Multi-Agent LLM Alzheimer's Disease CDR Prediction

1st Kuo Gong
Dept. of Electrical Engineering
Columbia University
New York, USA
kg3175@columbia.edu

2nd Feiyang Chen
Dept. of Electrical Engineering
Columbia University
New York, USA
fc2795@columbia.edu

3rd Humayun Syed
Dept. of Electrical Engineering
Columbia University
New York, USA
hhs2128@columbia.edu

Abstract—**The integration of large language models (LLMs) in healthcare has opened new avenues for enhancing data processing and predictive analytics. This project investigates the use of multi-agent LLM systems to preprocess electronic health records (EHR) and predict Alzheimer's Disease Clinical Dementia Ratings (CDR). Leveraging semi-structured datasets from Taiwan Hospital, we develop an automated pipeline utilizing advanced prompt engineering and tools like Google Cloud, Spark, and Airflow. Our methods include translation of mixed-language data, denoising, and structured analysis using both single-agent and multi-agent LLM frameworks. Experiments demonstrate that multi-agent systems outperform single-agent approaches, with accuracy increasing from 88% (GPT-4) to 91%. To make sure users without a computer background can utilize the model easily, we build a front-end website with Flask and Google Cloud Storage, where imagery charts and word cloud graphs are displayed. Our work highlights the potential of LLMs to enhance healthcare decision-making through improved EHR preprocessing and predictive modeling, setting a foundation for future research in multi-agent systems for clinical applications.**

I. Introduction

With the rapid advancements in artificial intelligence (AI) and natural language processing (NLP) the new machine learning models especially the ones based on transformer architecture have provided powerful tools for healthcare research and understanding Natural language health records. Large language models (LLMs) show remarkable capabilities in understanding and generating human-like language. Beyond their traditional use in conversational AI, LLMs are increasingly being used in specialized domains like clinical decision support, medical documentation, and diagnostic predictions. This study investigates the application of LLMs to preprocesses and detect Alzheimer's disease. In this study, we are trying to solve a classification problem, where we try to predict the labels for data points. The labels correspond to the 4 numbers of CDR levels. The Clinical Dementia Rating (CDR) is a scale used to evaluate the severity of dementia symptoms in individuals.

In our work, we are focusing on dealing with raw semi-structure NLP Alzheimer's Disease EHR. We are using LLM's agents to preprocess and denoise the data we have for Data with Alzheimer's Disease. We use prompt engineering skills to handle dirty and incomplete data. After cleaning the data, we test the performance of different models in predicting the Alzheimer's Disease EHR.

Our work is important for the future of Alzheimer's Disease EHR data preprocessing and its early detection. We show an efficient prompt skill of LLM to preprocess the Alzheimer's Disease EH, and then test different models to predict the CDR Score of Alzheimer's Disease EHR. The Work gives a guideline for model selection and Alzheimer's Disease EHR Preprocessing.

A. High level overview of process

We utilize Google Cloud, Airflow, and Spark to build a preprocessing pipeline. Additionally, we leverage GPT-4o for preprocessing tasks, including translating text from Chinese to English.

For our prediction task we explore the feasibility of using Multi Agent LLMs to predict the risk of Alzheimer's disease. The key components of our prediction study are as follows:

- **Baseline**: Using Data to train the BERT model, predict Test Data
- **BioBERT**
- **LLM-based Agent:** Single agent, Multi agents

II. Related Works

A. Traditional ML Models on EHR data

Research has been done using the Traditional ML models on medical data:

Li et al. [1] trained data-driven Gradient Boosted Trees (GBT) and expert-knowledge-driven GBT to predict Alzheimer's Disease (AD) risk based on electronic health record (EHR) data, including demographic data, diagnoses, and drug codes.

Prabhu et al. [2] explored the performance of deep neural networks, such as Multilayer Perceptrons (MLP) and Denoising Autoencoders (DAE), in extracting CN/MCI/AD symptoms from EHR data.

