Enhancing Early Detection of Cognitive Decline in the Elderly through Ensemble of NLP Techniques: A Comparative Study Utilizing Large Language Models in Clinical Notes

Xinsong Du[1,2*], John Novoa-Laurentiev[1], Joseph M. Plasek[1,2], Ya-Wen Chuang[3], Liqin Wang[1,2], Frank Chang[1], Surabhi Datta[4], Hunki Paek[4], Bin Lin[4], Qiang Wei[4], Xiaoyan Wang[4], Jingqi Wang[4], Hao Ding[4], Frank J. Manion[4], Jingcheng Du[4], Li Zhou[1,2]

[1]Division of General Internal Medicine and Primary Care, Brigham and Women's Hospital, Boston, Massachusetts 02115
[2]Department of Medicine, Harvard Medical School, Boston, Massachusetts 02115
[3]Division of Nephrology, Taichung Veterans General Hospital, Taichung, Taiwan, 407219
[4]Intelligent Medical Objects, Rosemont, Illinois, 60018

***Corresponding author:**

Xinsong Du, Ph.D.

Division of General Internal Medicine and Primary Care

Department of Medicine

Brigham and Women's Hospital and Harvard Medical School

399 Revolution Dr, Suite 777

Somerville, MA 02145

Tel: 3522811387 | E-mail: xidu1@bwh.harvard.edu

Summary: We found LLM, traditional machine learning, and deep learning had diverse error profiles on cognitive decline identification from clinical notes, and the ensemble of LLM, machine learning, and deep learning achieved state of the art performance.

Funding Source: This study was funded by NIH-NIA R44AG081006.

Conflict of Interest Statement: All authors have NO affiliations with or involvement in any organization or entity with any financial interest (such as honoraria; educational grants; participation in speakers' bureaus; membership, employment, consultancies, stock ownership, or other equity interest; and expert testimony or patent-licensing arrangements), or non-financial interest (such as personal or professional relationships, affiliations, knowledge or beliefs) in the subject matter or materials discussed in this manuscript.

Abstract

Background: Early detection of cognitive decline in elderly individuals facilitates clinical trial enrollment and timely medical interventions. This study aims to apply, evaluate, and compare advanced natural language processing techniques for identifying signs of cognitive decline in clinical notes.

Methods: This study, conducted at Mass General Brigham (MGB), Boston, MA, included clinical notes from the 4 years prior to initial mild cognitive impairment (MCI) diagnosis in 2019 for patients ≥ 50 years. Note sections regarding cognitive decline were labeled manually. A random sample of 4,949 note sections filtered with cognitive functions-related keywords were used for traditional AI model development, and 200 random subset were used for LLM and prompt development; another random sample of 1996 note sections without keyword filtering were used for testing. Prompt templates for large language models (LLM), Llama 2 on Amazon Web Service and GPT-4 on Microsoft Azure, were developed with multiple prompting approaches to select the optimal LLM-based method. Baseline comparisons were made with XGBoost and a hierarchical attention-based deep neural network model. An ensemble of the three models was then constructed using majority vote.

Results: GPT-4 demonstrated superior accuracy and efficiency to Llama 2. The ensemble model outperformed individual models, achieving a precision of 90.3%, recall of 94.2%, and F1-score of 92.2%. Notably, the ensemble model demonstrated a marked improvement in precision (from a 70%-79% range to above 90%) compared to the best performing single model. Error analysis revealed 63 samples were wrongly predicted by at least one model; however, only 2 cases (3.2%) were mutual errors across all models, indicating diverse error profiles among them.

Conclusion: Our findings indicate that LLMs and traditional models exhibit diverse error profiles. The ensemble of LLMs and locally trained machine learning models on EHR data was found to be complementary, enhancing performance and improving diagnostic accuracy.

Keywords: Cognitive Dysfunction, Natural Language Processing, Neurobehavioral Manifestations, Electronic Health Records, Early Diagnosis, Alzheimer Disease, Dementia

1. Introduction

Alzheimer's disease (AD) and related dementias (ADRD), the sixth leading cause of death in the US, affect 5.7 million Americans, with projections estimating an increase to 13 million by 2060.[1] The progression of AD/ADRD significantly diminishes patient quality of life and places substantial emotional and financial strain on families, communities, and healthcare systems,[2] with care costs anticipated to reach $1.1 trillion by 2050.[3] Existing treatments, which only temporarily alleviate symptoms and decelerate progression,[4] underscore the urgent need for breakthroughs in AD/ADRD therapy.[5] Cognitive decline, characterized by noticeable deficits in cognitive functions beyond what might be expected from normal aging[6] serves as a key early sign of AD/ADRD,[7] emphasizing the necessity of timely detection for slowing disease advancement, facilitating involvement in clinical trials, and initiating early interventions.[8,9]

Electronic health records (EHRs) provide comprehensive, both real-time and historical, patient data, making them a critical resource for identifying cognitive decline. However, challenges arise with structured data like International Classification of Diseases (ICD) codes, which may lack the granularity to capture the spectrum of cognitive impairments, risking clinician under-reporting.[10] Moreover, traditional diagnostic tools may fail to detect early stages of cognitive decline, increasing the risk of missed or incorrect diagnoses.[11] The variability in screening practices further complicates early detection during standard medical evaluations.[12] [13] Additionally, constraints such as limited access to specialized care providers, including neurologists and geriatricians, exacerbate delays in diagnosis and treatment initiation.

