both participant and interviewer speech concatenated together. The models evaluated from the Scikit-Learn framework were Support Vector Machines (SVMs), Gradient Boosting Decision Trees (GBDT), and Conditional Random Fields (CRFs). The deep learning models evaluated from the 'transformers' library included BERT, RoBERTa and DistilBERT/DistilRoBERTa. The models were used as embedding layers for the PAR and PAR+INV datasets. It was found that the top performing models were the Term Frequency-Inverse Document Frequency (TF-IDF) vectorizer as input into a SVM model and the pre-trained Transformer based model `DistilBERT' used as an embedding layer into simple linear models. The test dataset scores were 0.81-0.82 across classification metrics. This project continues the work of using machine learning for cognitive assessment, however, it focuses on comparing against standardized tests instead of speech or other non-testing features. In previous work, researchers were only predicting what the Mini Mental State Exam score would be based on how patients would talk. This project combines both giving the exam as well as interpreting the results of the exam.

Research Questions and Scope of Study

The research question to be addressed in this research is "How well can a large language model, which is not fine-tuned, perform a dementia assessment?" Six large language models will be compared and scored for their effectiveness at performing a full mini-cog assessment and a mini-mental state assessment including how well these models respond and redirect questions to incorrect or vague answers. This evaluation will determine whether fine-tuning needs to be performed to enhance performance. The scope of the research included the creation of a web-based framework that integrated a web-based 3D avatar with computer-generated voice with each of six large language models that were to be evaluated.

Assumptions and Limitations

The research had the following assumptions: 1) The web portion of the project was to use PHP, HTML, and JavaScript, 2) models used in testing the implementation would come from either Hugging Face or OpenAI, 3) models tested and implemented were the latest version, 4) models showed some level of competence without fine tuning in interacting with users, 5) the web avatar and generated voice were as close to human as possible, and 6) the models were not to be fine-tuned or altered meaning they would be tested on out of the box performance. The research had the following limitations: 1) It was limited by the level of sophistication of modern large language models, 2) it was limited by the current level of infrastructure related to online chat bots, 3) it was limited by no available datasets with answers that dementia patients gave during standard assessments - this meant that testing required the researchers judgement to determine correct/incorrect and vague answers involving dementia patient answers.

Methodology

The scope of this study included: 1) the design of a webpage with an avatar, generated voice, and a user input field linked to a LLM to process input, and 2) testing and scoring six different LLM models against two standard mental assessments. The models were tested and scored based on a custom scoring system while the avatar-LLM framework was created prior to testing. The web-based framework was created in PHP, JavaScript, and CSS. PHP and JavaScript were used as the development language to accommodate the APIs used for components of the website. The reason for doing a web-based application is because a web-based application would be easier to implement on a mass market scale than a downloadable application.

APIs for both the model and video portion of the website were used to save time in creating the infrastructure. The website features a 3D avatar with a simulated human-like voice integrated with a large language model to provide the conversation portion of the website. The avatar and voice were added to make it easier and more appealing for patients to interact with the system and also to address the issue of social isolation that is noted as one of the many risk factors related to dementia (Livingston 2020). The avatar was generated through an AI photo generation service known as Midjourney that specializes in AI images. Midjourney was chosen due to the need for an avatar that was human like but also wasn't a real human to avoid any issues related to using someone's likeness. A generated avatar also allows for more specificity in terms of the overall look of the avatar.

The voice service and service to make the avatar appear to be speaking was done through D-ID. The reason for this was because D-ID was identified as one of the only services to provide real time video generation. Interaction with the large language model happens when the user inputs their query into a text field and presses a submit button. This button calls a function that queries the large language model via an API. The model found to be the most effective for cognitive assessment was ChatGPT 4 (Ray, 2023). The reason for this was that ChatGPT 4 was the most flexible to work with and provided a full Mini Cognitive Assessment and Mini Mental State Exam very well (see the Results section for comparison of LLMs). Once the user's

query was sent to ChatGPT 4, the model generated a response. This response was then captured and sent to another function that called the D-ID video service. From there, the D-ID service would take the large language model's text and convert it into audio as well as create a video of the avatar to appear as if it was speaking to the end user.

