npj | digital medicine

Article

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-025-01834-5

A GPT-4o-powered framework for identifying cognitive impairment stages in electronic health records

Yu Leng[1,8], Yingnan He[1,8], Samad Amini[2], Colin Magdamo[1,3], Ioannis Paschalidis[2], Shibani S. Mukerji[1,3], Lidia M. V. R. Moura[1,3], M. Brandon Westover[3,4], Ana-Maria Vranceanu[3,5], Christine S. Ritchie[3,6], Deborah Blacker[3,5,7], John R. Dickson[1,3] & Sudeshna Das[1,3] ✉

Alzheimer's Disease and Related Dementias (ADRD) pose a major public health challenge, with a critical need for accurate and scalable tools for detecting cognitive impairment (CI). Readily available electronic health records (EHRs) contain valuable cognitive health data, but much of it is embedded in unstructured clinical notes. To address this problem, we developed a GPT-4o-powered framework for CI stage classification, leveraging longitudinal patient history summarization, multi-step reasoning, and confidence-aware decision-making. Evaluated on 165,926 notes from 1002 Medicare patients from Mass General Brigham (MGB), our GPT-4o framework achieved high accuracy in CI stage classification (weighted Cohen's kappa = 0.95, Spearman correlation = 0.93), and outperformed two other language models (weighted Cohen's kappa 0.82–0.85). Our framework also achieved high performance on Clinical Dementia Rating (CDR) scoring on an independent dataset of 769 memory clinic patients (weighted Cohen's kappa = 0.83). Finally, to ensure reliability and safety, we designed an interactive AI agent integrating our GPT-4o-powered framework and clinician oversight. This collaborative approach has the potential to facilitate CI diagnoses in real-world clinical settings.

Alzheimer's Disease and Related Dementias (referred to hereafter as dementia) describe a group of related neurodegenerative disorders affecting over 6 million people over the age of 65 in the United States and represent a large and growing problem in the 21st century[1]. Timely diagnosis of dementia is crucial for implementing effective interventions and treatment strategies to enhance the quality of life for both persons living with dementia and their families[2,3]. However, dementia remains under-recognized, under-diagnosed, and under-reported in healthcare records[1,4–6]. Additionally, staging cognitive impairment (CI), particularly in cases of mild cognitive impairment (MCI), is crucial for understanding disease progression and guiding precise treatment for patients. However, MCI is particularly challenging to diagnose due to its subtle symptoms, variability in presentation, and compensatory mechanisms that may mask symptoms, as well as fundamentally arbitrary boundaries[1].

Automated identification of CI has the potential to facilitate clinical diagnosis and dementia research. The Electronic Health Record (EHR), which contains detailed and comprehensive health history, clinical notes, and other health-system interaction information, offers readily available data with great potential for such healthcare applications[7]. Signs of cognitive or behavioral dysfunction are often present in clinical notes, but clinicians may not make a formal diagnosis or prescribe medication for multiple reasons, including lack of time or expertise, patient resistance, or concerns about the emotional impact on patients and families[6,8–13]. Thus, critical insights are often buried in unstructured clinician notes, making them not readily accessible for clinical decision-making or research. Traditional methods for extracting this information involve labor-intensive manual reviews, which are not only time-consuming but also prone to inconsistencies and errors. To address this gap, several studies have leveraged natural language processing (NLP) techniques on EHR notes for CI detection. For example, Gilmore-Bykovskyi et al. applied NLP to notes to detect CI in acute-care patients with stroke and hip fractures[14]. Similarly, NLP was used to identify CI in patients undergoing formal cognitive

[1]Department of Neurology, Massachusetts General Hospital, Boston, MA, USA. [2]Boston University, Boston, MA, USA. [3]Harvard Medical School, Boston, MA, USA. [4]Department of Neurology, Beth Israel Deaconess Medical Center, Boston, MA, USA. [5]Department of Psychiatry, Massachusetts General Hospital, Boston, MA, USA. [6]Mongan Institute Center for Aging and Serious Illness and the Division of Palliative Care and Geriatric Medicine, Massachusetts General Hospital, Boston, MA, USA. [7]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA. [8]These authors contributed equally: Yu Leng, Yingnan He. ✉e-mail: sdas5@mgh.harvard.edu

evaluations through the Mayo Clinic Study of Aging[15] and in participants of the UCLA Alzheimer's and Dementia Care program[16]. These studies along with others demonstrated the potential of incorporating NLP on clinical notes for automated text analysis and CI detection[14–18]. However, they primarily relied on fixed pattern recognition techniques for extraction of CI from clinical notes. More recent investigations have also explored the application of transformer-based and large-language models (LLMs), which have achieved significant success in various fields including healthcare[19]. These advanced models better capture context and complex relationships among words, showing considerable promise in identifying CI within clinical notes[20–26].

Despite these advancements, to our knowledge, no previous efforts have applied the latest generation of generative pre-trained transformer (GPT) models to the task of identifying CI stages from EHR clinical notes. In this study, we introduce a GPT-4o-powered framework for automating the extraction and interpretation of cognitive data from EHRs and evaluated it using two datasets. First, we evaluated performance in classification of CI stages—Cognitively Unimpaired (CU), Mild Cognitive Impairment (MCI), and Dementia—using a dataset of 1002 Medicare fee-for-service patients from Mass General Brigham (MGB) Healthcare Accountable Care Organization (ACO). The performance of GPT-4o in the framework was compared with several other deep learning models for classifying CI stages, allowing us to evaluate them against GPT-4o's language understanding capabilities and gain insights into the potential of GPT-4o for automated CI stage classification in clinical settings. Second, we evaluated the GPT-4o-powered framework for automated assignment of global Clinical Dementia Rating (CDR) score using specialist notes from patients who visited the memory clinic at MGB. Further, given the broader opportunities for AI agents in medicine[27], we designed an interactive AI agent workflow that integrates our GPT-4o-powered framework, aiming to enable real-time interaction and decision support for cognitive diagnoses.

Results

Study dataset

We developed our GPT-4o-powered framework for classifying stage of CI using a dataset of 1002 patients from a previous study by Moura et al.[28]. These patients were sampled from three groups: (1) those with at least one dementia-related ICD-10 diagnosis code in their EHR; (2) those seen by at least one dementia-related specialist (neurologist, neuropsychologist, psychiatrist, geriatric psychiatrist, geriatrician) but with no such code; (3) those meeting neither of these conditions[28]. Each patient's EHR data between 01/01/2016 and 12/31/2018 was reviewed by an expert physician to screen for cognitive concerns (i.e., any documented suspicion/concern of memory or cognitive decline). Patients were also assigned a syndromic diagnosis: "No MCI or Dementia Diagnosis (Normal)", "Normal vs. MCI", "MCI", "MCI vs. Dementia", "Dementia", "Unknown". The syndromic diagnoses were assigned a confidence level of 1–4 to reflect the degree of certainty in assigning syndromic diagnoses: 1 = not at all confident (barely guessing), 2 = mildly confident, 3 = moderately confident, and 4 = highly confident. The borderline cases Normal vs. MCI and MCI vs dementia patients were excluded to get three final categories: CU (Normal), MCI, and Dementia, resulting in a dataset of 814 patients. In a second experiment, all five diagnostic categories were included, yielding a dataset of 904 patients.