Clinical notes within EHRs contain critical information on cognitive decline, detailing symptoms like memory loss, language difficulties, and impaired daily activities. Natural language processing (NLP) offers a powerful tool to identify these early signs of decline, which may not be coded in diagnoses. NLP facilitates the rapid and efficient analysis of large datasets, outpacing manual review. Researchers have investigated NLP techniques for detecting cognitive decline from clinical notes.[14] For example, Penfold et al. utilized Python scripts to identify phrases indicative of cognitive function, predicting future mild cognitive impairment (MCI).[15] Wang et al. created and validated a deep learning model to detect evidence of cognitive decline within EHR clinical notes.[10] Nevertheless, no study has been done to investigate the effectiveness of large language model (LLM) on cognitive decline identification.

LLMs like GPT-4 and Llama 2, with their extensive pre-training on diverse text corpora, excel in understanding complex clinical narratives, offering advantages over traditional rule-based and machine learning approaches that are often trained from scratch on narrower clinical datasets. Their deep contextual comprehension enables the detection of subtle indicators of cognitive status within text.[16] This research leverages LLMs within secure cloud computing environments for a pioneering exploration of EHR note analysis for cognitive decline detection. It also evaluates LLMs' effectiveness and interpretability against conventional machine learning methods and examines the synergy of LLMs with machine learning to enhance diagnostic accuracy. To the best of our knowledge, this initiative is the first of its kind to employ LLMs in such a capacity, representing a substantial innovation and contribution to the field.

2. Methods

2.1. Setting and Datasets

This study was conducted at Mass General Brigham (MGB), a large, integrated health care delivery system in Massachusetts. MGB has established secure HIPAA-compliant cloud environments for deploying and evaluating LLMs using actual EHR data. Two LLMs were included in this study: GPT-4[17] via Microsoft Azure OpenAI Service Application Programming Interface (API) and Llama 2[18] via an Amazon Elastic Compute Cloud (EC2) instance. Details about the two cloud environments were described in **Supplementary Material Section 1 and Table S1**. We set the temperature to 0 for LLMs to eliminate variability.

We used two datasets annotated from a previous study, which included sections of clinical notes over the four years leading up to the initial diagnosis of mild cognitive impairment (MCI) (ICD-10-CM code G31.84) in 2019 for patients aged 50 years or older.[10] Because the positive case rate across any sections were low, a list of expert-curated keywords (**Table 2**) were used to screen sections that likely contain indications of cognitive decline. Dataset I included 4,949 sections filtered by keywords and were used for training two baseline models. For prompt development and LLM selection, we used a 200 random samples from Dataset I, and we name this dataset as Dataset I-S. Dataset II, a random sample of 1,996 sections not subjected to keyword filtering, were used for final testing.

The identification of cognitive decline encompasses various stages, ranging from subjective cognitive decline (SCD) to MCI to dementia, with a focus on the progressive nature of cognitive decline that is likely to correspond to or result in MCI. Cases considered transient (e.g., memory loss attributed to medication) or reversible were labeled as negative for cognitive decline.[10]

The study was approved by MGB Institutional Review Board with waiver of informed consent from study participants owing to secondary use of EHR data.

2.2.Prompt Development and LLM Selection

As illustrated in **Figure 1**, to optimize LLM, we first selected the best prompt template for each model via manual template engineering[19], then selected the best prompt template and model combination, finally, we tested if prompt augmentation[19] with few shot and error analysis-based instructions[20] could improve the model performance. Detailed steps for LLM development optimization are illustrated below.

Our task was to identify if a clinical note contains information that could identify the presence of cognitive decline, which is text classification. To complete the task using large language models, we developed requirements for prompts based on our needs, and then iteratively refined the prompts through team discussion. Since our problem belongs to clinical notes binary classification task, we need to measure the classification performance and interpret the result. Our requirements of the prompt include: 1) identify whether the clinical note has evidence showing cognitive decline or not; 2) output keywords that helped the LLM to make the judgment; and 3) ask the LLM to output its response to a JSON format, which makes it easier to use code to parse the result. Our internal discussion for the prompt improvement included: 1) to ask the LLM to output its reason for the judgment; 2) to include our definition of cognitive decline in the prompt. The temperature hyperparameter is used to control randomness and creativity, with 0 corresponding to a deterministic solution. During testing, we set the temperature to 0 to minimize randomness in response generation[20]. We categorized LLM responses into three categories as shown in **Supplementary Table S2**: 1) effective and parseable, 2) effective but not parseable, 3) not effective.

To develop the optimal prompt and model, we implement a structured approach as depicted in **Figure 1**. The process begins with prompt template engineering, where for each template, we manually tested 10 random samples from the training set to evaluate the percentage of effective responses. We then manually adjusted the prompt by paraphrasing or adding/removing optional content, such as definitions or requests for explanation of decisions, for each model. This tuning is stopped once there's no improvement in the effective response rate after three consecutive attempts. Then, for the most effective prompt template identified for each LLM model, we used Dataset I-S to evaluate the percentage of parseable responses. We then assessed the accuracy, and chose the LLM that demonstrated superior accuracy. Next, we explored prompt augmentation to see if the inclusion of five examples (five-shot prompting) enhances performance. We used Template 6 in Supplementary **Table S3** for the augmented prompt. Finally, we tested the effectiveness of error analysis-based instruction.