The second portion of this research included testing six large language models to see which of the six models would do the best at performing both a mini-cog and mini-mental state cognitive assessment and thereby be the best candidate for any future fine-tuning of the model. The research also included looking into what kinds of data sets would be needed for fine-tuning and what tasks the chat system would need to be able to achieve regarding daily conversation for brain exercises for any future studies of using such a chat bot as a conversational companion and aid for dementia patients.

The six models chosen for testing were Llama 2 70B Chat, ChatGPT 3.5, ChatGPT 4, Gemma 7B, Qwan 72 B, and Mistral 7B v0.2. The reason for these choices was due to the popularity of the models, their performance regarding standard scoring metrics, the companies who developed these models, API support, and overall ease of implementation. These models were tested against two standard cognitive tests. The first was a Mini Cognitive Assessment (Borson, 2000). The second was a Mini Mental State Exam (Folstein, 1975). The reason for choosing these two tests is that these are both common and standard tests when identifying dementia. Both of these tests have patients answering questions related to different functionalities such as memory or awareness and output a score based on how the patient answered. The Mini Cognitive Assessment is used as a quick method for testing patients and the Mini Mental State Exam is used for further in-depth testing and for repeated testing to track a patient's changes over time.

Table 1: Mini Cognitive Assessment Questions

Category	Questions
Three Word Registration	Please listen carefully. I am going to say three words that I want you to repeat back to me now and try to remember. The words are [select a list of words]. Please say them for me now.
Clock Drawing	I want you to draw a clock for me. First, put in all of the numbers where they go. Then set the hands to 10 past 11.
Three Word Recall	What were the three words I asked you to remember?

Table 2: Mini Mental State Exam Questions

Category	Questions
Orientation	What is the (year) (season) (date) (day) (month)?
Orientation	Where are we (state) (country) (town) (hospital) (floor)?
Registration	Name 3 objects and have the patient repeat the 3 objects
Attention and Calculation	Serial 7's OR spell world backward
Recall	Repeat the 3 previous objects
Language	Name a pencil and watch
Language	Repeat the following, "No ifs, ands, or buts"
Language	Follow a 3-stage command: "Take a paper in your hand, fold it in half, and put it on the floor
Language	Read and obey the following: CLOSE YOUR EYES
Language	Write a sentence
Language	Copy the shown design

LLM models were asked two different prompts. For the Mini Cognitive Assessment exam the prompt asked to the LLM was: *"I want you, [Model], to give me, the user, a Mini Cognitive Assessment. You are to act like my doctor and I will act like your patient. I want you to ask me questions and I will respond to them."* For the Mini Mental State Exam, the following prompt was used: *"I want you, [Model], to give me, the user, a mini-mental state examination. You are to act like my doctor and I will act like your patient. I want*

you to ask me questions and I will respond to them." The reason these were worded this way is due to the evaluation of prompts that can best invoke these tests in the LLM. Most modern models have filtering that makes it hard for them to provide medical assessments. The above prompts work in getting around these filters to have the models provide the assessment. These prompts also mostly got around the filtering for giving a verdict as to whether the user has or doesn't have dementia. Additionally, the prompts were further refined so that each question in the cognitive exam would be asked one at a time instead of all up front: *"I want you, [Model], to give me, the user, a mini-mental state examination. You are to act like my doctor and I will act like your patient. I want you to ask me questions and I will respond to them. I want you to ask me questions one at a time and I will respond to them."* For both tests, models were given three kinds of answers. These were correct or standard answers, fuzzy or not clear answers, and incorrect or nonstandard answers. Correct answers, in this case, mean answers that someone who doesn't have dementia would give in response to the assessment. Fuzzy answers were answers that weren't clear and could either be signs of dementia or could also be nothing at all. Finally, incorrect answers are answers that someone with dementia or cognitive issues would provide. The reason for having three different kinds of answers was to determine the full range and capabilities of the models out of the box. In total each model went through six trials. This number comes from the number of tests times the number of different kinds of answers. It should be noted that the answer categories were not mixed but done separately - meaning test 1 consisted of only correct answers, test 2 consisted of only fuzzy answers, and test 3 only consisted of incorrect answers.

