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

Xinsong Du[1,2], John Novoa-Laurentiev[1], Joseph M. Plasaek[1,2], Ya-Wen Chuang[3], Liqin Wang[1,2], Gad Marshall[2,4], Stephanie K. Mueller[1,2], Frank Chang[1], Surabhi Datta[5], Hunki Paek[5], Bin Lin[5], Qiang Wei[5], Xiaoyan Wang[5], Jingqi Wang[5], Hao Ding[5], Frank J. Manion[5], Jingcheng Du[5], David W. Bates[1,2], 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]Department of Neurology, Brigham and Women's Hospital, Boston, Massachusetts 02115
[5]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

Abstract

Background: Large language models (LLMs) have shown promising performance in various healthcare domains, but their effectiveness in identifying specific clinical conditions in real medical records is less explored. This study evaluates LLMs for detecting signs of cognitive decline in real electronic health record (EHR) clinical notes, comparing their error profiles with traditional models. The insights gained will inform strategies for performance enhancement.

Methods: This study, conducted at Mass General Brigham in Boston, MA, analyzed clinical notes from the four years prior to a 2019 diagnosis of mild cognitive impairment in patients aged 50 and older. We used a randomly annotated sample of 4,949 note sections, filtered with keywords related to cognitive functions, for model development. For testing, a random annotated sample of 1,996 note sections without keyword filtering was utilized. We developed prompts for two LLMs, Llama 2 and GPT-4, on HIPAA-compliant cloud-computing platforms using multiple approaches (e.g., both hard and soft prompting and error analysis-based instructions) to select the optimal LLM-based method. Baseline models included a hierarchical attention-based neural network and XGBoost. Subsequently, we constructed an ensemble of the three models using a majority vote approach.

Results: GPT-4 demonstrated superior accuracy and efficiency compared to Llama 2, but did not outperform traditional models. The ensemble model outperformed the individual models, achieving a precision of 90.3%, a recall of 94.2%, and an F1-score of 92.2%. Notably, the ensemble model showed a significant improvement in precision, increasing from a range of 70%-79% to above 90%, compared to the best-performing single model. Error analysis revealed that 63 samples were incorrectly predicted by at least one model; however, only 2 cases (3.2%) were mutual errors across all models, indicating diverse error profiles among them.

Conclusions: LLMs and traditional machine learning models trained using local EHR data exhibited diverse error profiles. The ensemble of these models was found to be complementary, enhancing diagnostic performance. Future research should investigate integrating LLMs with smaller, localized models and incorporating medical data and domain knowledge to enhance performance on specific tasks.

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

1. Introduction

Large Language Models (LLMs), neural models with billions of parameters trained on extensive, diverse text corpora, have exhibited remarkable capabilities in clinical language understanding tasks.[1–5] They offer distinct advantages over traditional rule-based and machine learning approaches, which are often trained from scratch on narrower clinical datasets.[6–8] Previous studies showed that LLMs achieved impressive performance in a variety of clinical natural language processing (NLP) tasks such as question answering, named entity recognition, and information extraction[1,2]. However, the effectiveness of LLMs in identifying specific clinical conditions within real medical records remains less explored. Their lack of explicit training on specific medical records may affect their accuracy.[9] This study aims to evaluate LLMs' performance in detecting signs of cognitive decline within clinical notes. We use this as a use case to explore their effectiveness and compare their error profiles with those of traditional models trained on a domain-specific corpus. The insights gained from this study will inform strategies for further enhancement.

Alzheimer's disease (AD) and related dementias (ADRD) affect millions of Americans,[10] significantly reducing patient quality of life and imposing substantial emotional and financial burdens,[11] with care costs projected to reach $1.1 trillion by 2050.[12] Existing treatments offer only temporary relief,[13] underscoring the urgent need for breakthroughs in AD/ADRD therapy.[14] Timely detection of cognitive decline signs can facilitate early interventions and clinical trial involvement for AD/ADRD.[15–18] Electronic health records (EHRs), particularly clinical notes, are critical resources for identifying early indicators of disease, yet traditional diagnostic tools and variability in screening practices complicate detection.[19–22] NLP offers a promising solution by efficiently analyzing large datasets and identifying subtle signs of decline not easily captured in traditional diagnostics.[23] Although studies have been conducted to identify cognitive decline using NLP,[7,24–26] the effectiveness of LLMs specifically in identifying cognitive decline through EHRs remains under-explored.