Wang et al. [3] utilized Long Short-Term Memory (LSTM) networks to analyze EHR clinical notes for detecting early evidence of MCI/AD, enabling healthcare professionals to identify and intervene with potential patients in advance.

However, these studies primarily focus on well-structured and cleaned data, which often require human-expert-identified risk factors. This approach can be costly for physicians and involve lengthy processes.

The introduction of large language model (LLM) Multi-Agent systems enhances user autonomy, reducing workflow complexity. These systems also facilitate communication among agents, leading to improved task coordination. This paper discusses the advantages and disadvantages of employing LLM Multi-Agent systems, providing a deeper understanding of their applications and potential risks. Xu et.el [4]

B. LLM for EHR Analysis

Research utilizing LLMs for analyzing EHR notes includes the following:

Feng et al. [5] integrated textual EHR analysis with GPT and image analysis, demonstrating the capability of LLMs to process multimodal data and enhance detection accuracy.

Chen et al. [6] proposed a patient-level transcript profiling framework that leverages LLM-based reasoning augmentation to systematically extract linguistic deficit attributes. These attributes are embedded and fed into a BERT model, improving detection performance. Their work highlights the reasoning power of LLMs in extracting and quantifying patient linguistic data.

Du et al. [7] evaluated GPT's performance with various prompts for detecting AD from clinical records. They observed that while simple LLMs may not outperform traditional ML models in isolation, combining LLMs with other models can yield optimal accuracy.

C. Multi-Agent LLM for Prediction

Guo et al. [8] Multi-agent LLM refers to a framework where multiple large language models (LLMs) interact and communicate with one another to accomplish complex tasks. This approach is particularly advantageous for scenarios that require multi-stage reasoning, as each agent can specialize in distinct sub-tasks or perspectives, collectively enhancing the system's overall problem-solving capacity.

Previous research has explored the potential of multi-agent LLMs in various domains, such as embodied manipulation, collaborative problem solving, and decision-making. For example, in embodied manipulation tasks, multi-agent LLMs can simulate agents collaboratively controlling robotic systems to perform tasks in physical environments.

Despite these advancements, most existing work has focused on domains that are either physically interactive or strategically adversarial, leaving significant room for exploration in other domains, especially those requiring nuanced interpretation of structured and unstructured textual data.

D. Gap and Potential of Multi-Agent LLM

Wang et al. [9]. Although multi-agent LLM frameworks have shown promise in various applications, no prior work has introduced their capabilities to the domain of text prediction tasks, particularly in healthcare settings such as Electronic Health Records (EHR) data analysis. Text-based prediction in EHR data requires the synthesis of complex, multi-modal information such as patient history, diagnostic codes, medication records, and clinical notes. This task is inherently challenging due to the heterogeneity and volume of medical data, making it an ideal candidate for multi-agent LLM frameworks.

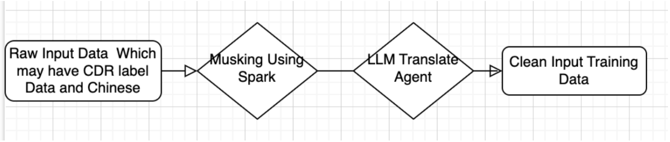
Raw Input Data Which may have CDR label Data and Chinese
Musking Using Spark
LLM Translate Agent
Clean Input Training Data

Fig. 1. BlockDiagram for Data Preprocessing

The potential of multi-agent LLMs in this context lies in their ability to divide and conquer the prediction task. For instance, one agent could specialize in expanding and reorganizing unstructured data like lab results and cylinical notes, while another focuses on analyzing the cleaned data from previous agent. These agents can communicate and collaborate to generate comprehensive and accurate predictions, such as forecasting patient outcomes, identifying potential risk factors, or recommending personalized treatment plans.