There were two forms of scoring - quantitative and qualitative. Quantitative scoring gave the models one point per standard assessment question asked, a half point if the question asked wasn't standard but achieved the same objective or was similar in wording to a standard question, and zero points if the question wasn't a standard question. The Mini Cognitive scores were out of two because the Mini Cognitive features only two questions which include recalling three words and drawing a clock. The Mini Mental State Exam was scored out of eleven using similar rules. Answers were scored out of the standard questions asked not the questions in the actual standard assessment. An example of this would be if the model only asked the "recall three words" question and not the "clock drawing" question. This means answers would be scored out of one not two since only one standard question was asked. This was done to not punish the model twice for the same error. Models were given one point per answer handled correctly and zero points per answer handled incorrectly. Finally, models were scored on their overall assessment of the user. This was scored out of one with one point for correctly identifying if the user had dementia, a half point if there was not a diagnosis but the model recommended the user go see a health care professional, and zero points if the diagnosis was incorrect. The reason the scoring system was constructed this way was because smaller numbers would be easier to work with and to identify which questions were asked, meaning each question should be weighted equally. There was a similar approach for answers. The qualitative scoring system consisted of the researcher reading through each question, answer, and diagnosis and giving it a score out of ten as well as a brief description of overall thoughts related to their experience. This was done because there are certain things numbers simply can't capture. An example includes situations such as when the model would ask a nonstandard but good question, meaning that even though the question might be good to ask, the model was not rewarded for asking the question. This could result in a model being overlooked because the numbers say it is not the right fit. Another issue that a qualitative system resolves is overall user experience. User experience is important when interacting with chat bots but becomes even more important when working with users who have cognitive issues.

Results

The terminology used in this section is as follows:

Correct Answers*:* These are answers that someone who doesn't have dementia would give. ***Fuzzy Answers****:* These are answers that can be perceived as unclear or fuzzy such as answering the recall of three words question wrong but only because a word was swapped around or replaced. These also include

mixing correct and incorrect answers.
Incorrect Answers*:* These are answers that someone with dementia would give.
Model*:* The large language model being tested.
Question Points*:* The first number is the points given for each standard question asked and the second number is the number of questions in a standard mini cognitive assessment. Example: In the "Mini Cog Correct Answers", the Llama LLM generated question points shown are 1 out of 2. This means Llama asked 1 standard question from the mini cognitive assessment out of the 2 standard questions on a traditional mini cognitive assessment.
Answer Points: The first number is the points given when the model correctly handles the answer from the user and the second number is the total points possible given the number of standard questions asked previously. Example: In the "Mini Cog Correct Answers", since Llama only asked one standard question, it is only graded on how it addresses the answer to that one question. To elaborate, if Llama asks the user to recall three words, and the user recalls the three words, Llama is given a point if it moves onto the next question or gives feedback indicating the answer from the user is correct.
Assessment Points: The first number is the points given for correctly identifying whether the patient has dementia and the second number being for the assessment itself. Example: In the "Mini Cog Correct Answers", Llama correctly identified the user didn't have dementia, so they were given one point.
Note**:** Notice that in each three-table set below, the total points may sometimes vary. The reason for the variance is because when querying a large learning model, answers from the model aren't always the same every time. Their answers can be similar to each other, but the wording may be different.

Mini Cognitive Assessment Exam

The following are the scores from quantitative testing:

Table 3: Mini Cognitive Assessment: Patient Gives the Correct Answers to the Questions Asked. †

Model	Question Points	Answer Points	Assessment Points
Llama-2-70B-Chat	1 out of 2	1 out of 1	1 out of 1
Chat GPT 3.5	1 out of 2	1 out of 1	1 out of 1
Chat GPT 4	1 out of 2	1 out of 1	1 out of 1
Gemma-7B*	0 out of 2	n/a	1 out of 1
Mistral-7B-v0.2	2 out of 2	1 out of 2	1 out of 1
Qwen-72B-Chat	0.5 out of 2	1 out of 1	1 out of 1

*Note: The reason Gemma still received assessment points is because it still gave the user a test like a standard mini cognitive test and determined the user did not have dementia.

Table 4: Mini Cognitive Assessment - Patient Gives Partially Incorrect or Unclear Answers to the Questions Asked. †

Model	Question Points	Answer Points	Assessment Points
Llama-2-70B-Chat	0.5 out of 2	0 out of 1	0 out of 1
Chat GPT 3.5	1 out of 2	0 out of 1	0 out of 1
Chat GPT 4	1 out of 2	1 out of 1	0.5 out of 1
Gemma-7B*	0 out of 2	n/a	0 out of 1
Mistral-7B-v0.2	2 out of 2	2 out of 2	0.5 out of 1
Qwen-72B-Chat	0.5 out of 2	1 out of 1	1 out of 1

*Note: The reason Gemma still received assessment points is because it still gave the user a test like a standard mini cognitive test and determined the user did not have dementia.