This research utilizes LLMs within HIPAA-compliant computing environments for a pioneering exploration of EHR note analysis for cognitive decline detection. It evaluates the effectiveness and interpretability of LLMs compared to conventional machine learning methods and examines the synergy between LLMs and machine learning to enhance diagnostic accuracy. To the best of our knowledge, this initiative is the first of its kind to employ LLMs in this capacity, representing a significant 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 healthcare system in Massachusetts, which has established secure, HIPAA-compliant cloud environments for deploying and evaluating LLMs with actual EHR data. Two LLMs were tested: the proprietary GPT-4[1] via Microsoft Azure OpenAI Service API, and the open-source Llama 2 (13B)[2] via an Amazon Elastic Compute Cloud (EC2) instance. Details on the cloud environments are provided in **Supplementary Material Section 1** and **Table S1**.

We utilized the same definition of cognitive decline and annotated datasets from a previous study.[19] Cognitive decline encompasses various progressive stages, from subjective cognitive decline (SCD) to mild cognitive impairment (MCI) to dementia. It can be identified through mentions of signs, symptoms, diagnostic evaluations, cognitive assessments, or treatment details in clinical notes. Transient cases, such as memory loss due to medication, were labeled as negative for cognitive decline.

The annotated datasets comprised sections of clinical notes from the four years prior to the initial diagnosis of mild cognitive impairment (MCI, ICD-10-CM code G31.84) in 2019, for patients aged 50 years or older.[19] Due to the low positive case rate across the sections, we used a list of expert-curated keywords (**Table 1**) to screen for sections likely indicating cognitive decline. **Table 2** shows that Dataset I, consisting of 4,949 keyword-filtered sections, was used to train two baseline models. For prompt development and LLM selection, 200 random samples from Dataset I (Dataset I-S) were used for performance assessment, while the remaining samples (Dataset I-A) were utilized for sample selection in prompt augmentation. Dataset II, which includes 1,996 random sections not subjected to keyword filtering, served for final testing.

The study received approval from the MGB Institutional Review Board, with a waiver of informed consent for study participants due to the secondary use of EHR data.

2.2. LLMs and Prompting Methods

Figure 1 (areas A and B) illustrates the two-step prompt engineering process: LLM selection and prompt improvement. Following previous studies, we divided the prompt into sections.[27] Supplementary **Figure S1** shows the prompt structure, which includes a required task description and optional sections for prompt augmentation, error analysis-based instructions, and additional task guidance. We were cautious about the potential impact of longer prompts, which might overwhelm the model, negatively affecting performance, response speed, and cost efficiency.[28–30]. Therefore, as an initial step, we evaluated the performance of the two LLMs using manual template engineering and a smaller sample size. This approach enabled us to select the superior model and its corresponding prompt template for further analysis.[31] The selection criterion was $accuracy = \frac{true\ positive + ture\ negative}{all\ cases}$, based on Dataset I-S. Using this metric and guided by the accuracy from Dataset I-S, we explored whether common prompt augmentation methods (both hard and soft prompting)[31] and error analysis-based instructions[32] could improve model performance. To ensure control over randomness and creativity, we adjusted the LLM's temperature hyperparameter to 0, providing a deterministic solution.[32]

2.2.1. LLMs Comparison and Selection

We utilized an intuitive manual template engineering approach to fine-tune the task description and additional task guidance for each LLM.[31] During the iterative refinement process, we focused on the following task descriptions for each LLM: 1) identifying evidence of cognitive decline in clinical notes; 2) displaying which keywords in the clinical notes informed its judgment on the assigned task; and 3) requiring LLM responses in

JSON format to facilitate straightforward parsing. Furthermore, we explored the possibility of adding additional task guidance to assist the LLM in its reasoning and enhance performance. Specifically, we considered two approaches: 1) requesting the LLM to provide reasoning for its judgments, and 2) incorporating our definition of cognitive decline directly into the prompt.

Manual Template Engineering: We fed each prompt to GPT-4 and Llama 2 separately. The responses from these LLM were classified into three categories (Supplementary **Table S2**): 1) effective and parseable: the LLM's response provides answers to both questions—whether cognitive decline was identified and which keywords were used for the decision—using a standard JSON format; 2) effective but not parseable: the LLM's response answers both questions, but does not adhere to the standard JSON format; 3) not effective: the LLM's response fails to answer either of the two questions. We assessed model effectiveness using 10 random samples from Dataset I. Our observations indicated that this sample size was sufficient for a meaningful comparison. If the effective response rate did not reach 100%, we manually adjusted the prompt template by paraphrasing or modifying optional content. This tuning process continued until no further improvement in the effective response rate was achieved after three consecutive attempts. Finally, we selected the prompt template that yielded the highest effective response rate for GPT-4 and Llama 2 separately.