Introducing multi-agent LLM frameworks to EHR text prediction tasks could not only fill this gap in current research but also pave the way for significant advancements in healthcare AI. By leveraging the collaborative nature of multi-agent systems, researchers and practitioners can harness the full potential of LLMs to address critical challenges in medical decision-making, ultimately improving patient care and outcomes.

III. Data

The dataset we use is a raw SOAP (Subjective-Objective-Assessment-Plan) data collected by Taiwan Hospital and need for Alzheimer's Disease Detection.

A. Data Introduction

The dataset is in a semi-structured format, comprising Electronic Health Records (EHRs) related to Alzheimer's Disease. It includes both textual fields and a numerical target field, combining unstructured (textual) and structured (numerical) data. The columns labeled as (S), (A), and (P) serve as input features .while the Clinical Dementia Rating (CDR) score is used as the target variable for prediction. The dataset consists of raw data collected over one year from Taiwan Hospital, containing 3,702 records and three features.
Column description:

- (S) = Subjective (patients describe their own situation)
- (A) = Assessment (assessment of the doctor)
- (P) = Plan (treatment plan for the patient)

B. Storage

The data was stored in a GCP bucket, we then download the data from GCP bucket and preprocess it using spark.

C. Preproccessing

We performed the following steps for data preprocessing:

- **Data Cleaning**: We removed irrelevant elements such as special symbols and extra spaces.

- **Label Data Removal**: Since the CDR score, our target variable for prediction, is recorded in column (A), it is essential to remove this information from (A) to prevent data leakage, which could unintentionally allow the model to access the target value beforehand.
- **Translation**: The raw data contained natural language text in a mix of Chinese and English, and we experimented with two approaches to achieve the best-translated results.

1) **Agent Translation** To standardize the data, we employed an LLM agent to translate each sentence into English. Missing values in columns (S) and (A) were handled, followed by text cleaning to remove Additionally, certain label data were embedded within the input column (A). To resolve this, the LLM agent was utilized to process each entry, identify and filter out records containing label data, and ensure thorough cleaning. For instance, if a column contained raw data formatted as "English ChineseWord English ChineseWord," the agent first interpreted the mixed-language sentence, preserving its context, and then translated it entirely into English.
2) **Selective Context Translation**: First, we identified and extracted the Chinese text segments from the context, saving them in a separate file while removing them from the original content. Next, we utilized a translation-specific model, Helsinki-NLP/opus-mt-zh-en, to process the extracted Chinese segments. The translated content was then appended to the end of each context to preserve semantic integrity.

   This method outperformed direct agent translation, as isolating Chinese segments minimized noise that could arise when the LLM processes large datasets containing both Chinese and English text. While this approach altered the positions of Chinese phrases within the context, our primary focus was to maintain content accuracy and coherence. Since LLMs typically decompose sentences into smaller units for analysis, slight positional shifts were deemed insignificant for preserving meaning and overall comprehension.

IV. Methods

A. Traditional Machine Learning

We conduct a sample run on a small dataset to test the ability of traditional machine learning models mentioned by Li et al. [1]. The resulting prediction (classification) accuracy is significantly low which is probably because these models rely on cleaned datasets without noise and the help of human experts. As a result, we will focus on LLM models in later experiments.

To address this, we utilize a Large Language Model (LLM) to preprocess and filter out irrelevant data. Additionally, the agent translates any Chinese text in the input rows into English. A carefully designed prompt instructs the LLM to understand the Chinese text and generate its English equivalent. Our findings indicate that standardizing the input data to English improves the accuracy of the prediction model.