Table 5: Mini Cognitive Assessment - Patient Gives Incorrect Answers to the Questions Asked. †

Model	Question Points	Answer Points	Assessment Points
Llama-2-70B-Chat	0 out of 2	n/a	0 out of 1
Chat GPT 3.5	0.5 out of 2	0 out of 1	1 out of 1
Chat GPT 4	1 out of 2	1 out of 1	0.5 out of 1
Gemma-7B*	0 out of 2	n/a	0 out of 1
Mistral-7B-v0.2	2 out of 2	2 out of 2	0 out of 1
Qwen-72B-Chat	0.5 out of 2	0 out of 1	0.5 out of 1

*Note: The reason Gemma still received assessment points is because it still gave the user a test like a standard mini cognitive test and determined the user did not have dementia.

Summary of Mini Cognitive Assessment Results

Table 6: Mini Cognitive Assessment - Average Score (Displayed as a percentage) †

Model	Question Points	Answer Points	Assessment Points
Llama-2-70B-Chat	25%	50%	~33%
Chat GPT 3.5	~42%	~33%	~66%
Chat GPT 4	50%	100%	~66%
Gemma-7B	0%	n/a	~33%
Mistral-7B-v0.2	100%	~83%	50%
Qwen-72B-Chat	25%	~66%	~83%

†Note. a) All percentages are rounded unless stated otherwise. Rounding was done using standard rounding rules, b) The best in a category is represented in blue, and c) the worst in a category is represented in red.

Table 7: Mini Cognitive Assessment - Average Score (Ranked) †

Model	Question Points	Answer Points	Assessment Points
Llama-2-70B-Chat	**T4th**	**4th**	**T4th**
Chat GPT 3.5	**3rd**	**5th**	**T2nd**
Chat GPT 4	**2nd**	**1st**	**T2nd**
Gemma-7B	**6th**	**6th**	**T4th**
Mistral-7B-v0.2	**1st**	**2nd**	**3rd**
Qwen-72B-Chat	**T4th**	**3rd**	**1st**

†Note. a) The best in a category is represented in blue, b) the worst in a category is represented in red, and c) the "T" prefix corresponds to a tie.

Mini Mental State Exam

Table 8: Mini Mental State Exam – Patient Gives the Correct Answers to the Questions Asked. †

Model	Question Points	Answer Points	Assessment Points
Llama-2-70B-Chat	2 out of 11	1 out of 1	1 out of 1
Chat GPT 3.5	6 out of 11	6 out of 6	1 out of 1
Chat GPT 4	10 out of 11	10 out of 10	1 out of 1
Gemma-7B	2 out of 11	2 out of 2	1 out of 1
Mistral-7B-v0.2	3 out of 11	3 out of 3	1 out of 1
Qwen-72B-Chat	7 out of 11	7 out of 7	1 out of 1

Table 9: Mini Mental State Exam – Patient Gives Partially Incorrect or Unclear Answers to the Questions Asked. †

Model	Question Points	Answer Points	Assessment Points
Llama-2-70B-Chat	1 out of 11	0 out of 1	0 out of 1
Chat GPT 3.5	9 out of 11	6 out of 9	0.5 out of 1
Chat GPT 4	7 out of 11	5 out of 7	0.5 out of 1
Gemma-7B	2 out of 11	0 out of 2	0 out of 1
Mistral-7B-v0.2	3 out of 11	3 out of 3	1 out of 1
Qwen-72B-Chat	2 out of 11	1 out of 2	0.5 out of 1

Table 10: Mini Mental State Exam – Patient Gives Incorrect Answers to the Questions Asked. †

Model	Question Points	Answer Points	Assessment Points
Llama-2-70B-Chat	3 out of 11	1 out of 3	0 out of 1
Chat GPT 3.5	9 out of 11	0 out of 9	0.5 out of 1
Chat GPT 4	9 out of 11	4 out of 9	0.5 out of 1
Gemma-7B	2 out of 11	0 out of 2	0 out of 1
Mistral-7B-v0.2	3 out of 11	3 out of 3	0.5 out of 1
Qwen-72B-Chat	4 out of 11	3 out of 4	0.5 out of 1

Summary of Mini Mental State Exam Results

Table 11: Average Score (displayed as a percentage) †

Model	Question Points	Answer Points	Assessment Points
Llama-2-70B-Chat	~18%	40%	~33%
Chat GPT 3.5	~72%	50%	~66%
Chat GPT 4	~79%	~73%	~66%
Gemma-7B	~18%	~33%	~33%
Mistral-7B-v0.2	~27%	100%	~83%
Qwen-72B-Chat	~39%	~85%	~66%

†Note. a) The best in a category is represented in blue, and b) the worst in a category is represented in red.