Performance Comparison with Manually Crafted Templates: To select the optimal LLM, we compared the accuracy of GPT-4 and Llama 2 on Dataset I-S by providing the LLMs with manually crafted task descriptions and guidance.

2.2.2. Prompt Improvement

Prompt Augmentation: We explored prompt augmentation to determine if including five examples (five-shot prompting) enhances performance. We adopted five-shot prompting due to the maximum token limitation of GPT-4. Since the selection of examples for few-shot prompting can significantly affect model performance[31,33], we tested four different strategies, including both hard and soft prompting. To select the best strategy, we chose examples from Dataset I-A and evaluated model performance on Dataset I-S. The four example selection strategies were: 1) Hard Prompting - Random Selection: This strategy involves randomly selecting five samples. 2) Hard Prompting - Targeted Selection: We selected examples where the model had previously performed poorly, aiming to directly address its weaknesses. 3) Hard Prompting - K-Means Clustering-Aided Selection: This strategy involves selecting five samples from that are the centers of five clusters generated by k-means clustering. We utilized OpenAI's embedding model, *text-embedding-ada-002*[33], as features to ensure the examples are diverse and representative, which could be crucial for performance improvement. 4) Soft Prompting - Dynamic Selection: For each case in Dataset I-S, we automatically identified the top five most similar samples from Dataset I-A using OpenAI's embedding model, *text-embedding-ada-002,*[33] based on the k-nearest-neighbors algorithm. This process enabled us to provide the LLM with five samples that most closely resemble the current case, thereby guiding its decision-making.

Error Analysis-Based Instructions: We tested whether incorporating error analysis-based instruction into the prompt could improve performance.[32] To achieve this, we first conducted an error analysis of the LLM on Dataset I-S. Subsequently, we added a paragraph describing common errors that the LLM made and instructed it to pay attention to those errors when generating its response.

2.3. Baseline Machine Learning Models

We compared the performance of the LLM with two baseline machine learning models developed from our previous study: XGBoost[34] and a four-layer attention-based deep neural network (DNN),[7,35] which incorporated elements of a convolutional neural network, a bidirectional long-short term memory (LSTM) network, and an

attention model. These two models were the top performers compared to other traditional models in identifying cognitive decline in clinical notes.[19]

2.4.Ensemble Model

Finally, we investigated whether an ensemble model that combines predictions from both the LLM and traditional machine learning models could achieve better performance. The ensemble learning, which involves combining several different predictions from various models to formulate the final prediction, has proven to be an effective approach for enhancing performance.[36,37] To create the ensemble model, we determined the label by taking the majority vote from the LLM, the attention-based DNN, and XGBoost. The high diversity of the models included may enable the ensemble to correct errors made by individual models.[38]

2.5.Model Evaluation

We evaluated and compared the selected LLM, traditional models and the ensemble model on Dataset II using standard metrics: $precision = \frac{true\ positive}{true\ positive + false\ positive}$, $recall = \frac{true\ positive}{true\ positive + false\ negative}$, and $f1 - score = \frac{2 \times precision \times recall}{precision + recall}$. We used 0.5 as the cutoff point for calculating precision, recall, and F1 score for the baseline models.

2.6.Interpretation

Regarding interpretation, we listed keywords from the LLM's output that appeared more frequently than the average appearance time plus two standard deviations. We also identified keywords whose deep learning attention weights exceeded the mean weights by more than two standard deviations within individual sections, and keywords with an XGBoost information gain higher than the average value plus two standard deviations. Additionally, we included expert-curated keywords developed in our previous study as a reference.[19]

2.7.Error Analysis

We conducted two levels of error analyses. The first analysis assessed the selected LLM using various prompting strategies, including zero-shot, the best few-shot method, and the prompt with error analysis-based instructions. The second analysis evaluated the best-performing LLM with its optimal prompt, alongside the attention-based DNN, and XGBoost. Errors made by each model were analyzed and discussed by two biomedical informaticians and a physician. We quantified unique and overlapping errors made by each model using a Venn diagram.

3. Results

Dataset characteristics are illustrated in **Table 2**. The average length of the Dataset I sections was 850 characters (range: 26-9393), and that of the Dataset II sections was 464 characters (range: 26-14740). Dataset I contained 29.4% positive cases and Dataset II contained 3.5% positive cases.