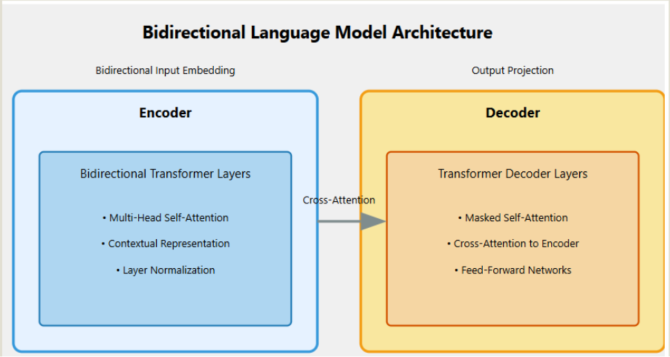
Bidirectional Language Model Architecture
Bidirectional Input Embedding
Output Projection
Encoder
Bidirectional Transformer Layers
• Multi-Head Self-Attention
• Contextual Representation
• Layer Normalization
Cross-Attention
Decoder
Transformer Decoder Layers
• Masked Self-Attention
• Cross-Attention to Encoder
• Feed-Forward Networks

Fig. 2. Bio Bert functional diagram

B. Single-agent LLM for Text Classification

We use two main LLMs to predict the CDR. The first one is BERT, a pre-trained bidirectional language model. It implements only the encoder architecture of transformers (refer fig. 2) and is good at understanding the text (in particular, good at classification and annotation). BioBERT is a variant of BERT pre-trained on thousands of medical articles and data. It performs especially well on natural language data containing medical terminologies and specialized knowledge, which is exactly our case. To achieve potential improvement in accuracy, we will also integrate fine-tuned BERT models.

The second one is GPT-4o implemented by OPENAI. It is trained with billions of parameters (which is much larger than BERT) and is good at generalization (performs well on unseen datasets even without fine-tuning). The GPT-4o architecture only contains the decoder part of the transformer (refer fig. 3) and we are interested in how it will perform on text prediction (classification) tasks.

C. Multi-agent LLM for Text Classification

In the multi-agent approach, we set up two agents to help the user make medical predictions. The first agent is the doctor's agent, which is the agent that makes the prediction. The critical evaluator is a research agent which uses ToolKits. When the doctor's agent makes his first prediction, he will send the result and reasoning to the critical agent. The critical agent will be based on the reasoning to find the related research articles, summarize the key conclusion, and send that summary back to the doctor agent. The Doctor's agent will based on the new information to do a new prediction. In future work, we may set up a chat room, which can many doctor agents and research agents to let them talk to each other and make a final prediction. The Block diagram is described in figure 4.

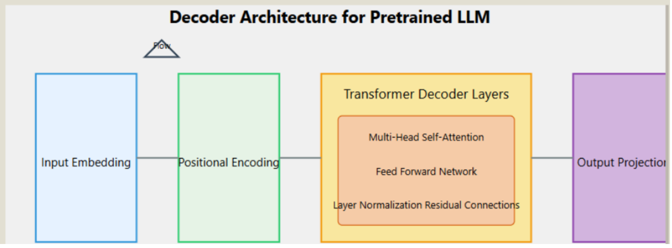
Decoder Architecture for Pretrained LLM
Flow
Input Embedding
Positional Encoding
Transformer Decoder Layers
Multi-Head Self-Attention
Feed Forward Network
Layer Normalization Residual Connections
Output Projection

Fig. 3. GPT functional diagram

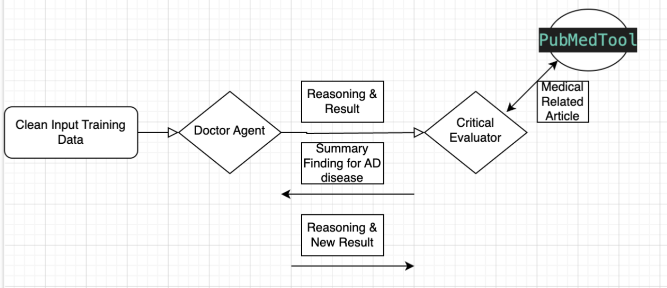

Fig. 4. Multi functional diagram

V. System Overview

A. Acceleration with Pyspark

The EHR dataset can be very large (up to 20GB and over 200,000 rows) and can easily block the server execution. To offer a good user experience under possible large input data, we introduce Spark (pySpark) running on powerful Dataproc clusters to accelerate the reading and processing of data.