We also evaluated the GPT-4o-powered framework's potential in automatically assigning a global CDR, using a dataset comprised of visit notes from 769 unique patients seen in the memory clinic at MGB between February 2016 and July 2019. These patients consented to be part of a registry which recorded the global CDR score and diagnoses at their visit, along with other data. To construct the evaluation dataset, we restricted the data to office visits and telemedicine encounters at the MGB memory clinic and selected the latest note for each unique patient to ensure a single representative entry per individual. We filtered out any encounters with missing text. Since these memory specialist notes have detailed information on cognitive evaluation, our goal was to evaluate whether we could automatically create structured datasets of the more detailed cognitive staging in expert clinician's global CDR scores.

Table 1 presents the demographic and clinical characteristics of the two study datasets. For CI stage classification (Table 1a), the final cohort after screening consisted of 814 Medicare fee-for-service patients[28], with 528 (64.9%) categorized as CU (Normal), 62 (7.6%) with MCI, and 224 (27.5%) with dementia. The mean age increased across the groups, from 74.8 years in the CU (Normal) group to 82.1 years in the dementia group. The demographics and clinical characteristics of the second experiment which included all five diagnostic categories—CU (Normal), Normal vs. MCI, MCI, MCI vs. Dementia, and Dementia—are provided in Supplementary Table 1. For assigning global CDR (Table 1b), the dataset comprised of the latest visit notes of 769 unique patients, demographics are stratified by global CDR scores (0, 0.5, 1, 2, 3).

Development of GPT-4o-powered framework and evaluation in CI stage classification

We designed an end-to-end GPT-4o approach that integrates multi-note summarization and structured multi-step reasoning to support long-term patient history tracking and multiple clinical encounters assessment for cognitive diagnoses, and compared it to three other frameworks (Fig. 1). Our framework processes unstructured clinical notes from EHRs in two key stages. First, all available clinical notes for each patient are aggregated and segmented into smaller chunks, ensuring comprehensive coverage while adhering to token limits. Using multi-note summarization, GPT-4o generates concise representations of these note chunks, capturing critical cognitive and diagnostic details at each time point. These intermediate summaries are then further distilled into a "summary of summaries", enabling the model to integrate longitudinal information and maintain contextual continuity. This structured summary serves as the foundation for GPT-4o's final classification, where it infers the CI stage—Cognitively Unimpaired (CU), Mild Cognitive Impairment (MCI), or Dementia—along with an associated confidence level. By implementing this two-step summarization strategy, our framework enhances GPT-4o's ability to track complex patient histories, perform multi-step reasoning, and utilize long-term memory across multiple encounters. Detailed examples of the prompts and corresponding responses are provided in Supplementary Note 1a (summarization) and Supplementary Note 1b (classification).

Figure 2a illustrates the confusion matrix of our GPT-4o-powered framework for CI stage classification, demonstrating strong agreement between actual labels and GPT-4o-generated labels. The overall weighted kappa score for the framework's classification on CI stages is 0.95. Of the 260 patients that were correctly identified as cognitively impaired (MCI or dementia) by the framework, 58 (22.3%) did not have a dementia-related diagnosis code[28]. In addition, per-class performance metrics of CI stage classification are reported in Supplementary Table 2. Adjudicator confidence scores were substantially lower for MCI (mean = 2.52) compared to CU (3.22) and Dementia (3.63), confirming the diagnostic uncertainty in this category (Kruskal-Wallis $H = 84.95$, $p < 0.001$).

A stratification analysis based on the clinical adjudicator's confidence levels revealed a clear trend: cases adjudicated with higher confidence by physicians demonstrated stronger alignment between the framework's classification and the physician's diagnosis (Fig. 2b). This trend indicates that cases rated with higher confidence by physicians were also those where GPT-4o-powered framework performs exceptionally well. Figure 2c displays the confusion matrix between physician's confidence level in the adjudication and confidence level generated by GPT-4o in the classification module of the framework. There was strong agreement between physicians and framework at the highest confidence level, suggesting that GPT-4o is consistent with physicians' reasoning patterns when assigning high confidence. We also tested an alternate measure of uncertainty, log-probabilities

Table 1 | Patient characteristics at baseline

(a) Dataset for Classification of CI Stage						
Characteristics		**Cognitive Impairment Stage**				
	Total *N* = 814	**CU (Normal) *N* = 528 (64.9%)**		**MCI *N* = 62 (7.6%)**		**Dementia *N* = 224 (27.5%)**
Age (Mean and SD)	76.9 (7.5)	74.8 (6.5)		76.8 (6.3)		82.1 (7.5)
Sex						
Female	482 (59.2%)	302 (57.2%)		30 (48.4%)		150 (67.0%)
Male	332 (40.8%)	226 (43%)		32 (52%)		74 (33%)
Race						
Asian	6 (0.7%)	4 (0.8%)		1 (1.6%)		1(0.4%)
Black or African American	22 (2.7%)	12 (2.3%)		3 (4.8%)		7 (3.1%)
Other	10 (1.2%)	8 (1.5%)		0 (0.0%)		2 (0.9%)
White	758 (93.1%)	496 (93.9%)		54 (87.1%)		208 (92.9%)
Unavailable or Not Reported	3 (0.4%)	1 (0.2%)		2 (3.2%)		0 (0.0%)
Ethnicity						
Hispanic or Latino	11 (1.4%)	8 (1.5%)		1 (1.6%)		2 (0.9%)
Not Hispanic or Latino	740 (90.9%)	484 (91.7%)		56 (90.3%)		200 (89.3%)
Unavailable or Not Reported	63 (7.7%)	36 (6.8%)		5 (8.1%)		22 (9.8%)
Education						
High School or Less	225 (27.6%)	124 (23.5%)		16 (25.8%)		85 (37.9%)
Some College	44 (5.4%)	33 (6.2%)		4 (6.5%)		7 (3.1%)
College	295 (36.2%)	202 (38.3%)		22 (35.5%)		71 (31.7%)
Graduate or More	107 (13.1%)	74 (14.0%)		9 (14.5%)		24 (10.7%)
Unavailable or Not Reported	142 (17.4%)	95 (18.0%)		11 (17.7%)		37 (16.5%)
(b) Dataset for Assignment of Global CDR						
Characteristics		**Global CDR Score**				
	Total *N* = 769	**0 *N* = 38 (5%)**	**0.5 *N* = 267 (35%)**	**1 *N* = 218 (28%)**	**2 *N* = 179 (23%)**	**3 *N* = 67 (9%)**
Age (Mean and SD)	77.9 (8.0)	72.3 (9.5)	76.6 (7.5)	78.3 (7.7)	80.2 (7.4)	79.1 (8.6)
Sex						
Female	397 (51.6%)	21 (55.3%)	111 (41.6%)	120 (55.0%)	109 (60.9%)	36 (53.7%)
Male	372 (48.4%)	17 (44.7%)	156 (58.4%)	98 (45.0%)	70 (39.1%)	31 (46.3%)
Race						
American Indian or Alaska Native	1 (0.1%)	0 (0.0%)	1 (0.4%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Asian	6 (0.8%)	0 (0.0%)	1 (0.4%)	2 (0.9%)	2 (1.1%)	1 (1.5%)
Black or African American	11 (1.4%)	0 (0.0%)	2 (0.7%)	3 (1.4%)	4 (2.2%)	2 (3.0%)
Other	2 (0.3%)	0 (0.0%)	0 (0.0%)	1 (0.5%)	0 (0.0%)	1 (1.5%)
White	285 (37.1%)	15 (39.5%)	95 (35.6%)	69 (31.7%)	75 (41.9%)	31 (46.3%)
Unknown or Not Reported	464 (60.3%)	23 (60.5%)	168 (62.9%)	143 (65.6%)	98 (54.7%)	32 (47.8%)
Ethnicity						
Hispanic or Latino	23 (3.0%)	1 (2.6%)	4 (1.5%)	8 (3.7%)	5 (2.8%)	5 (7.5%)
Not Hispanic or Latino	245 (31.9%)	10 (26.3%)	80 (30.0%)	63 (28.9%)	64 (35.8%)	28 (41.8%)
Unknown or Not Reported	501 (65.1%)	27 (71.1%)	183 (68.5%)	147 (67.4%)	110 (61.5%)	34 (50.7%)
Education						
High School or Less	171 (22.2%)	6 (15.8%)	42 (15.7%)	57 (26.1%)	49 (27.4%)	17 (25.4%)
College	91 (11.8%)	2 (5.3%)	33 (12.4%)	26 (11.9%)	23 (12.8%)	7 (10.4%)
Graduate or More	131 (17.0%)	11 (28.9%)	48 (18.0%)	26 (11.9%)	30 (16.8%)	16 (23.9%)
Unknown or Not Reported	376 (48.9%)	19 (50.0%)	144 (53.9%)	109 (50.0%)	77 (43.0%)	27 (40.3%)