3.1.LLM Selection and Prompt Selection

The effective response rate varied for each LLM using five different prompt templates (**Figure 2**). For GPT-4, Template 1, which includes a task description section and additional task guidance section as shown in Supplementary **Figure S1**, achieved a 100% effective response rate. Llama achieved its highest effectiveness at 80% when using Template 2, which only includes the task description section. GPT-4 and Llama 2, with their most effective prompts, achieved accuracies of 86.5% and 52.0% respectively on Dataset I-S. We therefore chose GPT-4 for subsequent analysis.

Prompt improvement result on Dataset I-S shows that the best prompt augmentation approach (Template 6) was soft prompting – dynamic five-shot, which had an 85% accuracy. However, adding error analysis-based instructions (Template 7) surpassed this, reaching an accuracy of 93%. Therefore, we decided to adopt error analysis-based instructions as our prompting strategy for subsequent analyses.

3.2.Performance Evaluation

GPT-4 achieved a precision of 71.6%, a recall of 91.3%, and an F1 score of 80.3%. Optimized hyperparameters for attention-based DNN and XGBoost are illustrated in Supplementary **Table S8**. Attention-based DNN achieved a precision of 77.1%, recall of 92.8, and F1 score of 84.2%. XGBoost model achieved a precision of 79.0%, recall of 92.8%, and F1 score of 85.3%. Notably, the ensemble model significantly improved overall performance, achieving a precision of 90.3%, a recall of 94.2%, and an F1 score of 92.2%.

3.3.Interpretation

Table 1 contains keywords identified through expert curation and exported by GPT-4, the attention-based DNN, and XGBoost. These keywords encompass a range of topics, including memory-related issues such as recall and forgetfulness, cognitive impairments, and dementia, with terms like "dementia" and "Alzheimer's." They also cover evaluation and assessment methods, referencing tools like the MoCA and MMSE. Compared to traditional AI models and expert-selected keywords, GPT-4 highlighted specific treatment options, notably "Aricept" and "donepezil," (Supplementary **Table S9**) which are important in managing dementia and Alzheimer's disease. Furthermore, GPT-4 explicitly identified specific diagnoses or conditions more than other models, with terms such as "mild neurocognitive disorder," "major neurocognitive disorder," and "vascular dementia." Additionally, GPT-4 exported keywords regarding the emotional and psychological effects of cognitive disorders, such as "anxiety," thus addressing aspects sometimes overlooked by other models.

3.4.Error Analysis

As illustrated in Supplementary **Figure S2**, when using different prompting strategies with GPT-4, some errors may be mitigated, while new ones could emerge that were not previously observed. Notably, adding error analysis-based instructions to the prompt yielded the best performance, with only 31 wrongly predicted cases in Dataset II. In contrast, the error profiles of GPT-4, attention-based DNN, and XGBoost exhibited much higher diversity (**Figure 3)**. We found that 63 cases were wrongly predicted by one or more models. GPT-4 accounted for 31 incorrect predictions, the attention-based DNN made 23 wrong predictions, and XGBoost was responsible

for 22 incorrect predictions. However, only 2 (3.2%) cases were wrongly predicted by all models. Four errors were common between GPT-4 and the attention-based DNN, three were common between GPT-4 and XGBoost, and eight were shared between the attention-based DNN and XGBoost.

All models were susceptible to misinterpreting signs or symptoms as indicative of unrelated clinical conditions. GPT-4 excelled in handling ambiguous terms and interpreting nuanced information, a frequent challenge for traditional AI. Unlike traditional AI, GPT-4 was not confused by negations and contextual details. However, it could sometimes overinterpret nuanced information or be overly conservative, failing to recognize cognitive decline despite strong evidence. It might also overlook underlying causes of clinical events like treatments or visits related to cognitive decline. Both GPT-4 and attention-based DNNs occasionally misread clinical testing results.

4. Discussion

Recently, LLMs have demonstrated remarkable performance on various NLP tasks, yet their ability to analyze clinical notes from EHR data remains underexplored, partly due to data privacy concerns. In this study, we established HIPAA-compliant secure environments for LLMs and used cognitive decline identification as a use case to test LLMs' capabilities in clinical note classification, thereby enhancing diagnostic tasks. Our contributions are threefold: 1) This study is the first to set up a secure cloud environment for GPT-4 and tested its ability to identify cognitive decline from clinical notes in EHR data; 2) We introduced a novel method for implementing NLP models for cognitive decline identification, achieving state-of-the-art performance with a significant lead over existing methods; 3) We discovered that although existing LLMs may not outperform traditional AI methods trained on a local medical dataset, their error profile differs distinctly, underscoring the significant potential of combining LLM with traditional AI models.