2.3. Baseline Models

We used XGBoost [21] as the machine learning model, and Convolutional Neural Network + Long-Short Term Memory with Hierarchical Attention as the deep learning model [22,23], since they achieved state-of-the-art performances according to our previous study[10]. XGBoost stands for Extreme Gradient Boosting and represents a scalable, distributed machine learning library based on gradient-boosted decision trees (GBDT). Renowned for its parallel tree boosting capabilities, it stands as the foremost choice for addressing regression, classification, and ranking challenges within the machine learning domain [21]. A convolutional neural network (CNN) is a deep learning model specifically designed to process structured grid-like data, such as images or sequences [24]. Long Short-Term Memory (LSTM) is a type of recurrent neural network (RNN) architecture designed to overcome the vanishing gradient problem in traditional RNNs and effectively capture long-range dependencies in sequential data [25]. Connecting a LSTM neural network with an attention layer enhances interpretability by allowing the model to focus on relevant parts of the input sequence while making predictions[23]. For model parameter optimization, we used five-fold cross-validation on training data.

2.3.1. Ensemble Model

To create the ensemble model, we take the majority vote of the LLM, deep learning model, and machine learning model as the label. The LLM is the best model selected from the above procedures. The deep learning model is the same as our previous study, which connects LSTM with attention. We employed XGBoost as our machine learning model since it was the best-performing machine learning model based on our previous study [10].

2.3.2. Model Evaluation

Machine learning and deep learning models were trained and optimized using Dataset I. LLM and prompt were optimized using a subset (200 sectioned notes) of the 4,949 keyword filtered sections (Dataset I-S) due to budget limit. We tested each AI model on Dataset II, and reported precision, recall, and F1 score.

2.4.Error Analysis and Interpretation

Errors made by each model were analyzed and discussed among two biomedical informatician and a physician. For errors made by LLM, we also read the explanation output from the model. We also quantified unique and overlapping errors made by each model using a Venn Diagram. Regarding interpretation, we listed keywords that had a high frequency of appearance from LLM's output (i.e., the number of appearances is higher than the average

appearance time plus two standard deviations); keywords whose deep learning attention weights above the mean weights plus 2 standard deviations within individual sections; and keywords with an XGBoost information gain higher than the average value plus 2 standard deviations. We also listed expert curated keywords developed in our previous study as a reference[10].

3. Results

Detailed dataset characteristics and model's optimized hyperparameters are illustrated in the Supplementary Material. We highlighted model evaluation, interpretation, and error analysis below.

3.1.Performance evaluation

As illustrated in **Table 1** and **Figure 2,** we tested the Optimized LLM, XGBoost, and deep learning model on the test set. It turned out the LLM achieved a precision of 71.6%, a recall of 91.3%, and an F1 score of 80.3%. Optimized parameters for the deep learning model and XGBoost model are illustrated in **Supplementary Table S8**. The optimized XGBoost model achieved a precision of 79.0%, recall of 92.8%, and F1 score of 80.3%; the optimized deep learning model obtained a precision of 77.1%, recall of 92.8, and F1 score of 84.2%. Notably, an ensemble consisting of the LLM, deep learning model, and machine learning model boosted the performance (over 10% precision increase and around 7% F1-score improvement), which achieved a precision of 90.3%, recall of 94.2%, an F1 score of 92.2%.

3.2.Interpretation and Error Analysis

Table 2 contains keywords that have been identified through expert curation, exported by XGBoost, the deep learning model, and the LLM. These keywords cover a range of topics including memory-related issues such as recall and forgetfulness, cognitive impairments, and dementia with terms like dementia and Alzheimer, and evaluation and assessment methods, referencing tools like MoCA and MMSE. Comparing with traditional AI models and expert-selected keywords, GPT-4 highlighted specific treatment options, notably "Aricept" and "donepezil," which are essential in managing dementia and Alzheimer's disease. GPT-4 also identified specific diagnoses or conditions more explicitly than other models, with terms like "Mild Neurocognitive Disorder," "major neurocognitive disorder," and "vascular dementia." Additionally, GPT-4 exported some keywords regarding emotional and psychological effects of cognitive disorders, such as "anxiety," thus covering aspects sometimes overlooked by other models.

An error analysis performed on LLM with test data is highlighted in **Figure 3**. We found all models could be misled by signs or symptoms as being caused by unrelated clinical conditions or be confounded by negations and other contextual details. Notably, GPT-4 stood out for its clarity in handling ambiguous terms, a common pitfall for traditional AI models. It shows an adeptness at inferring or emphasizing nuanced information that traditional AI methods typically overlook. However, GPT-4 could infer/amplify the nuanced information while traditional AI models did not make such mistakes. Additionally, GPT-4 was sometimes overly conservative, failing to identify cognitive decline despite compelling evidence. GPT-4 might also miss the underlying reasons for specific clinical events, such as visits or treatments related to cognitive decline. Both GPT-4 and deep learning models might misinterpret clinical testing results, marking an area for future improvement.

4. Discussion

Early detection of precursor stages of AD/ADRD, such as cognitive decline, becomes extremely important, as it can introduce treatment or intervention earlier to effectively reduce progression to AD/ADRD [8,9]. In this study, we aimed to develop an NLP model for cognitive decline identification from clinical notes in EHR. We used data collected in our previous study for model development. Our study found that LLM did not outperform traditional AI for cognitive decline identification. This is because LLM was not trained specifically for a task, it is a more powerful tool than traditional AI models, and its output can have a wide range of responses, possibly giving ambiguous, inconsistent, and sometimes out-of-context answers[16]. We found although three types of models achieved similar performances, they made different types of errors. Notably, the ensemble of the three types of models boosted the performance (e.g., there was an over 10% precision improvement). Regarding interpretation, we found that LLM could identify keywords that were not identified by experts or the deep learning model, such as words related to medications. As illustrated in Supplementary **Table S9**, during error analysis, we found LLM could be very helpful to help physicians quickly identifying cognitive-decline related medications from a long medication report. Overall, our developed method can be a valuable tool for screening older patients to identify those with evidence of cognitive decline in clinical notes.