(*logprobs*), that resulted in worse weighted kappa scores (0.17) (Supplementary Fig. 1) compared to the GPT-4o-assigned confidence scores (0.25).

For the second experiment including all five diagnostic categories — CU (Normal), Normal vs. MCI, MCI, MCI vs. Dementia, and Dementia— the evaluation yielded a weighted Cohen's kappa of 0.91. As expected, model performance was strongest for clearly defined categories: CU (Normal) and Dementia, and lowest for borderline cases: Normal vs. MCI and MCI vs. Dementia (Supplementary Table 3). Interestingly, this mirrors the average confidence levels assigned by expert adjudicators, indicating that these transitional stages are inherently ambiguous even for clinicians (Supplementary Table 3).

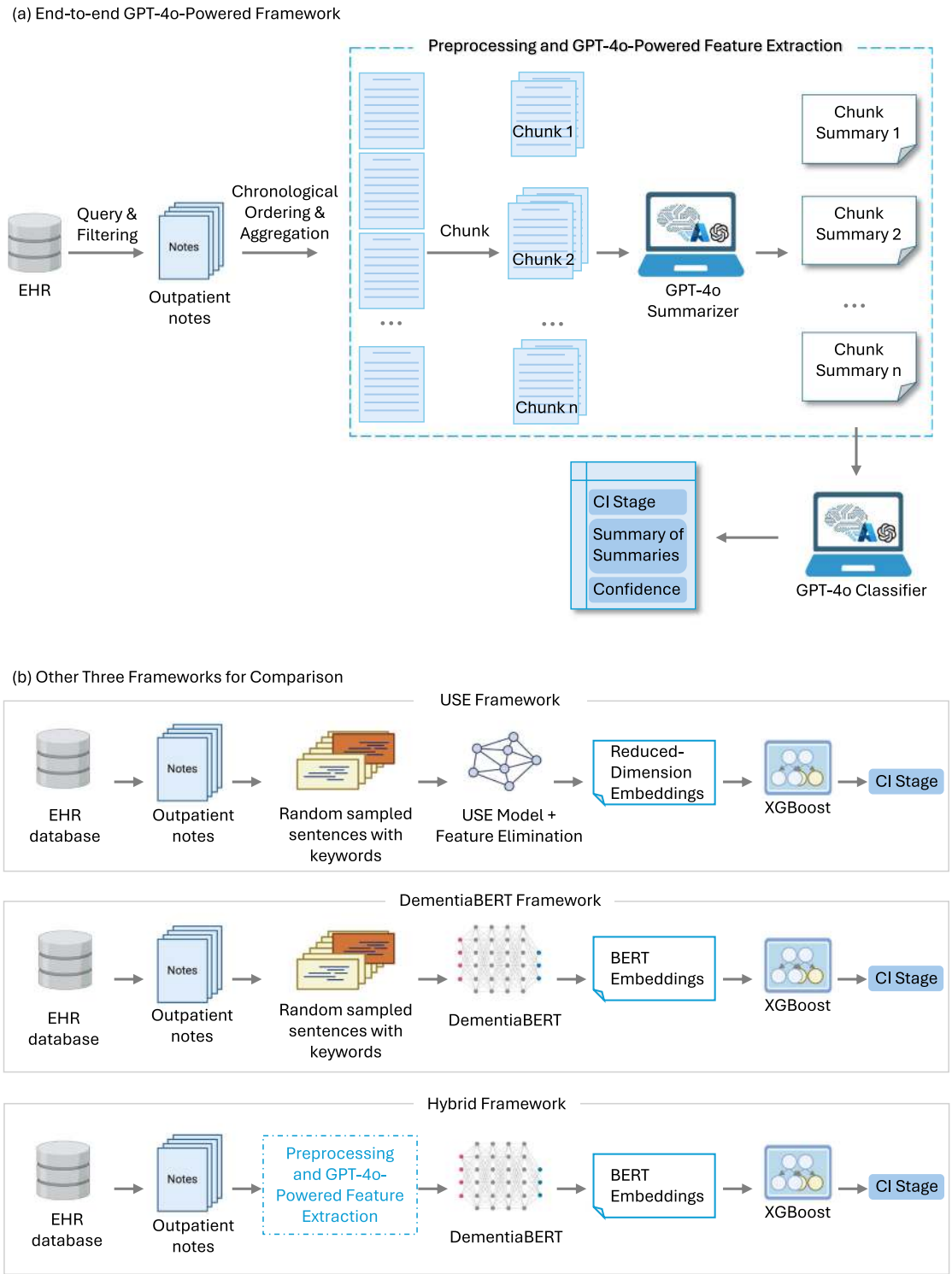


Fig. 1 | Overview of the Workflow for Cognitive Impairment (CI) Staging Across Four Frameworks. a End-to-End **GPT-4o-Powered Framework**: our framework, an End-to-end GPT-4o approach using multi-note summaries, chunked and summarized further into a "summary of summaries," outputting CI stage, summary and confidence level. **b** Other Three Frameworks for Comparison. **USE Framework**: Keyword-based sentence extraction with Universal Sentence Encoder (USE) embeddings, Recursive Feature Elimination (RFE), and XGBoost for classification. **DementiaBERT Framework**: Keyword-based sentence extraction with DementiaBERT embeddings (fine-tuned on dementia-related clinical language) and XGBoost classification. **Hybrid Framework**: GPT-4o-generated summaries of clinical notes, chunked and embedded using DementiaBERT, with XGBoost for classification. The figure was partially created in BioRender. He, Y. (2025) https://BioRender.com/4i93myx.