Our research demonstrated that prompt engineering using error analysis-based instructions significantly enhanced performance compared to zero-shot and prompt augmentation approaches, as it directly targeted the LLM's weaknesses. Nevertheless, the LLM did not surpass traditional AI in identifying cognitive decline, primarily because it was not specifically trained for this task.[9,39] While the LLM can generate a range of responses, it is prone to producing plausible but incorrect hallucinations. Nonetheless, it is valuable for its ability to operate without task-specific training, thereby complementing traditional AI, which requires specific training but often does not suffer from hallucinations.[40] In terms of interpretation, the LLM identified keywords overlooked by experts and traditional AI models, such as medications related to cognitive decline. Error analysis revealed that the LLM demonstrated superior handling of ambiguous or contextually complex information due to its transformer architecture.[3,4] However, LLMs misinterpreted or overlooked certain domain-specific medical tests and treatments. Future research should explore the integration of the LLM with smaller, localized models and knowledge bases to enhance performance on specific tasks.

Although our study has many strengths—for instance, it is the first to employ LLMs on unstructured EHR data for detecting cognitive decline—the results should be considered in light of several limitations. The LLMs used may not represent the most recent advancements (e.g., the recently released Llama 3 model) due to the rapid evolution of LLM technologies. While utilizing LLMs with a larger number of parameters (e.g., Llama 2-70 billion) may lead to better performance, this improvement comes with trade-offs, including higher computational demands and greater memory needs, posing challenges due to resource constraints. Additionally, our data are record-based and not patient-based (i.e., longitudinal), thus, the developed model may struggle to distinguish between reversible and progressive cognitive decline, and it remains unclear if patients recovered later based solely on a note from one time point. Therefore, developing an LLM-based early warning system for cognitive decline using longitudinal data would be a valuable direction for future research.

5. Conclusion

This study is among the first to utilize LLM within HIPAA-compliant cloud environments, leveraging real EHR notes for detecting cognitive decline. 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, significantly enhancing performance and improving diagnostic accuracy. Future research could investigate methods for incorporating domain-specific medical knowledge and data to enhance the capabilities of LLMs in healthcare-related tasks.

Acknowledgments

This study was funded by NIH-NIA R44AG081006.

Tables

Table 1. Keywords Contributing to the Identification of Positive Cognitive Decline Cases, Curated by Domain Experts and Extracted from AI Models.*

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

*The table lists keywords that had a high frequency of appearance in the LLM's output (i.e., the number of appearances is higher than the average appearance time plus two standard deviations); keywords whose attention weights (from the attention-based DNN) exceeded the mean weights plus two standard deviations within individual sections; and keywords with an information gain (XGBoost) higher than the average value plus two standard deviations. We found that keywords identified by AI models could significantly enrich the expert-curated keyword set. Notably, only GPT-4 identified keywords related to medications for cognitive decline.

Table 2. Dataset Characteristics.

Dataset	Description	Mean (range), characters	Positive Rate
Dataset I	4,949 note sections filtered with cognitive decline-related keywords	850 (26-9323)	29.4%
Dataset I-A	A random subset of Dataset I, containing 4,749 samples	848.8 (26-9323)	29.6%
Dataset I-S	A random subset of Dataset I, comprising 200 samples that do not overlap with Dataset I-A	870.7 (34-8353)	23.5%
Dataset II	1,996 random note sections without keyword filtering	464 (26-14740)	3.5%