Table 12: Average Score (Ranked) †

Model	Question Points	Answer Points	Assessment Points
Llama-2-70B-Chat	T5th	5th	T3rd
Chat GPT 3.5	2nd	4th	T2nd
Chat GPT 4	1st	3rd	T2nd
Gemma-7B	T5th	6th	T3rd
Mistral-7B-v0.2	4th	1st	1st
Qwen-72B-Chat	3rd	2nd	T2nd

†Note. a) The best in a category is represented in blue, b) the worst in a category is represented in red, and c) the "T" prefix corresponds to a tie.

Qualitative scoring system

Llama-2-70B-Chat – 6th

Overall, the Llama 2 model seemed to not ask the right questions and would sometimes forget what it was set to do. This would cause tests to continue for a long time with similar kinds of questions being asked. The researcher would sometimes have to interrupt Llama to end the assessment and receive a diagnosis.

ChatGPT 3.5 – 2nd

ChatGPT 3.5 is a top three choice when looking at which model should serve as the backbone for the chat bot. Chat GPT does a great job of asking the right questions as well as labeling each section, so the user

knows what kind of question they are being asked. The only downside is that not all questions were asked, and questions could be slightly inconsistent per test.

ChatGPT 4 – 1st

Chat GPT 4 would be the best choice for the model to serve as the driving force of the chat bot. Chat GPT 4 is the latest model from open ai and while Chat GPT 3.5 performed better in certain categories overall Chat GPT 4 performed the best. It handled each situation appropriately.

Gemma-7B – 4th

Gemma quantitatively performed nearly the worst with the questions not necessarily being bad but rather not what was expected for a standard assessment. While close with Mistral in score, Gemma outperformed Mistral in how it displayed the questions and interacted with the researcher.

Mistral-7B-v0.2 – 5th

Mistral seemed to ask good questions however it seemed to mix up Mini Cognitive questions with Mini Mental State Exam questions. The one edge I would give it over models like gemma or Llama is that it did explain each question's answer and would state if it was correct, incorrect, or in between. It also seemed to do well in the Mini Cognitive portion.

Qwen-72B-Chat – 3rd

Qwen is a model that is in contention for future fine tuning. Qwen performed near the upper half of results and feedback during interactions felt simple. Qwen also did a decent job of answer analysis.

Discussion

Table 13: Scoring Summary Ranked†

Model	Question Points	Answer Points	Assessment Points
Llama-2-70B-Chat	T5th	5th	T3rd
Chat GPT 3.5	2nd	4th	T2nd
Chat GPT 4	1st	3rd	T2nd
Gemma-7B	T5th	6th	T3rd
Mistral-7B-v0.2	4th	1st	1st
Qwen-72B-Chat	3rd	2nd	T2nd

†Note. a) The best in a category is represented in blue, b) the worst in a category is represented in red, and c) the "T" prefix corresponds to a tie.

Table 14: Scoring Summary by Percent†

Model	Question Points	Answer Points	Assessment Points
Llama-2-70B-Chat	~18%	40%	~33%
Chat GPT 3.5	~72%	50%	~66%
Chat GPT 4	~79%	~73%	~66%
Gemma-7B	~18%	~33%	~33%
Mistral-7B-v0.2	~27%	100%	~83%
Qwen-72B-Chat	~39%	~85%	~66%

†Note. a) The best in a category is represented in blue, and b) the worst in a category is represented in red.

The results from these findings indicate that out of the box large language models can provide both a Mini Cognitive Assessment and Mini Mental State Exam however significant work will need to be done to have the models perform at an acceptable level to be used in a non-academic setting. While models were able to

ask either standard questions or questions like standard questions their interpretation of answers would sometimes be wrong, or the models could be tricked into believing a wrong answer was the right answer. An example is being able to lie about which three words or objects the patient was told to remember. Models could be tricked by the researcher into believing that the three words the researcher gave were correct answers.