Comparison of performance across embedding and large language models in CI stage classification

We compared the performance of four frameworks across ten folds using weighted Cohen's kappa as the evaluation metric (Fig. 3a). The models are USE framework (Sentence extraction + USE Model + XGBoost), DementiaBERT framework (Sentence extraction + DementiaBERT + XGBoost), Hybrid framework (GPT-4o Summaries + DementiaBERT + XGBoost), and our GPT-4o-powered framework (see Fig. 1 and Methods). Overall, all

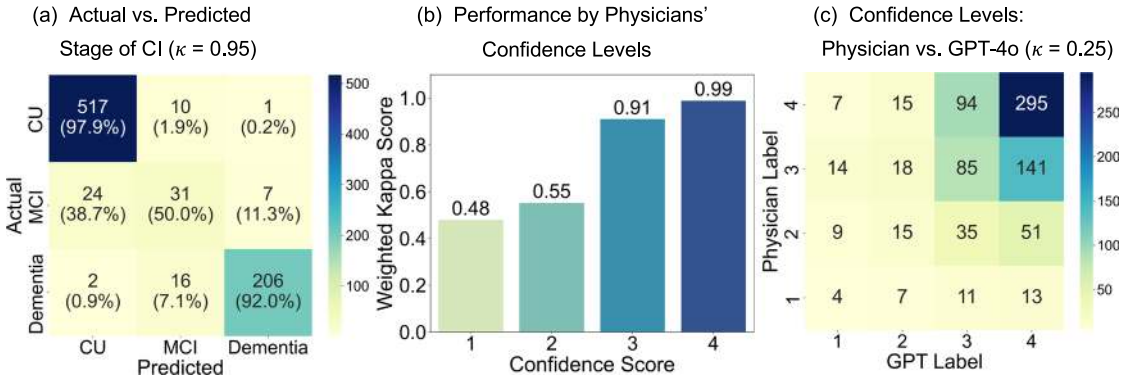
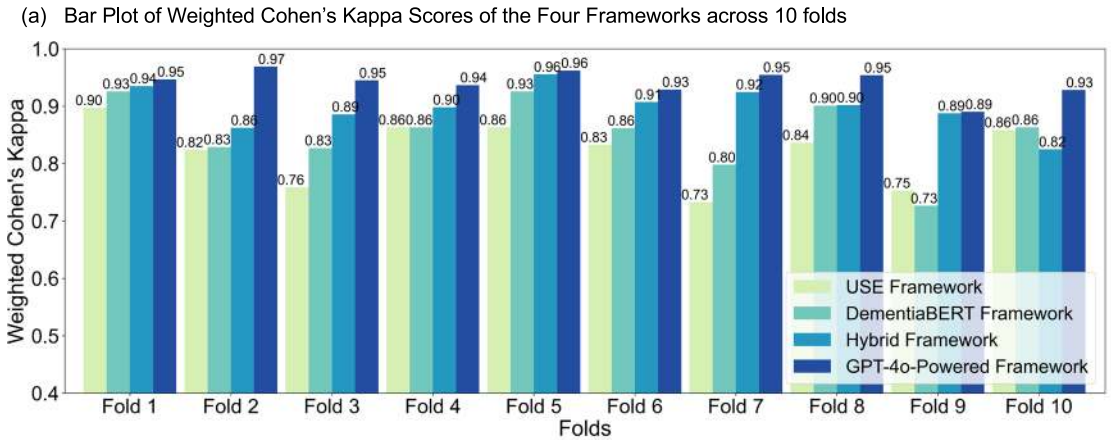


Fig. 2 | Performance and Confidence Analysis of GPT-4o-Powered Framework. **a** Framework Performance: Confusion matrix comparing actual versus GPT-4o predicted cognitive impairment (CI) stages: CU, MCI, Dementia. Darker colors indicate higher counts. **b** Performance Analysis Stratified by Physician Confidence scores: Bar plot of weighted Cohen's kappa scores stratified by physicians' confidence levels. Higher confidence scores (3 and 4) correspond to greater alignment with ground truth. **c** Comparison of Physician and GPT-4o confidence scores: Heatmap comparing confidence levels assigned by physicians versus GPT-4o. Darker colors represent higher case counts. Abbreviations: CU Cognitively Unimpaired, MCI Mild Cognitive Impairment.



(b) Multi-Metric Evaluation of the Four Frameworks Performance

Frameworks	Evaluation Metrics (Mean ± SD over 10 Folds)		
	Cohen's Kappa Score	Spearman's Rank Correlation	Baccianella's adapted MSE
USE	0.82 (0.06)	0.82 (0.05)	0.07 (0.02)
DementiaBERT	0.85 (0.06)	0.85 (0.06)	0.06 (0.02)
Hybrid	0.90 (0.04)	0.89 (0.04)	0.04 (0.01)
GPT-4o-Powered	0.95 (0.02)	0.93 (0.02)	0.02 (0.01)

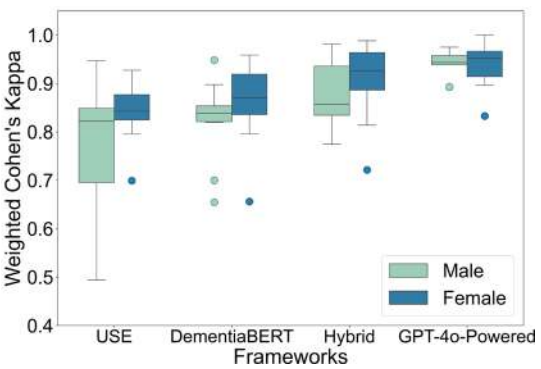


Fig. 3 | Comparison of Framework Performance. USE Framework: Keyword-based sentence extraction with Universal Sentence Encoder (USE) embeddings and XGBoost classification. **DementiaBERT Framework**: Keyword-based sentence extraction with DementiaBERT embeddings (fine-tuned on dementia-related clinical language) and XGBoost classification. **Hybrid Framework**: GPT-4o-generated summaries with DementiaBERT embeddings and XGBoost classification. **GPT-4o-Powered Framework**: an End-to-end GPT-4o approach using GPT-4o-generated summaries and GPT-4o classification. **a** Comparison of Weighted Cohen's Kappa Scores of the Four Models: Bar plot of weighted Cohen's kappa scores for four models across 10 cross-validation folds. Each bar represents the kappa score for a specific model on each fold. **b** Multi-Metric Evaluation of the Four Models Performance: Table summarizing the performance of each model across three evaluation metrics: Cohen's kappa score, Spearman's Rank Correlation, and Baccianella's adapted MSE. Mean and standard deviation values are provided over 10 folds. **c** Box Plot of the Weighted Cohen's Kappa Scores of the Four Models Stratified by Sex: Comparison of kappa scores across the four models, stratified by sex (Male and Female), with p-values indicating statistical tests for differences in performance between male and female groups.

frameworks demonstrate strong and consistent performance across folds, with Cohen's Kappa values predominantly above 0.80, indicating high agreement with syndromic diagnoses labeled by expert physicians. Among the four frameworks, GPT-4o-powered framework consistently shows the highest performance across almost every fold, suggesting that our framework consistently outperforms the others in terms of agreement. This indicates that the end-to-end GPT-4o-powered approach for summarization and classification captures nuanced information on cognitive status within clinical notes. The two DementiaBERT-based frameworks also perform strongly. Specifically, DementiaBERT framework uses the same preprocessing methods and classifier as USE framework but outperforms it across most folds, indicating that the domain-specific fine-tuned DementiaBERT as embedding model captures more meaningful representations of the extracted sentences compared to the general-purpose USE model. Additionally, while both DementiaBERT framework and Hybrid framework are based on DementiaBERT, the Hybrid framework which uses GPT-4o summaries instead of traditional sentence extraction shows a notable improvement over DementiaBERT framework across most folds, further highlighting GPT-4o's effective summarization capabilities. In summary, the results indicate that the GPT-4o-powered framework consistently achieves the highest performance, followed by the DementiaBERT-based frameworks with domain-specific fine-tuning, with the USE model showing comparatively lower performance.