Figures

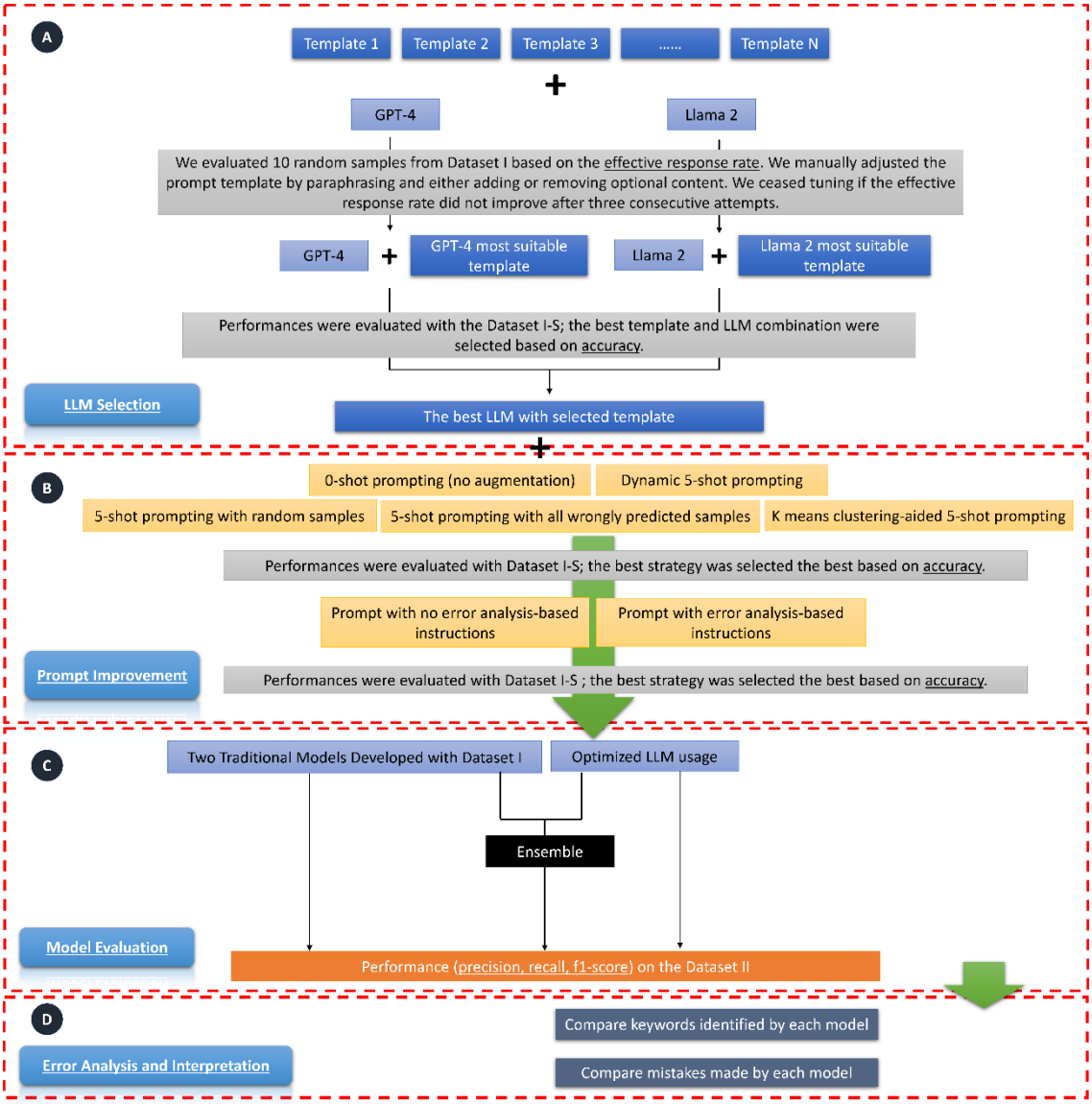

Figure 1. Study Design Overview.

The workflow consists of five parts: A) LLM Selection: We fed prompts, which contain task descriptions and may also include additional task guidance as illustrated in Supplementary **Figure S1**, to GPT-4 and Llama 2 separately. We used 10 random samples from Dataset I to select the most suitable template for each LLM. During this process, if the effective response rate (i.e., the rate at which the response answered the questions in the prompt) was not 100%, we manually adjusted the template for each model. If the effective response rate did not improve after three consecutive attempts, we ceased tuning and used the template that led to the highest effective response rate. We then selected the best LLM based on their accuracy on Dataset I-S. B) Prompt Improvement: This step includes two sub-steps: prompt augmentation and adding error analysis-based instructions. During prompt augmentation, we tested whether five-shot prompting could improve accuracy. We then assessed whether incorporating instructions following an error analysis of the LLM's output on Dataset I-S could enhance accuracy. C) Model Evaluation: We evaluated the selected LLM and two traditional machine learning models. We also tested the performance of an ensemble model, which took the majority vote of the three models as the predictive label. D) Interpretation and Error Analysis: For interpretation, we examined and compared keywords used by each model for prediction, in conjunction with those curated by domain experts. Lastly, we analyzed and compared errors made by each model.