B. Automatical Workflow with Airflow

Airflow is a easy workflow and enables us to trigger a workflow preprocess .csv file and uploaded the clean data file to our server. This framework is especially helpful when there are multiple python executions and complex data flows between Python files.

C. Easy-to-use Front-end APP

Our data comes from hospitals and aims at solving the problem of medical workers. Since the medical workers may not have a computer science background, we will build a front-end website that anyone can surface and receive the predicted CDR and other results. This website will be a Flask APP that receives HTTP requests, puts the received data in the right place to trigger Airflow execution, and sends the generated graphs and tables to the user's browser. The workflow diagram is shown in figure 5.
To get the analyzed results, users only need to surface our website address, select and upload the .csv file they want, and wait for the feedback. A sample UI exposed to users is shown in figure 6.

D. Data Visulization API

We use wordcloud library and matplotlib.pyplot Library for our data visualization. We create a word Cloud for most frequent important keyword in input data, and we use matplot to generate the confusion matrix for each model prediction. The data visualization sample is shown in figure 8, 7

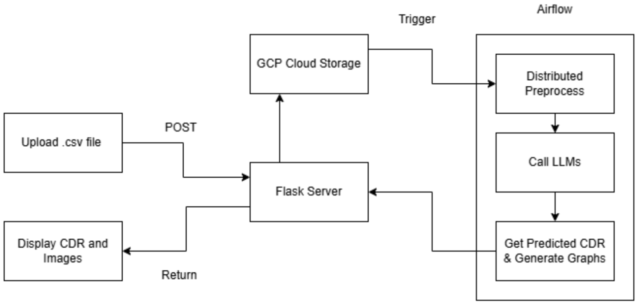

Fig. 5. Workflow of the Front-end Website

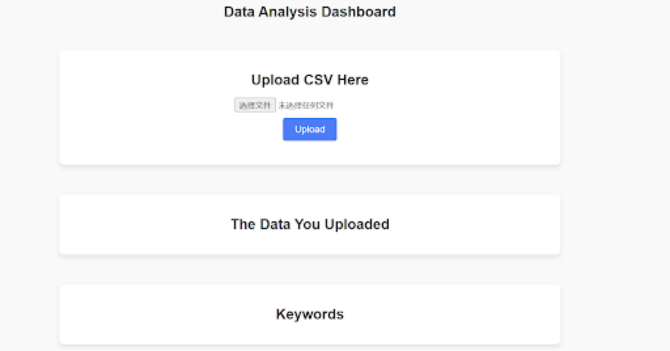

Fig. 6. Sample UI Interface that Users Can See

VI. Experiments

- **Baseline**: We used BERT as our baseline model to get the prediction results and used them for comparison.
- **Domain-Specific Model Prediction:** We deployed domain-specific model, such as BioBERT, to see if adding