In Fig. 3b, we present the performance of four models in CI stage classification, evaluated across three metrics to provide a comprehensive assessment of their performance in ordinal classification. The Spearman correlation scores for Models 1 to 4 are 0.82, 0.85, 0.89, and 0.93, respectively. This is consistent with Cohen's kappa score and reinforces that the model effectively maintains the correct ordering of classes. The Baccianella's MSE of the four models are 0.07, 0.06, 0.04, and 0.02, respectively. The performance of every model remains consistent across multiple evaluation metrics, with each model achieving high agreement (kappa), strong rank-order correlation (Spearman), and low error rates (Baccianella's MSE), underscoring the reliability of the methods applied. Additionally, the trend of performance improvement from USE framework, DementiaBERT framework and Hybrid framework to GPT-4o-powered framework, as observed in Cohen's kappa scores, is also reflected in both Spearman's Rank Correlation and Baccianella's MSE. This trend across all metrics highlights a clear progression in model performance and collectively reinforces GPT-4o's strong capabilities for classification of CI stages.

Figure 3c presents a subgroup analysis of sex for each framework, comparing performance across male and female patients. For all four frameworks, the p-values of paired t-test were all greater than 0.05, suggesting that the differences in performance between male and female subgroups were not statistically significant. That is, no framework demonstrated evidence of sex-based bias in their classifications of CI stages.

Evaluation of GPT-4o-powered framework in CDR assignment and prompt engineering

As the GPT-4o-powered framework demonstrated a strong performance in identifying CI stage, we further evaluated its potential in assigning global CDR, a more complex task requiring detailed measure of cognitive and functional performance across six domains. To address this complexity, we explored advanced GPT prompt engineering and Retrieval-Augmented Generation (RAG) techniques to optimize the framework's ability to capture information for assigning global CDR. Figure 4a–c present the normalized confusion matrices for the three approaches, illustrating agreement between actual and predicted CDR scores and misclassifications for each method (see Supplementary Fig. 2 for count-based confusion matrices). Across all approaches, GPT-4o-powered framework consistently predicted a higher CDR than the actual condition. The framework achieved weighted Cohen's kappa scores of 0.79, 0.80, and 0.83 with three different prompting and augmentation methods: structured answer template, RAG-enabled, and confidence level with domain count, respectively (Fig. 4d). The Spearman correlation for these approaches were 0.81, 0.81, and 0.83, while the MSE values were 0.12, 0.11, and 0.10, respectively. Although some errors remain, GPT-4o-powered framework's performance in assigning global CDR is consistent and robust across multiple evaluation metrics, collectively demonstrating its effectiveness in handling ordinal classification. Furthermore, the performance trends are consistent across experiments, demonstrating that improvements in one metric are mirrored by improvements in the others. We also included detailed per-class performance metrics in Supplementary Table 4. Stratification analysis of confidence levels generated by GPT-4o with prompting revealed that decisions made with high confidence had the highest weighted Cohen's kappa, while predictions with low or medium confidence had lower values, as expected (Low: 0.40; Medium: 0.56; High: 0.84). Interestingly, GPT-4o exhibited "overconfidence," with more than two-thirds of its predictions ($n = 618$) rated as high confidence and only a few ($n = 2$) rated as low. Detailed examples of the prompts and corresponding responses can be found in Supplementary Note 2a (GPT-4o with Structured Guidance), Supplementary Note 2b (RAG-enabled GPT-4o) and Supplementary Note 2c (GPT-4o with Domain Count and Confidence Level).

Finally, we investigated the association of the documentation of the six CDR domains in the visit note and GPT-4o's confidence levels (Medium, High) in assigning global CDR, using a single logistic regression model with six domains as predictors. Figure 4e shows coefficients of CDR domains in GPT-4o's confidence of assigning global CDR. The domains of Orientation, Judgment, and Problem Solving, and Home and Hobbies play a more substantial role in determining the model's confidence. Notably, the Memory domain's coefficient was not statistically significant ($p > 0.05$) and showed an unusually large standard error. Further analysis revealed quasi-complete separation: nearly all notes mentioning memory corresponded to high-confidence predictions. To address this, we applied Lasso-penalized logistic regression. The Memory coefficient was substantially reduced—indicating instability due to separation—but remained higher than others. Importantly, the coefficients for the other domains remained largely unchanged, suggesting that memory was influential but less stable compared to the other predictors (Supplementary Table 5).

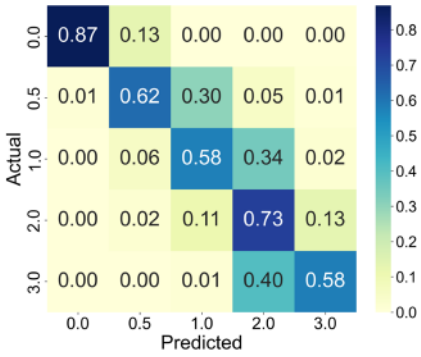
Design of GPT-4o-powered AI agent for enhancing clinical decision processes

We designed an interactive AI agent workflow that integrates our GPT-4o-powered framework to assist clinicians and researchers in cognitive diagnosis using EHRs (Supplementary Fig. 3). The AI agent operates through three core components: (1) an LLM-based Query Builder, which translates human prompts into structured queries to extract relevant patient notes and data from the EHR database; (2) our GPT-4o-powered framework, which processes clinical notes and generates CI staging and summaries; and (3) an LLM-based chatbot, which facilitates real-time interaction by responding to user inquiries using patient cognitive status summarized by GPT-4o-powered framework, and providing decision support. This design enables an efficient AI-augmented workflow for cognitive diagnoses related tasks, such as identifying undiagnosed cognitive impairment, tracking disease progression, etc., ultimately enhancing physician decision-making as well as researchers' understanding of CI patterns.