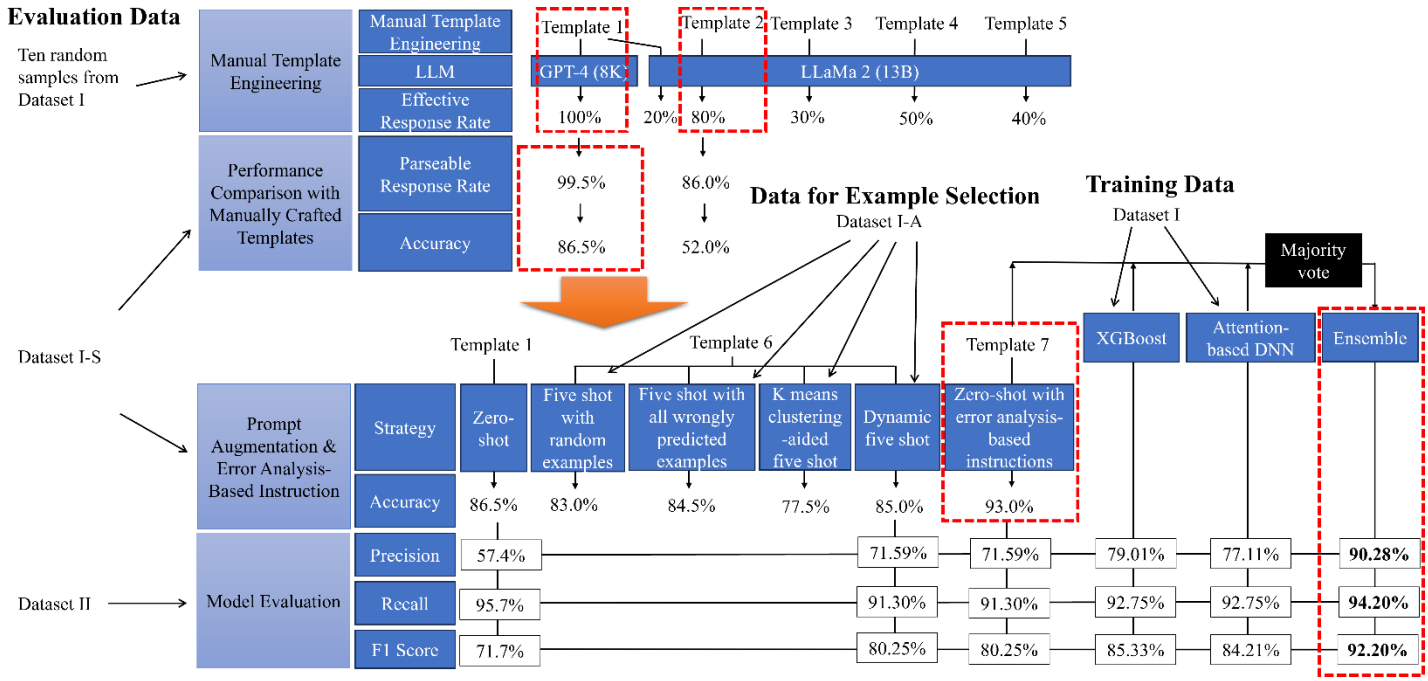

Figure 2. Evaluation Results Summary.

During the prompt template selection, Template 1 was selected for the GPT-4 model due to a 100% effective response rate; Template 2 was selected for the Llama 2 model as the effective response rate (80%) did not improve after three tuning attempts. Subsequently, we compared the two combinations with 200 samples from Dataset I-S and found that GPT-4 and Template 1 combination achieved significantly better accuracy (86.0%). We also discovered that five-shot prompting did not lead to improved performance; however, adding error analysis-based instructions (i.e., GPT-4 and Template 7 combination) increased the accuracy to 93% on Dataset I-S. Consequently, we opted to use Template 7 as the prompt template and GPT-4 as the LLM. In tests, we evaluated the performance of the XGBoost, the attention-based DNN, and the LLM. We found that XGBoost performed better: precision – 79.01%, recall – 92.75%, and F1 score – 85.33%. Notably, after assembling the three models using a majority vote, the ensemble model demonstrated significantly improved performance: precision – 90.11% (an 11.1% improvement), recall – 94.20% (a 1.45% improvement), and F1 score – 92.20% (a 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	The pt reports significant deficits that affect daily life. Referral does not indicate any cognitive deficits.	√	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 his son states he is working on putting a plan in place.	X	√	√	
Too inclusive 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 specific interventions on symptoms and problems related to cognitive decline	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 results	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) All models were susceptible to misinterpreting signs or symptoms as indicative of unrelated clinical conditions. 2) GPT-4 excelled in handling ambiguous terms and interpreting nuanced information, challenges that traditional AI frequently encounters. 3) Unlike traditional models, GPT-4 handles negations and contextual details more efficiently. 4) However, GPT-4 could sometimes overinterpret nuanced information or be overly conservative, failing to recognize whether a patient has cognitive decline despite strong evidence. 5) GPT-4 might also overlook certain medical domain knowledge, such as treatments or visits related to cognitive decline. 6) Both GPT-4 and attention-based DNNs occasionally misread clinical testing results, highlighting an opportunity for further improvement.