Many existing studies focused on using AI techniques to predict later stage of cognitive decline [7,9,32–34]. A few studies were about NLP techniques for detecting cognitive decline. Moreira et al. (2018) [35] used unsupervised learning approach to cluster clinical notes and get pathological features, and then integrated structured EHR data for cognitive decline identification. They achieved a recall of 0.8 and an AUROC of 0.87. Wang et al. (2021) [10] developed a deep learning model (i.e., LSTM with attention model) for cognitive decline identification on clinical notes, and achieved a precision of 0.77 precision and a recall of 0.93. The study also found the attention-based deep learning model was able to identify useful keywords missed by experts. In 2022, Penfold et al. [15] developed a LASSO model using clinical notes and reported a precision of 0.70 and a recall of 0.17. More recently, Fouladvand et al. (2023) [36] integrated NLP techniques to identify cognitive decline using both structured and unstructured EHR data, and achieved an AUROC of 0.68 meanwhile ranked important predictors from both structured data and unstructured data. Compared with existing studies, our study achieved state-of-the-art performance (precision 0.90, recall 0.94, and f1 score 0.92) meanwhile identified keywords that were missed by experts and traditional AI models.

Our study has several strengths. Firstly, we set up private computational environments and tested two LLMs using cloud platforms. We noticed that Microsoft Azure GPT-4 outperformed AWS EC2 Llama 2 in terms of price, execution speed, and accuracy. However, as an open-source model, Llama 2 may have a better reproducibility [37] while GPT-4 may provide slightly different answers over time due to model updates from OpenAI [38]. Secondly, we tested several prompting strategies, and found doing error analysis using some training data, and then asking LLM to avoid summarized common errors by revising the prompt can be an effective way to improve LLM's performance. Additionally, we developed a prompt template for identifying cognitive decline from clinical note using LLM, and the template can also be used as a good reference when developing prompts for identifying other diseases. To the best of our knowledge, this is the first study employing LLM and unstructured EHR data for cognitive decline detection.

However, the results of this study should be considered along with some limitations. We used data from our previous study, which has several limitations including: 1) the data is from a single healthcare system, thus, external validation is still needed; 2) the data is retrospective and include notes written four years preceding the diagnose, but detecting cognitive decline beyond the 4-year window is still needed; 3) the data is unimodal, integrating other types of data (e.g., image, genomics) may further improve the performance. Additionally,

considering the budget issue, the Llama 2 model used in our study is not the best one, which contains 70 billion parameters and requires more expensive hardware. Furthermore, our data is record-based and not patient based (I.e., longitudinal), thus, the developed model may not be able to distinguish well between reversable and progressive cognitive decline, since we do not know if the patient recovered later or not based only on a note section at one time point. An LLM-based early warning system of cognitive decline developed with longitudinal data would be a valuable future work.

5. Conclusion

This study is among the initial endeavors to utilize LLMs within HIPAA-compliant cloud environments, leveraging real EHR notes for the detection of cognitive decline. Our findings indicate that LLMs and traditional models exhibit diverse error profiles in comparison. The ensemble of LLMs and locally trained machine learning models on EHR data was found to be complementary, enhancing performance and improving diagnostic accuracy. Future work can leverage longitudinal and multimodal data to further improve performance.

Tables

Table 1. Performance evaluation of the best LLM and prompting combination and machine learning models. XGBoost is the best performing individual model, which achieved a precision of 79.01%, recall of 92.75%, and f1 score of 85.33%. The ensemble model boosted the best performance to precision - 90.28% (11.27% higher than XGBoost), recall – 94.20% (1.45% higher than XGBoost), f1 score – 92.20 (6.87% higher than XGBoost).

Model	Prompt Prefix Text Version	Precision	Recall	F1 Score	GPT-4 Execution Date
GPT4-8K	0 shot improved with error analysis	71.59%	91.30%	80.25%	01/03/2024
Deep Learning	NA	77.11%	92.75%	84.21%	-
XGBoost	NA	79.01%	92.75%	85.33%	-
Ensemble (vote of above 3 rows)	0 shot improved with error analysis	**90.28%**	**94.20%**	**92.20%**	01/03/2024

Table 2. Keywords contributing to identifying positive cognitive decline cases extracted from experts, deep learning model, and large language model. The table listed keywords that had a high frequency of appearance from LLM's output (i.e., the number of appearances is higher than the average appearance time plus two standard deviations); keywords whose deep learning attention weights above the mean weights plus 2 standard deviations within individual sections; and keywords with an XGBoost information gain higher than the average value plus 2 standard deviations. We found keywords identified by AI models could significantly enrich the expert curated keyword set. Notably, only LLM identified keywords regarding medications for cognitive decline.