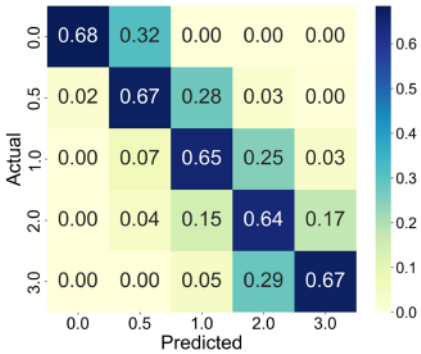
Discussion

In this study, our GPT-4o-powered framework demonstrated strong performance (weighted kappa 0.95) in classifying stages of CI. The performance comparison across the four LLM frameworks highlights the advantages of integrating GPT-4o for both note summarization and classification of CI stage. A clear trend was demonstrated in the results: as the level of GPT-4o integration increases, so does model performance in staging CI. This may be due to GPT-4o's advanced ability to interpret and summarize complex, unstructured clinical language. Moreover, GPT-4o outputs clear rationale for the classification (see examples in Supplementary Note 1b), thus providing evidence that can be checked for accuracy. To support reproducibility and transparency, we have included detailed descriptions of the GPT-4o setup, prompting and parsing strategies, and modeling parameters for USE,

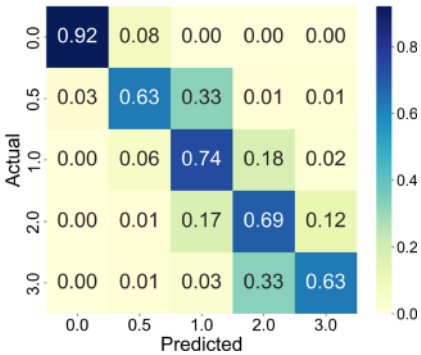
(a) GPT with Structured Guidance (κ = 0.79)



(b) RAG-Enabled GPT (κ = 0.80)



(c) GPT with Domain Count and Confidence Level (κ = 0.83)



(d) Multi-Metric Evaluation of Models Performance

Method	Evaluation Metrics		
	Cohen's Kappa Score	Spearman's Rank Correlation	Baccianella's Adapted MSE
GPT with Structured Guidance	0.79	0.81	0.12
RAG-Enabled GPT	0.80	0.81	0.11
GPT with Domain Count and Confidence Level	0.83	0.83	0.10

(e) Association Between CDR and GPT Confidence Levels in Assigning CDR

Domain	β Coefficient	Standard Error	p-value
Memory	16.21	965.72	0.98
Orientation	1.34	0.29	<0.001
Judgment and Problem Solving	1.30	0.31	<0.001
Community Affairs	1.01	0.33	0.0023
Home and Hobbies	1.50	0.31	<0.001
Personal Care	1.14	0.30	<0.001

Fig. 4 | Performance and Statistical Analysis of GPT-4o-Powered Framework in Assigning Global CDR. Normalized confusion matrices for three GPT-4o-based approaches in cognitive impairment staging: **a** GPT-4o with Structured Guidance, **b** RAG-Enabled GPT-4o, and **c** GPT-4o with Confidence Level and Domain Counts; each matrix shows the proportion of actual vs. predicted CDR scores within each row, with darker colors indicating higher proportions. Rows are normalized to sum to 1. **d** Multi-Metric Evaluation of Model Performance for Assigning Global CDR: Table summarizing model performance across multiple evaluation metrics—Cohen's kappa score, Spearman's Rank Correlation, and Baccianella's adapted MSE. **e** Association Between CDR Domains and GPT-4o Confidence Levels in Assigning Global CDR: Table showing the statistical association between CDR domains (binary variable indicating documentation of domain in the note) and GPT-4o confidence levels (Medium, High) in assigning global CDR. β coefficients indicate the effect size of the association of each domain with standard errors and p-values.

DementiaBERT, and XGBoost. These inclusions are intended to facilitate future replication and extension of our work.

While these results are encouraging for GPT-4o's potential for automated chart reviews and facilitating diagnosis in clinical settings, there are some instances of misclassification that need to be addressed. We identified three key factors contributing to misclassifications through a detailed error analysis: (1) GPT-4o tends to overly rely on specific cognitive assessment scores, such as MoCA and MMSE, directly making decisions based on whether the scores fall within the MCI or dementia range, instead of integrating more comprehensive cognitive and functional evaluation beyond these cognitive tests scores; (2) in some cases, GPT-4o may not account for the absence of neurodegeneration evidence, leading to incorrect classifications; (3) the model sometimes demonstrates a focus on individual factors—overestimating impairment based on a single affected domain, such as decision-making, or underestimating it by categorizing individuals with independence in any activities of daily living as MCI or CU, without considering other cognitive domains. Moreover, the model's lower performance on MCI reflects this category's inherent diagnostic complexity. This also aligns with adjudicator confidence scores, which were significantly lower for MCI than for CU (Normal) or Dementia, suggesting that even expert raters find MCI cases more difficult to classify.

This challenge in classifying MCI underscores the importance of the confidence scores. Although *logprobs* provide objective token-level certainty, our findings suggest that GPT-4o's self-reported confidence more closely aligns with expert assessments. One possible explanation is that *logprobs* reflect token-level linguistic certainty, whereas GPT-4o-assigned

confidence may better capture internal reasoning aligned with human-like diagnostic processes. We propose that a dual-confidence framework—combining the GPT-4o-assigned confidence and *logprobs*-derived certainty—could improve reliability and identify cases for human review.

Given our GPT-4o-powered framework's strong performance in classifying CI stage, we further evaluated its potential in assigning global CDR. While the framework also performed well in assessing global CDR (weighted kappa 0.79–0.83), its scores were slightly lower than in classifying CI stage. This difference likely reflects the increased complexity of the task of assigning global CDR, as the CDR requires detailed assessment of cognitive and functional abilities across six domains, whereas staging provides a broader categorization of cognitive status. Notably, the highest weighted Cohen's kappa score was achieved through the use of prompt engineering techniques that incorporated confidence levels and a count of the documented domains. However, even the best approach had several errors, so co-author JRD, who is a memory specialist, manually reviewed the cases with incorrectly assigned CDRs. The mismatches between the global CDR in the notes and GPT-4o-determined global CDR were attributed to three key factors: (1) gestalt vs. formal CDR calculation – some clinicians estimate a "gestalt" CDR based on their overall assessment rather than using the formal algorithm, leading to potential discrepancies (for example, a patient clinically presenting with mild dementia (CDR 1.0) may have a formally calculated CDR of 0.5); (2) bias towards amnestic presentations – the formal CDR scoring, which heavily weights memory changes, may underestimate functional severity in non-amnestic dementia cases; (3) reliance on history vs. exam data – GPT-4o may rely heavily on history sections and miss key information from the exam, such as orientation, leading to miscalculations.

The use of RAG-enabled GPT-4o did not significantly enhance the scores in this study; it is possible that the source of information being utilized is not truly 'external' for GPT-4o. That is, we queried GPT-4o directly about the National Alzheimer's Coordinating Center (NACC) documents which were used as sources for augmentation, and GPT-4o was able to answer with content from these documents, indicating that GPT-4o had already been trained on similar or identical information. Consequently, the retrieved documents merely reinforced existing knowledge rather than introducing new insights. Specialists' decision-making notes or related sources could be highly useful for augmenting GPT-4o via RAG in our task, as they are likely not part of the common datasets used to train GPT and can provide more nuanced expert reasoning. In future work, we could potentially explore published case reports, clinical decision-making articles, or hospital case conferences where clinicians discuss their reasoning for diagnosis and treatment. In short, while our results did not show significant performance improvements with the RAG approach, we believe it would be premature to conclude that the RAG architecture itself is not useful based on our analyses. With non-pretrained knowledge, RAG might still demonstrate its full potential, which we plan to explore in future work.