While this didn't happen every time it is still an issue that will need to be addressed in some capacity. One common issue that the models had that will need to be either circumvented or a solution will need to be created is that for a standard Mini Mental State Exam the patient is supposed to have their level of consciousness noted and considered when giving a diagnosis. This would be hard for a model to achieve without some form of facial recognition software or sentiment analysis.

Another issue related to answers was that models would rarely restate the question if the researcher answered wrong or stated they weren't sure what the answer was. This could be perceived as a non-issue, however certain questions such as the name 3 objects question from the Mini Mental State Exam require the assessment giver to repeat the question until the patient repeats the objects. They are to also count the number of times this takes. This means that a solution would be needed for this issue in its current form.

Future work would also have to go into providing the assessments in a way that doesn't overwhelm the user as some models would ask all the questions at once without allowing the user to answer each question one at a time. Some models, such as the GPT models, would allow the user to answer one at a time but would still give all the questions at once while occasionally realizing the user wanted the questions one at a time. We were able to get around this by adding additional information in the prompt to have it ask questions one at a time (see the Methodology section).

Models also were found to not have consistent questions for each test. While most questions asked were similar across each portion of testing there were some slight variations in either questions asked or wording of questions. This is expected from large language models as they are generating new answers from user input each time.

More work will need to be done to make the models more consistent in the questions they ask. Models also seemed to struggle with certain kinds of questions related to drawings, however this is not entirely their fault. A standard question in the Mini Cognitive Assessment is to draw a clock and a standard question in the Mini Mental State Exam is to have the patient draw a design. Both questions require something that, while theoretically possible, would be hard to do. For the models to understand if a drawing is good or bad, they would need to first be able to see the image followed by being able to determine if the image is good or not.

Conclusion

Future work will need to be done in creating a dataset for fine-tuning the top LLM model. One limitation that was discovered during early planning stages was that it was hard to find datasets related to the objective of the research. While medical datasets exist, they were datasets related to broad medical questions and lacked the specificity needed for a fine-tuned dementia chat bot. A dataset will need to be created consisting of questions to ask for each kind of assessment. This dataset could feature the same question several times but slightly rephrased with the end goal of eliminating the model missing standard questions. This dataset could be assembled by hand or depending on the dataset creator's resources could be done using AI.

Another dataset may be needed to help the models overcome user variability in the user's response. While it is near impossible to create a dataset with every possible answer to a question on an assessment, it is possible to leverage current large language models to understand when an answer is not sufficient. This was evidenced through testing with the large language models where in certain circumstances, the models

correctly identified that an answer was not sufficient or understood that they needed to ask the question again. This means that a dataset could be created consisting of examples of questions, answers, and follow-up actions which are labeled as wrong or right.

Regarding what questions are asked, ideally the chat bot should ask standard assessment questions. Problems arise with certain types of questions that ask the user to follow a certain command or draw a certain design. These types of questions will either need to be replaced or the model will need to be trained or supported with computer vision to be able to interpret whether a command is followed, or a design is acceptable. The use of computer vision also introduces both the need for an additional dataset of drawings for the model to be trained on but also works on the assumption that every user has a camera or some method to upload their photo. While it is possible to use computer vision it may be more efficient to simply replace the questions.

Once the dataset is assembled, a model will need to be fine-tuned. The current recommendation based on the researcher's findings is that future work should utilize ChatGPT 4 or the latest version of the ChatGPT model. While other models and companies shouldn't be ruled out, ChatGPT4 in its current form did reasonably well at performing the cognitive assessment exams. Fine-tuning the model can be done in python or through a paid service. The benefit of a paid service is that it would require less coding as fine-tuning services ask for your data and some parameters and would not require you to find a third party to supply GPUs for training. The downside, and what plays into the strength of training from scratch, is that you lose out on flexibility when you code the fine-tuning yourself. By coding the fine-tuning program by hand, you have more flexibility and customization, but at the expense of programmer development time and GPU processing cost.

Finally, once the model is trained and ready to be deployed it should be used in place of the current non-fine-tuned LLM. Future work for the web portion would include a more realistic avatar as well as more realistic facial movements as the technology progresses. It is also recommended that future work on the web portion should include input via speech using speech to text translation. This would allow users who aren't as comfortable with a keyboard or can't use a keyboard to be able to still interact with the chat bot.