Model	Keywords
Expert Curated	Memory, delirium, dementia, psycha, neuroa, mental, alzheimer, confusa, mood, cognita, forgeta, agitate, moca, montreal, mmse, remember, difficult, recall, function, word, evaluata, score, drive, attention, mild, impairment, speech, question, disorientation, orientation, sleep, alter, exam, decline, worse, loss
XGBoost	Cognitive, memory, dementia, forgetful
Deep Learning	Memory, cognitive, dementia, impairment, neurocognitive, recall, decline, word, forgetful, cognition, MCI, executive, forgetfulness, alzheimer, remembering, MoCA, attention, deficits, recalling, forgetting, finding, words, visual, forgets, naming, difficulties, fluency, delay, alzheimers, retrieval, visuospatial, repetition, remember, hearing, cog, trails, language, FTD, frailty, encoding, developmental, delayed, behavioral, amnestic, phonemic, MMSE, falls, errors, attentional, speech, span, processing, neurodegenerative, lapses, HOH, deficit, correctly, auditory, years, spatial, solving, semantic, personality, perseveration, names, multidomain, moderately, linguistic, learning, LBD, items, insight, impaired, immediate, global, functioning, functional, expressive, died, cube, comprehension, clock, challenges, category, BNT, aphasia, amyloid, age, aforementioned, abstraction
GPT4-8K	dementia, cognitive impairment, memory loss, memory, mild cognitive impairment, neuropsychological evaluation, confusion, confused, Memory loss, forgetful, deficits, cognitive decline, Mild cognitive impairment, Cognitive impairment, Dementia, cognitive changes, Aricept, current level of cognitive functioning, memory impairment, Mild Neurocognitive Disorder, executive functioning, memory issues, attention, cognitive deficits, Memory impairment, MCI, neuropsychological tests, cognition, poor safety awareness, forgetfulness, Cognitive decline, cognitive difficulties, MOCA, memory problems, mild neurocognitive disorder, donepezil, neuropsychological testing, Impaired, memory concerns, major neurocognitive disorder, cognitive concerns, mild dementia, weakness, memory difficulties, working memory, altered mental status, concerns, MCI (mild cognitive impairment), neurodegenerative process, language, cognitive symptoms, neurocognitive status, Altered mental status, short term memory loss, cognitive issues, neurocognitive disorder, executive function, battery of neuropsychological tests, neuropsych testing, problem solving, verbal fluency, memory complaints, vascular dementia, word-finding difficulties, processing speed, cognitive-linguistic therapy, delirium, word finding difficulties, delayed recall, anxiety