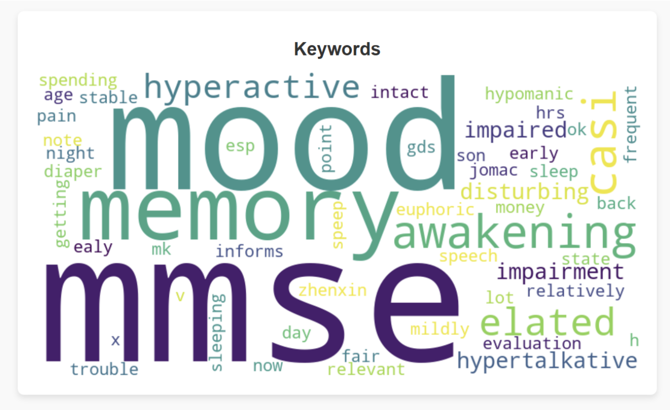

Fig. 7. Resulting Word Cloud Generated from User Upload

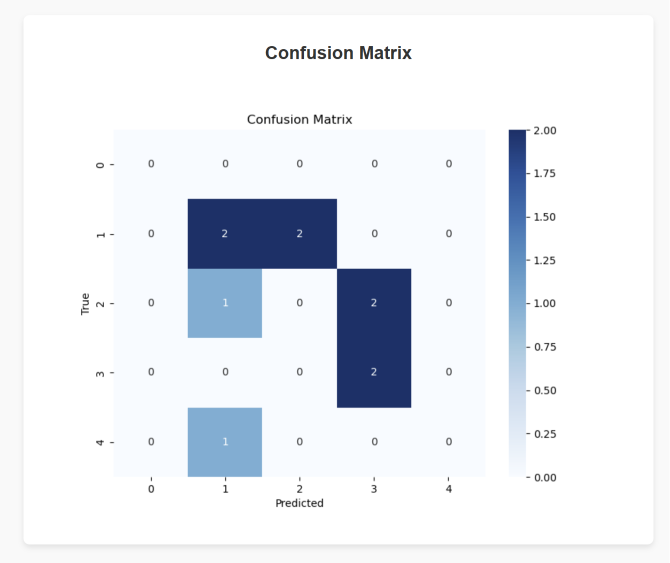

Fig. 8. Confusion Matrix for Model Prediction

domain knowledge would yield a better prediction accuracy compared to the baseline BERT.

- **LLM-based Agent**: We came up with single-agent and multi-agents to do the prediction separately. Providing instructive prompting and tools for them to perform the classification and prediction tasks. More specifically:
  1) **Single Agent:** We use a single agent with GPT-4.0 to do the prediction based on the input column.
  2) **Multi Agents:** We developed two agents—one dedicated to prediction and the other focused on evaluation and providing feedback using external resources. For the evaluator agent, we utilized the PubMed website as our primary external resource. Upon receiving a message, the agent activates its tools to search for relevant information on PubMed by leveraging embedding similarities. It retrieves the most relevant articles, generates summaries, and delivers feedback on the predictions made by the prediction agent. This iterative process facilitates result refinement and enhances overall accuracy.

In our experiment, we used different models to increase the accuracy of prediction for Alzheimer's Disease. We measure the accuracy of each model and results are shown in Table I. In before research, people used different LLM models to predict Alzheimer's Disease and measure the accuracy. Our work started a new idea that uses the multi-agent LLM model to predict Alzheimer's Disease. The architecture figure is described in 4.

Besides, in our experiment, we also experiment two ways to do the big data translation. The first one is agent translation, and second is the selective context translation. We found that the selective context translation have lower cost than agent translation.

TABLE I
PERFORMANCE TABLE FOR DIFFERENT KERNEL FUNCTIONS(MS)

Model Name	***Description***	***Accuracy***
BERT	Baseline Model	73
BioBERT	Updated Kernel	77
GPT-4o	One agent Prediction	88
Multi-agent	More Agent Prediction	91

VII. CONCLUSION

A. *Summary of Key Result*

In this study, we addressed challenges in big data processing, specifically focusing on natural language translation. Translating data row by row incurs significant API costs while uploading the entire dataset for translation introduces hallucination issues with models like GPT-4.0. To mitigate these challenges, we extracted Chinese text from the context, translated it into English, and reintegrated the English text back into the original context.

In our experiments, we evaluated various models for predicting the Clinical Dementia Rating (CDR) of Alzheimer's Disease patients. The results demonstrated that GPT-4.0 significantly outperformed traditional models like BioBERT and BERT. Furthermore, multi-agent systems showed improved performance compared to a single GPT-4.0 agent. However, the stability of multi-agent systems remains a concern; in some cases, their accuracy approaches that of the standalone GPT-4.0. We attribute this variability to the dynamic interactions and probabilistic nature of multi-agent systems, which can result in fluctuating performance.