Our study has a few limitations. A potential limitation is that in the evaluation dataset, one-third of the patients were seen by specialists and another third had a dementia-related diagnosis code in their medical records[28]. These patients may already be under suspicion for or treatment of cognitive issues and are more likely to have detailed relevant documentation in their EHRs, potentially inflating the performance metrics. Moreover, it is important to acknowledge that although we examined sex as a potential source of bias in model performance and found no significant difference in model performance between male and female subgroups, other socio-demographic factors—such as access to specialists, healthcare utilization, symptom reporting, and documentation practices in clinical notes—may introduce biases that were not evaluated in this study[29,30]. Additionally, the cohort demographics are not representative of the broader U.S. or regional populations—for instance, 93.1% of the dataset consisted of individuals who identified as non-Hispanic White. However, the cohort closely reflects the MGH Medicare population[28], making it appropriate for evaluating model performance in a real-world academic medical setting. Within this context, and despite the noted limitations, the GPT-4o-powered framework consistently outperformed other large language models.

Future work is essential to evaluate and mitigate potential biases in EHR data to ensure responsible and equitable deployment at scale. To deploy and improve reliability in sensitive clinical settings, we plan to collaborate with more specialists to refine the prompt engineering process and explore incorporating additional relevant documents via RAG, such as hospital case conferences. This approach will help reduce misclassifications and ensure more comprehensive assessments in future iterations. Additionally, larger studies at multiple healthcare institutions are required to validate our GPT-4o-powered framework as a tool for dementia chart reviews, and to investigate whether GPT-4o-assisted cognitive diagnoses in clinical settings can influence patient outcomes.

In conclusion, our proposed GPT-4o framework achieves strong performance, particularly in clearly defined categories. However, ambiguity in borderline and MCI cases remains a challenge for real-world use. We propose a clinician-in-the-loop framework where GPT-4o confidence scores are considered alongside clinical judgment[31] to guide decisions about further cognitive evaluation or specialist referral. Rather than fully automating decisions in ambiguous cases, this approach uses confidence scores to flag uncertain or borderline predictions for expert review, ensuring that deployment remains safe, interpretable, and aligned with real-world clinical practice. Moreover, considering the broader potential of AI agents in medicine[27], we conceptualized an interactive AI agent workflow that integrates our framework with human feedback to facilitate automatic data queries from EHR, real-time user engagement and decision support for cognitive diagnosis. While this workflow has not yet been fully implemented, we believe that its design and underlying concept can serve as a valuable foundation and guide deployment in real-world clinical settings.

Methods

Notes preprocessing

To ensure scalability and relevance, our pipeline restricted the input to notes from the most recent three years containing cognition-related keywords, following a strategy used in our prior work[32]. For CI stage classification, we employed two methods in the text representation extraction step to capture relevant clinical information from unstructured notes: keyword-based sentence extraction and GPT-4o summarization, each adopted in separate classification frameworks to evaluate its effectiveness (see Fig. 1). In the first approach, we performed sentence segmentation on clinical notes and extracted all sentences containing specific keywords. We used two comprehensive collections of keywords: (I) dementia-related words (Dem) such as "memory", "cognition", "MMSE", "Donepezil"[32], and (II) activities of daily living related words (ADL) such as "cook", "clean", "drive", "walk", "bath"[32]. In the second approach, we first aggregated visit notes in chronological order for each patient. The context of the note (i.e., the date, department, specialty) was added to each note and summarized by GPT-4o (Supplementary Note 1a). The note summaries were then combined chronologically into one document.

For automatically assigning global CDR, we first redacted any sentences mentioning CDR from the clinical notes using regular expressions. This was done before the notes were evaluated using GPT-4o to ensure that the answer would not be contained within the provided clinical notes.

GPT-4o framework and prompt engineering

We accessed GPT-4o via Azure OpenAI Services (API version 2024-05-13) through a private deployment within Mass General Brigham's firewall, ensuring compliance with institutional privacy and protected health information (PHI) protection policies. The model was run with temperature = 0 and max_tokens = 4096 to ensure deterministic and complete outputs.

For CI stage classification, we developed a GPT-4o-powered framework to perform ordinal classification of CI stages (CU, MCI, and Dementia). GPT-4o generated summaries of chronological notes and was further prompted to generate a "summary of summaries" to make a final diagnosis (Supplementary Note 1b). Additionally, GPT-4o was prompted to provide a rationale for its classification.

GPT-4o was also prompted to provide a confidence level from 1 to 100, which is the default range for generative language models when producing probabilistic outputs. To enable comparison with clinician confidence ratings (1–4), we applied quantile mapping to align the distributions. We also tested an alternative measure of model certainty using log-probabilities (*logprobs*) of the predicted tokens. These were transformed into probabilities and then mapped to the 1–4 clinician scale using quantile mapping.

Our end-to-end GPT-4o-powered approach, combined GPT-4o-generated summaries with direct GPT-4o classification. We used a two-stage prompting approach: (1) summarizing note chunks, and (2) generating a diagnosis and confidence score based on the summary of summaries. Responses were parsed using regex-based scripts. Approximately 1% of outputs required manual correction due to formatting deviations.

For assigning global CDR, we tried three approaches to enhance the GPT-4o-powered framework. The first approach utilized a structured answer template to guide GPT-4o's analysis of patient visit notes. The response format asked GPT-4o to provide observations and summaries across six key domains in the CDR scoring system[33]: (i) Memory, (ii) Orientation, (iii) Judgment and Problem Solving, (iv) Community Affairs, (v) Home and Hobbies, and (vi) Personal Care. The model was required to conclude with an explicit CDR score. This structured format helped standardize the output and ensured that each relevant domain was systematically considered before making the final decision. Second, to further enhance GPT-4o's ability to determine stages of CI, we implemented a Retrieval-Augmented Generation (RAG) approach[34]. We extracted information from the NACC UDS v3 CDR Dementia Staging Instrument[33,35], such as, "CDR of 0.5: slight impairment in solving business/financial affairs and judgment" and indexed it for efficient retrieval. When processing patient visit notes, the model searched for the top three chunks that are most similar to the notes using cosine similarity as a metric. These relevant pieces of information were then used to augment GPT-4o, providing the model with domain-specific guidance from human experts. Third, we asked GPT-4o to include a self-assessment of confidence and a count of explicitly mentioned domains. The model was asked to review the visit notes with a focus on identifying information within the six specific CDR domains and summarize the number of domains with explicit information. We tested multiple prompt variations to improve response structure, elicit domain-specific scores, and reduce overestimation, ultimately arriving at the final prompts. Examples of these prompts are provided in Supplementary Note 2a–c. Based on the clarity and consistency of the evidence across these domains, GPT-4o was asked to assign a confidence level (low, medium, or high). This method aimed to enhance the reliability of the predictions by integrating a self-evaluation component into the model's decision-making process.