Figures

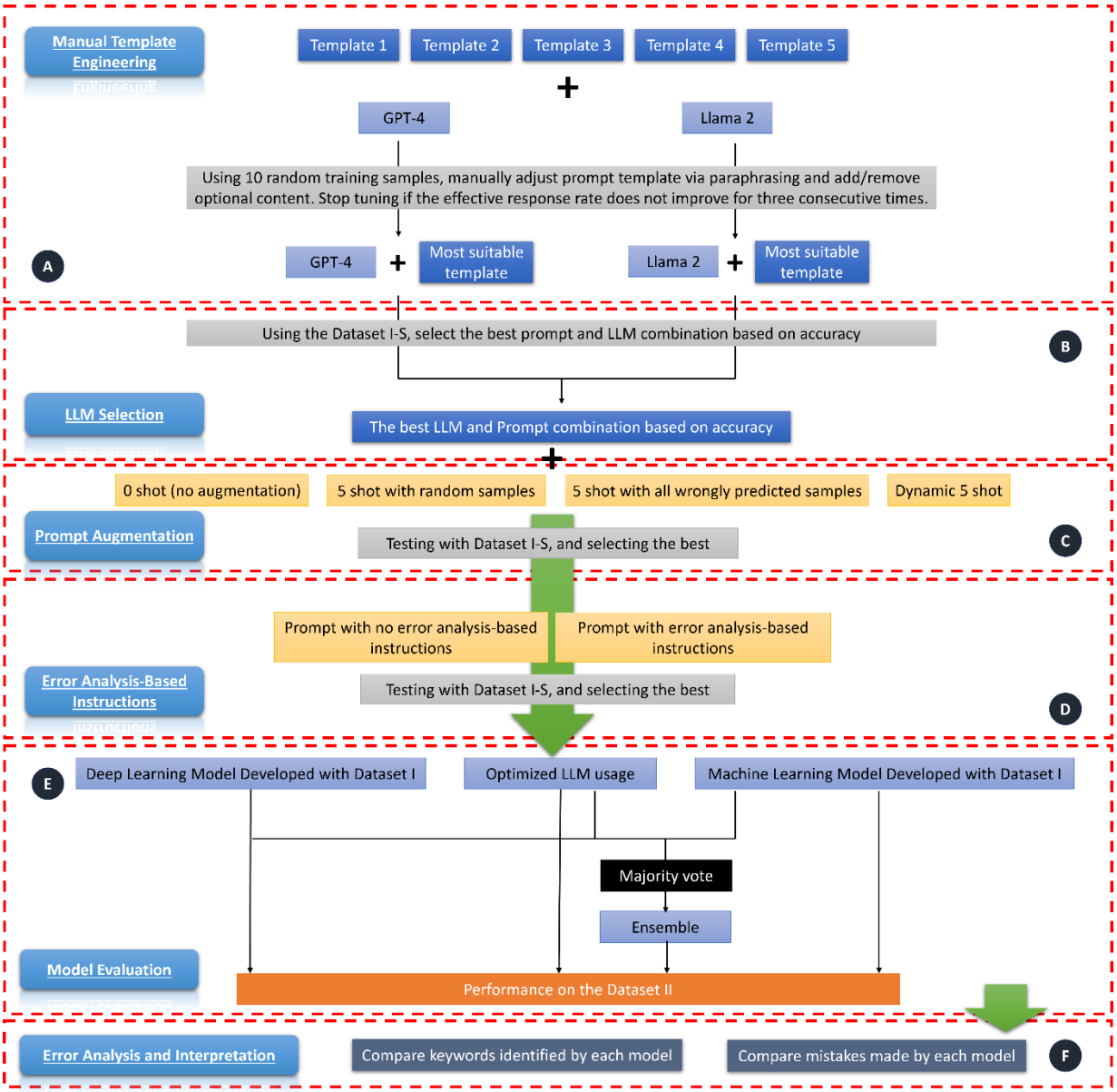

Figure 1. Study design overview. The workflow consists of five parts. A) Manual template engineering: we fed each prompt to GPT-4 and Llama 2 separately. If the effective response (i.e., the response answered the questions in the prompt) rate was not 100%, then manually adjust the template. If the effective response rate does not improve after tuning three times, then stop tuning and use the template leading to the best effective response rate. B) LLM selection: We used Dataset I-S and tested the accuracy of results produced by GPT-4 + its selected template and Llama 2 + its selected template. We used the best performing model and template version combination for following experiments. C) Prompt augmentation: we tested if five-shot prompting could improve the accuracy. D) Error analysis-based instructions: we tested if adding instructions following error analysis of GPT-4's output on the Dataset I-S could improve the accuracy. E) Model evaluation: we evaluated the developed deep learning model, large language model, and machine learning model. We also tested the performance of ensemble model, which took the majority vote of the three models as the predictive label. F) Error analysis and interpretation: we exported keywords used by each model for the prediction, and compared the keywords with expert curated keywords, then we analyzed and compared errors made by each model.

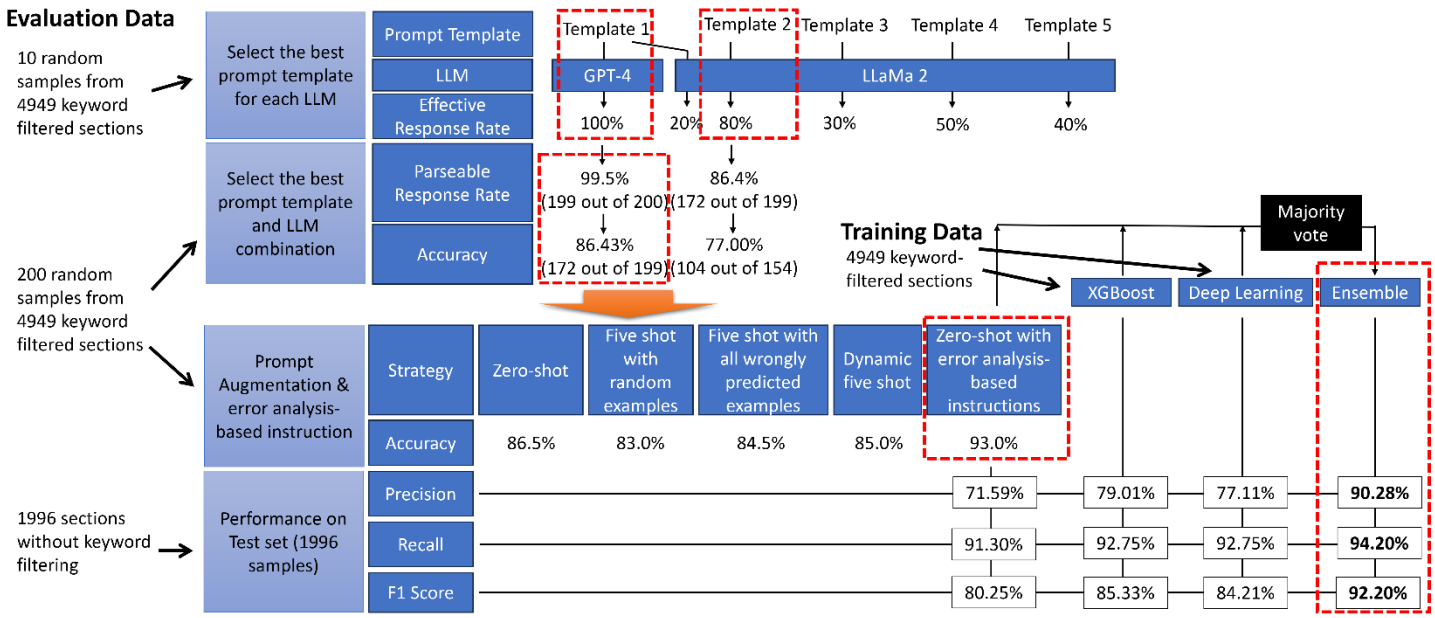

Figure 2. Evaluation results summary. During the prompt template selection, Template 1 was selected for GPT-4 model due to an 100% effective response rate; Template 2 was selected for Llama 2 model since the effective response rate (80%) was not improved after tuning it for three times. Then, we compared the two combinations with 200 samples (Dataset I-S), and found GPT-4 + Template 1 had a much better accuracy (86.43%). Next, we found five shot prompting did not lead to a better performance, but adding error analysis-based instructions (i.e., GPT-4 + Template 7) could improve the accuracy to 93% on the Dataset I-S. Therefore, we selected using Template 7 as the prompt template, and GPT-4 as the LLM. On the test set, we evaluated the performances of developed machine learning model (XGBoost), developed deep learning model, and developed LLM. We found the XGBoost had a better performance: precision – 79.01%, recall – 92.75%, f1 score – 85.33%. Notably, after ensemble the three models using majority vote, we found the ensemble model had a much better performance: precision – 90.11% (11.1% improvement), recall – 94.20% (1.45% improvement), and f1 score – 92.20% (6.87% improvement).