In future work, we aim to explore diverse agent networks to enhance prediction accuracy and stability. Additionally, we plan to investigate which agent network configurations are most suitable for addressing the specific challenges of our problem.

B. *Our Contribution*

- Constrained Format Output of LLM
- Big Data Natural Language Translation
- Data Visualization for Key Word

ACKNOWLEDGMENT

We would like to thank Taiwan hospital for allowing us to use the data. Thank TA and Professor for hard-working in this semester.

REFERENCES

[1] Q. Li, X. Yang, J. Xu, Y. Guo, X. He, H. Hu, T. Lyu, D. Marra, A. Miller, G. Smith *et al.*, "Early prediction of alzheimer's disease and related dementias using real-world electronic health records," *Alzheimer's & Dementia*, vol. 19, no. 8, pp. 3506–3518, 2023. [Online]. Available: https://alz-journals.onlinelibrary.wiley.com/doi/epdf/10.1002/alz.12967

[2] S. S. Prabhu, J. A. Berkebile, N. Rajagopalan, R. Yao, W. Shi, F. Giuste, Y. Zhong, J. Sun, and M. D. Wang, "Multi-modal deep learning models for alzheimer's disease prediction using mri and ehr," in *2022 IEEE 22nd International Conference on Bioinformatics and Bioengineering (BIBE)*. IEEE, 2022, pp. 168–173. [Online]. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9973481

[3] L. Wang, J. Laurentiev, J. Yang, Y.-C. Lo, R. E. Amariglio, D. Blacker, R. A. Sperling, G. A. Marshall, and L. Zhou, "Development and validation of a deep learning model for earlier detection of cognitive decline from clinical notes in electronic health records," *JAMA network open*, vol. 4, no. 11, pp. e2 135 174–e2 135 174, 2021.

[4] X. Xu, B. Yao, Y. Dong, S. Gabriel, H. Yu, J. Hendler, M. Ghassemi, A. K. Dey, and D. Wang, "Mental-llm: Leveraging large language models for mental health prediction via online text data," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 8, no. 1, pp. 1–32, 2024. [Online]. Available: https://dl.acm.org/doi/pdf/10.1145/3643540

[5] Y. Feng, X. Xu, Y. Zhuang, and M. Zhang, "Large language models improve alzheimer's disease diagnosis using multi-modality data," in *2023 IEEE International Conference on Medical Artificial Intelligence (MedAI)*. IEEE, 2023, pp. 61–66.

[6] C.-P. Chen and J.-L. Li, "Profiling patient transcript using large language model reasoning augmentation for alzheimer's disease detection," *arXiv preprint arXiv:2409.12541*, 2024.

[7] X. Du, J. Novoa-Laurentiev, J. M. Plasek, Y.-W. Chuang, L. Wang, G. A. Marshall, S. K. Mueller, F. Chang, S. Datta, H. Paek *et al.*, "Enhancing early detection of cognitive decline in the elderly: a comparative study utilizing large language models in clinical notes," *EBioMedicine*, vol. 109, 2024.

[8] T. Guo, X. Chen, Y. Wang, R. Chang, S. Pei, N. V. Chawla, O. Wiest, and X. Zhang, "Large language model based multi-agents: A survey of progress and challenges," *arXiv preprint arXiv:2402.01680*, 2024. [Online]. Available: https://arxiv.org/pdf/2402.01680

[9] J. Wang, S. Ahn, T. Dalal, X. Zhang, W. Pan, Q. Zhang, B. Chen, H. H. Dodge, F. Wang, and J. Zhou, "Augmented risk prediction for the onset of alzheimer's disease from electronic health records with large language models," *arXiv preprint arXiv:2405.16413*, 2024. [Online]. Available: https://arxiv.org/pdf/2405.16413