The last part of the recommendation portion of this research is a bit more speculative but will hopefully be useful in future work. The future dementia chatbot should go beyond doing a cognitive assessment, but also interact with patients to stimulate the brain through a mix of casual conversation and game playing. The future dementia chat bot should also be able to repeatedly give users frequent assessments and track their cognitive state without the need for frequent caregiver visits. The current stage of research is simply to have the large language models give an initial assessment with this initial assessment serving as a baseline for future assessments. The future chat bot will need to be able to either independently or with minimal human interaction recognize when a patient is getting worse than expected and either take corrective actions or inform the necessary caregiver.

This research has created a framework that uses a 3D avatar-based chatbot that integrates an out-of-the-box LLM to help users interact with the chatbot in a more natural, conversational way. The research evaluated several models for performing cognitive assessments and helped to establish which models should be further fine-tuned for this purpose. The implications of this work are that it may be possible to replicate some human-to-human interaction and evaluation in a medical setting using AI chatbots.

References

Agbavor, F., & Liang, H. (2022). Predicting dementia from spontaneous speech using large language models. *PLOS Digital Health*, *1*(12), e0000168.

Borson, S., Scanlan, J., Brush, M., Vitaliano, P., & Dokmak, A. (2000). The mini-cog: a cognitive 'vital signs' measure for dementia screening in multi-lingual elderly. *International journal of geriatric psychiatry*, *15*(11), 1021-1027.

Bouazizi, M., Zheng, C., Yang, S., & Ohtsuki, T. (2023). Dementia detection from speech: what if language models are not the answer?. Information, 15(1), 2.

Folstein, M. F., Folstein, S. E., & McHugh, P. R. (1975). "Mini-mental state": a practical method for grading the cognitive state of patients for the clinician. *Journal of psychiatric research*, *12*(3), 189-198.

Li, C., Knopman, D., Xu, W., Cohen, T., & Pakhomov, S. (2022). Gpt-d: Inducing dementia-related linguistic anomalies by deliberate degradation of artificial neural language models. arXiv preprint arXiv:2203.13397.

Lima, M. R., Wairagkar, M., Gupta, M., y Baena, F. R., Barnaghi, P., Sharp, D. J., & Vaidyanathan, R. (2021). Conversational affective social robots for ageing and dementia support. *IEEE Transactions on Cognitive and Developmental Systems*, *14*(4), 1378-1397.

Ilias, L., & Askounis, D. (2022). Explainable identification of dementia from transcripts using transformer networks. *IEEE Journal of Biomedical and Health Informatics*, *26*(8), 4153-4164.

Livingston, G., Huntley, J., Sommerlad, A., Ames, D., Ballard, C., Banerjee, S., ... & Mukadam, N. (2020). Dementia prevention, intervention, and care: 2020 report of the Lancet Commission. *The Lancet*, *396*(10248), 413-446.

Müller, C., Paluch, R., & Hasanat, A. B. M. (2022). Care: A chatbot for dementia care.

Öhman, F., Hassenstab, J., Berron, D., Schöll, M., & Papp, K. V. (2021). Current advances in digital cognitive assessment for preclinical Alzheimer's disease. Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring, 13(1), e12217.

Ray, P. P. (2023). ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*.

Searle, T., Ibrahim, Z., & Dobson, R. (2020). Comparing natural language processing techniques for Alzheimer's dementia prediction in spontaneous speech. arXiv preprint arXiv:2006.07358.

So, A., Hooshyar, D., Park, K. W., & Lim, H. S. (2017). Early diagnosis of dementia from clinical data by machine learning techniques. *Applied Sciences*, *7*(7), 651.

U.S. Bureau of Labor Statistics. (2024, April 17). *Home Health and Personal Care Aides : Occupational outlook handbook*. U.S. Bureau of Labor Statistics. https://www.bls.gov/ooh/healthcare/home-health-aides-and-personal-care-aides.htm

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Wittenberg, R., Hu, B., Barraza-Araiza, L., & Rehill, A. (2019). Projections of older people living with dementia and costs of dementia care in the United Kingdom, 2019–2040. *London: Care Policy and Evaluation Centre, London School of Economics and Political Science*, 79.