Potential reasons contributing to false predictions	Example	GPT-4	Deep Learning	XGBoost	Exist in areas
Signs/symptoms caused by other factors or clinical conditions	Assessment and Plan for pancreatitis, who is presenting with several days of malaise, poor appetite, slow cognition, and found to have pancreatitis c/b hypotension, now hemodynamically stable	X	X	X	
Errors caused by negation or other contextual information	1 reports significant deficits that affect daily life ?2 3 = Referral does not indicate any cognitive deficits.	X	X	X	
Errors caused by ambiguity	The patient was injected with 10 mCi and 32.6. mCi of Tc-99m Sestamibi at rest and during peak stress, respectively, and SPECT imaging was performed.	√	X	X	
Infer /amplify the nuanced information	Pt remains at heightened risk for falls and demos poor safety/judgment. I/E that pt hx of falls and strength and balance impairments place pt at heightened risk for another fall. Pt remains reluctant but son Carl states he is working on putting a plan in place.	X	√	√	
Too rigrous when identifying cognitive decline	Falls precautions in place. Mental Status, Impaired - Medical/Surgical Condition - Adult. Absence of continued neurologic deterioration signs and symptoms during hospitalization Progressing. Pt at baseline mental status. Pleasantly confused and forgetful. Bed alarm at all times. Pain - Medical/Surgical Condition - Adult.	X	√	√	
Overlooked cognitive decline related reason of a clinical event	Impressions Pt completed his first week of skilled individual SLP intervention focused on the above areas, specifically domains of memory and problem solving: Pt demonstrates strong gains towards goals.	X	√	√	
Misinterpreted clinical testing result	Test of Premorbid Functioning* 60 1.27. 119 90 Above Average. Cognitive Screen. Addenbrooke's Cognitive Examination - III (ACE-III). Total (/100) 84. Attention (/18) 18. Memory (/26) 21. Fluency (/14) 8. Language (/26) 25. Visuospatial (/16) 12.	X	X	√	

Figure 3. Venn Diagram highlighting unique and overlapping mistakes made by different models. √: correct prediction X: incorrect prediction. Some important findings include: 1) GPT-4 never predicts positive for cognitive decline if there is ambiguous information seemingly relevant to cognitive decline but not, while traditional AI approaches could make mistakes due to ambiguous terms; 2) GPT-4 could amplify the nuanced information traditional AI approaches, or be too rigorous, but traditional AI approaches do not have this issue; 3) all types of models could be misled by seemingly relevant symptoms/signs that normal people can have or very likely caused by signs/symptoms caused by other factors or clinical conditions.

References

1. 2023 Alzheimer's disease facts and figures. *Alzheimers Dement J Alzheimers Assoc*. 2023;19(4):1598-1695. doi:10.1002/alz.13016

2. Vu M, Mangal R, Stead T, Lopez-Ortiz C, Ganti L. Impact of Alzheimer's Disease on Caregivers in the United States. *Health Psychol Res*. 10(3):37454. doi:10.52965/001c.37454

3. Stefanacci RG. The costs of Alzheimer's disease and the value of effective therapies. *Am J Manag Care*. 2011;17 Suppl 13:S356-362.

4. Aducanumab to Be Discontinued as an Alzheimer's Treatment. Alzheimer's Disease and Dementia. Accessed March 9, 2024. https://alz.org/alzheimers-dementia/treatments/aducanumab

5. Cummings J, Zhou Y, Lee G, Zhong K, Fonseca J, Cheng F. Alzheimer's disease drug development pipeline: 2023. *Alzheimers Dement N Y N*. 2023;9(2):e12385. doi:10.1002/trc2.12385

6. Mitchell AJ, Beaumont H, Ferguson D, Yadegarfar M, Stubbs B. Risk of dementia and mild cognitive impairment in older people with subjective memory complaints: meta-analysis. *Acta Psychiatr Scand*. 2014;130(6):439-451. doi:10.1111/acps.12336

7. Kidd PM. Alzheimer's disease, amnestic mild cognitive impairment, and age-associated memory impairment: current understanding and progress toward integrative prevention. *Altern Med Rev J Clin Ther*. 2008;13(2):85-115.

8. Leifer BP. Early Diagnosis of Alzheimer's Disease: Clinical and Economic Benefits. *J Am Geriatr Soc*. 2003;51(5s2):S281-S288. doi:10.1046/j.1532-5415.5153.x

9. Veitch DP, Weiner MW, Aisen PS, et al. Understanding disease progression and improving Alzheimer's disease clinical trials: Recent highlights from the Alzheimer's Disease Neuroimaging Initiative. *Alzheimers Dement*. 2019;15(1):106-152. doi:10.1016/j.jalz.2018.08.005

10. Wang L, Laurentiev J, Yang J, et al. Development and Validation of a Deep Learning Model for Earlier Detection of Cognitive Decline From Clinical Notes in Electronic Health Records. *JAMA Netw Open*. 2021;4(11):e2135174. doi:10.1001/jamanetworkopen.2021.35174

11. He Z, Dieciuc M, Carr D, et al. New opportunities for the early detection and treatment of cognitive decline: adherence challenges and the promise of smart and person-centered technologies. *BMC Digit Health*. 2023;1(1):7. doi:10.1186/s44247-023-00008-1

12. Sabbagh MN, Boada M, Borson S, et al. Early Detection of Mild Cognitive Impairment (MCI) in Primary Care. *J Prev Alzheimers Dis*. 2020;7(3):165-170. doi:10.14283/jpad.2020.21

13. Whelan R, Barbey FM, Cominetti MR, Gillan CM, Rosická AM. Developments in scalable strategies for detecting early markers of cognitive decline. *Transl Psychiatry*. 2022;12(1):1-11. doi:10.1038/s41398-022-02237-w

14. Cummings J, Lee G, Nahed P, et al. Alzheimer's disease drug development pipeline: 2022. *Alzheimers Dement N Y N*. 2022;8(1):e12295. doi:10.1002/trc2.12295

15. Penfold RB, Carrell DS, Cronkite DJ, et al. Development of a machine learning model to predict mild cognitive impairment using natural language processing in the absence of screening. *BMC Med Inform Decis Mak*. 2022;22(1):129. doi:10.1186/s12911-022-01864-z