References

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

2. 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

3. Vaswani A, Shazeer N, Parmar N, et al. Attention is All you Need. In: *Advances in Neural Information Processing Systems*. Vol 30. Curran Associates, Inc.; 2017. Accessed April 11, 2024. https://papers.nips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

4. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Published online May 24, 2019. doi:10.48550/arXiv.1810.04805

5. Gemini Team, Anil R, Borgeaud S, et al. Gemini: A Family of Highly Capable Multimodal Models. Published online April 2, 2024. doi:10.48550/arXiv.2312.11805

6. Lemas DJ, Du X, Rouhizadeh M, et al. Classifying early infant feeding status from clinical notes using natural language processing and machine learning. *Sci Rep*. 2024;14(1):7831. doi:10.1038/s41598-024-58299-x

7. 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

8. Sheikhalishahi S, Miotto R, Dudley JT, Lavelli A, Rinaldi F, Osmani V. Natural Language Processing of Clinical Notes on Chronic Diseases: Systematic Review. *JMIR Med Inform*. 2019;7(2):e12239. doi:10.2196/12239

9. 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

10. 2023 Alzheimer's disease facts and figures. *Alzheimers Dement J Alzheimers Assoc*. 2023;19(4):1598-1695. doi:10.1002/alz.13016

11. 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

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

13. 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

14. 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

15. 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

16. 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.

17. 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

18. 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

19. 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

20. 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

21. 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

22. 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

23. Myszczynska MA, Ojamies PN, Lacoste AMB, et al. Applications of machine learning to diagnosis and treatment of neurodegenerative diseases. *Nat Rev Neurol*. 2020;16(8):440-456. doi:10.1038/s41582-020-0377-8

24. 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

25. 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

26. 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

27. Chen Q, Du J, Hu Y, et al. Large language models in biomedical natural language processing: benchmarks, baselines, and recommendations. Published online January 20, 2024. doi:10.48550/arXiv.2305.16326

28. Jiang H, Wu Q, Lin CY, Yang Y, Qiu L. LLMLingua: Compressing Prompts for Accelerated Inference of Large Language Models. In: Bouamor H, Pino J, Bali K, eds. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics; 2023:13358-13376. doi:10.18653/v1/2023.emnlp-main.825

29. Li L, Zhang Y, Chen L. Prompt Distillation for Efficient LLM-based Recommendation. In: *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*. CIKM '23. Association for Computing Machinery; 2023:1348-1357. doi:10.1145/3583780.3615017

30. Chen L, Zaharia M, Zou J. FrugalGPT: How to Use Large Language Models While Reducing Cost and Improving Performance. Published online May 9, 2023. doi:10.48550/arXiv.2305.05176

31. 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

32. 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

33. 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

34. 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

35. 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

36. Ganaie MA, Hu M, Malik AK, Tanveer M, Suganthan PN. Ensemble deep learning: A review. *Eng Appl Artif Intell*. 2022;115:105151. doi:10.1016/j.engappai.2022.105151

37. Mohammed A, Kora R. A comprehensive review on ensemble deep learning: Opportunities and challenges. *J King Saud Univ - Comput Inf Sci*. 2023;35(2):757-774. doi:10.1016/j.jksuci.2023.01.014

38. Mao S, Chen JW, Jiao L, Gou S, Wang R. Maximizing diversity by transformed ensemble learning. *Appl Soft Comput*. 2019;82:105580. doi:10.1016/j.asoc.2019.105580

39. Hughes A. Phi-2: The surprising power of small language models. Microsoft Research. Published December 12, 2023. Accessed April 18, 2024. https://www.microsoft.com/en-us/research/blog/phi-2-the-surprising-power-of-small-language-models/

40. Azamfirei R, Kudchadkar SR, Fackler J. Large language models and the perils of their hallucinations. *Crit Care*. 2023;27:120. doi:10.1186/s13054-023-04393-x