16. Caruccio L, Cirillo S, Polese G, Solimando G, Sundaramurthy S, Tortora G. Can ChatGPT provide intelligent diagnoses? A comparative study between predictive models and ChatGPT to define a new medical diagnostic bot. *Expert Syst Appl*. 2024;235:121186. doi:10.1016/j.eswa.2023.121186

17. OpenAI, Achiam J, Adler S, et al. GPT-4 Technical Report. Published online December 18, 2023. doi:10.48550/arXiv.2303.08774

18. Touvron H, Martin L, Stone K, et al. Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv.org. Published July 18, 2023. Accessed January 17, 2024. https://arxiv.org/abs/2307.09288v2

19. Liu P, Yuan W, Fu J, Jiang Z, Hayashi H, Neubig G. Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *ACM Comput Surv*. 2023;55(9):195:1-195:35. doi:10.1145/3560815

20. Hu Y, Chen Q, Du J, et al. Improving large language models for clinical named entity recognition via prompt engineering. *J Am Med Inform Assoc*. Published online January 27, 2024:ocad259. doi:10.1093/jamia/ocad259

21. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD ’16. Association for Computing Machinery; 2016:785-794. doi:10.1145/2939672.2939785

22. Yang J, Wang L, Phadke NA, et al. Development and Validation of a Deep Learning Model for Detection of Allergic Reactions Using Safety Event Reports Across Hospitals. *JAMA Netw Open*. 2020;3(11):e2022836. doi:10.1001/jamanetworkopen.2020.22836

23. Yang Z, Yang D, Dyer C, He X, Smola A, Hovy E. Hierarchical Attention Networks for Document Classification. In: Knight K, Nenkova A, Rambow O, eds. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics; 2016:1480-1489. doi:10.18653/v1/N16-1174

24. Lecun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. *Proc IEEE*. 1998;86(11):2278-2324. doi:10.1109/5.726791

25. Hochreiter S, Schmidhuber J. Long Short-Term Memory. *Neural Comput*. 1997;9(8):1735-1780. doi:10.1162/neco.1997.9.8.1735

26. Temsah O, Khan SA, Chaiah Y, et al. Overview of Early ChatGPT’s Presence in Medical Literature: Insights From a Hybrid Literature Review by ChatGPT and Human Experts. *Cureus*. 2023;15(4):e37281. doi:10.7759/cureus.37281

27. Ahmad N, Murugesan S, Kshetri N. Generative Artificial Intelligence and the Education Sector. *Computer*. 2023;56(6):72-76. doi:10.1109/MC.2023.3263576

28. Leippold M. Thus Spoke GPT-3: Interviewing a Large-Language Model on Climate Finance. Published online September 12, 2022. doi:10.2139/ssrn.4237242

29. Koga S, Martin NB, Dickson DW. Evaluating the performance of large language models: ChatGPT and Google Bard in generating differential diagnoses in clinicopathological conferences of neurodegenerative disorders. *Brain Pathol Zurich Switz*. Published online August 8, 2023:e13207. doi:10.1111/bpa.13207

30. Pillai J, Pillai K. Accuracy of generative artificial intelligence models in differential diagnoses of familial Mediterranean fever and deficiency of Interleukin-1 receptor antagonist. *J Transl Autoimmun*. 2023;7:100213. doi:10.1016/j.jtauto.2023.100213

31. Sensoy E, Citirik M. A comparative study on the knowledge levels of artificial intelligence programs in diagnosing ophthalmic pathologies and intraocular tumors evaluated their superiority and potential utility. *Int Ophthalmol*. 2023;43(12):4905-4909. doi:10.1007/s10792-023-02893-x

32. Adelson RP, Garikipati A, Maharjan J, et al. Machine Learning Approach for Improved Longitudinal Prediction of Progression from Mild Cognitive Impairment to Alzheimer's Disease. *Diagnostics*. 2024;14(1):13. doi:10.3390/diagnostics14010013

33. Fouladvand S, Noshad M, Periyakoil VJ, Chen JH. Machine learning prediction of mild cognitive impairment and its progression to Alzheimer's disease. *Health Sci Rep*. 2023;6(10):e1438. doi:10.1002/hsr2.1438

34. James C, Ranson JM, Everson R, Llewellyn DJ. Performance of Machine Learning Algorithms for Predicting Progression to Dementia in Memory Clinic Patients. *JAMA Netw Open*. 2021;4(12):e2136553. doi:10.1001/jamanetworkopen.2021.36553

35. Moreira LB, Namen AA. A hybrid data mining model for diagnosis of patients with clinical suspicion of dementia. *Comput Methods Programs Biomed*. 2018;165:139-149. doi:10.1016/j.cmpb.2018.08.016

36. Fouladvand S, Noshad M, Goldstein MK, Periyakoil V, Chen JH. Mild Cognitive Impairment: Data-Driven Prediction, Risk Factors, and Workup. *AMIA Summits Transl Sci Proc*. 2023;2023:167-175.

37. DeWitt PE, Rebull MA, Bennett TD. Open source and reproducible and inexpensive infrastructure for data challenges and education. *Sci Data*. 2024;11(1):8. doi:10.1038/s41597-023-02854-0

38. Chen L, Zaharia M, Zou J. How is ChatGPT's behavior changing over time? Published online October 31, 2023. doi:10.48550/arXiv.2307.09009

39. Nori H, Lee YT, Zhang S, et al. Can Generalist Foundation Models Outcompete Special-Purpose Tuning? Case Study in Medicine. Published online November 27, 2023. doi:10.48550/arXiv.2311.16452