Other frameworks

To compare our GPT-4o-powered framework to traditional embeddings and language models, we constructed three additional frameworks —USE, DementiaBERT, and a Hybrid framework—each with a distinct approach to feature extraction and classification. The USE framework used keyword-based sentence extraction with the Universal Sentence Encoder (USE)[36] to generate embeddings, followed by classification with XGBoost. The DementiaBERT framework used a similar keyword-based extraction but employed DementiaBERT[24] for embeddings, which were also classified with XGBoost. DementiaBERT is a domain-specific language model that we developed by fine-tuning ClinicalBERT on labeled dementia-relevant text spans from clinical notes[24]. We extracted 512-token contexts surrounding cognitive terms, labeled them into three categories (positive, negative, neither/no information), and fine-tuned the model using Huggingface + SimpleTransformers. Hyperparameter tuning was performed using Optuna. Recursive Feature Elimination (RFE)[37] was applied to the embeddings in both USE and DementiaBERT frameworks to remove weaker features. The Hybrid framework partially utilized GPT-4o for generating summaries in place of sentence extraction, with DementiaBERT and XGBoost for embedding and classification. We trained the XGBoost models for the three frameworks using default hyperparameters with the mlogloss objective function. No tuning was performed, as our goal was comparative evaluation across feature sets rather than model optimization.

For assessment and comparison of frameworks' robustness and generalizability, we adopted a half-nested 10-fold cross-validation strategy. An inner loop was used to optimize the RFE hyperparameters on a train-validation split within each fold, and the final model, trained on the combined training and validation data using the optimal hyperparameters, was then evaluated on the held-out test fold of the outer loop.

Evaluation

To assess the accuracy and reliability of decisions made by the GPT-4o-powered framework, we used several analytical methods. First, we created confusion matrices to give a detailed view of model performance, displaying the distribution of actual versus predicted class assignments and enabling us to identify patterns in misclassification. Given that our two tasks, staging CI and assigning CDR, are ordinal classification problems, we adopted three evaluation metrics to comprehensively assess model performance: quadratic weighted Cohen's kappa[38], Spearman's Rank Correlation[39], and Baccianella's adapted Mean Squared Error (MSE)[40]. We selected Baccianella's MSE over traditional MSE due to its suitability for ordinal tasks, as it applies greater penalties for larger discrepancies between predicted and actual stages.

Statistical analysis

For CI stage classification, a paired t-test was conducted for each model's kappa scores across the 10-fold cross-validation splits. To examine potential biases and performance differences between male and female patients, we conducted a sex-based subgroup analysis. For each model, predictions were stratified by sex across 10 cross-validation folds, and weighted Cohen's kappa scores were calculated separately for male and female groups within each fold. A paired t-test was then performed across the kappa scores from the 10 folds to assess whether there was a statistically significant difference in model performance between the two subgroups.

For assigning global CDR, understanding the relationship between GPT-4o's confidence levels and the six domains is crucial for providing insights into how specific aspects of cognitive assessment influence the model's confidence and model reliability in clinical decision-making. Therefore, we built a logistic regression model using the six domains as binary regressors (independent variables). Each domain was coded as 1 if it was recognized from notes by GPT-4o, and 0 otherwise. The target (dependent) variable for the model was GPT-4o's confidence level for the task, categorized as Medium or High; cases with Low confidence were excluded from the analysis, as only two cases fell into this category.

Data availability

The data analyzed in this study are sourced from the Mass General Brigham Healthcare System. Due to the presence of protected health information (PHI) and restrictions imposed by institutional privacy policies, these data are not publicly available.

Code availability

The underlying code developed for the tasks and evaluation in this study is available in the GitHub repository: https://github.com/mindds/GPT-CI-Staging.

Received: 22 February 2025; Accepted: 23 June 2025;
Published online: 03 July 2025

References

1. Association, A. S. 2024 Alzheimer's disease facts and figures. *Alzheimer's Dement.* **20**, 3708–3821, https://doi.org/10.1002/alz.13809 (2024).
2. Robinson, L., Tang, E. & Taylor, J.-P. Dementia: timely diagnosis and early intervention. *BMJ* **350**, h3029, https://doi.org/10.1136/bmj.h3029 (2015).

3. Borson, S. et al. Improving dementia care: the role of screening and detection of cognitive impairment. *Alzheimer's Dement.* **9**, 151–159, https://doi.org/10.1016/j.jalz.2012.08.008 (2013).
4. Amjad, H. et al. Underdiagnosis of dementia: an observational study of patterns in diagnosis and awareness in US older adults. *J. Gen. Intern. Med.* **33**, 1131–1138, https://doi.org/10.1007/s11606-018-4377-y (2018).
5. Taylor, D. H., Østbye, T., Langa, K. M., Weir, D. & Plassman, B. L. The accuracy of medicare claims as an epidemiological tool: the case of dementia revisited. *J. Alzheimer's Dis.***17**, 807–815, https://doi.org/10.3233/JAD-2009-1099 (2009).
6. Bradford, A., Kunik, M. E., Schulz, P., Williams, S. P. & Singh, H. Missed and delayed diagnosis of dementia in primary care: prevalence and contributing factors. *Alzheimer Dis. Associated Disord.* **23**, 306–314, https://doi.org/10.1097/WAD.0b013e3181a6bebc (2009).
7. Ghassemi, M., Naumann, T., Schulam, P., Chen, I. Y. & Ranganath, R. A Review of Challenges and Opportunities in Machine Learning for Health.*AMIA Jt Summits Transl. Sci. Proc* **2020**, 191–200 (2020).
8. Association, A. S. 2019 Alzheimer's disease facts and figures. *Alzheimer's Dement.* **15**, 321–387, https://doi.org/10.1016/j.jalz.2019.01.010 (2019).
9. Yarnall, K. S., Pollak, K. I., Ostbye, T., Krause, K. M. & Michener, J. L. Primary care: is there enough time for prevention?. *Am. J. Public Health* **93**, 635–641, https://doi.org/10.2105/ajph.93.4.635 (2003).
10. Boustani, M. et al. Who refuses the diagnostic assessment for dementia in primary care?. *Int J. Geriatr. Psychiatry* **21**, 556–563, https://doi.org/10.1002/gps.1524 (2006).
11. Fowler, N. R. et al. Traits of patients who screen positive for dementia and refuse diagnostic assessment. *Alzheimers Dement.* **1**, 236–241, https://doi.org/10.1016/j.dadm.2015.01.002 (2015).
12. Boustani, M. et al. Implementing a screening and diagnosis program for dementia in primary care. *J. Gen. Intern. Med.* **20**, 572–577, https://doi.org/10.1111/j.1525-1497.2005.0126.x (2005).
13. Bamford, C. et al. Disclosing a diagnosis of dementia: a systematic review. *Int. J. Geriatr. Psychiatry* **19**, 151–169, https://doi.org/10.1002/gps.1050 (2004).
14. Gilmore-Bykovskyi, A. L. et al. Unstructured clinical documentation reflecting cognitive and behavioral dysfunction: toward an EHR-based phenotype for cognitive impairment. *J. Am. Med. Inform. Assoc.***25**, 1206–1212, https://doi.org/10.1093/jamia/ocy070 (2018).
15. Amra, S. et al. Derivation and validation of the automated search algorithms to identify cognitive impairment and dementia in electronic health records. *J. Crit. Care* **37**, 202–205, https://doi.org/10.1016/j.jcrc.2016.09.026 (2017).
16. Reuben, D. B., Hackbarth, A. S., Wenger, N. S., Tan, Z. S. & Jennings, L. A. An automated approach to identifying patients with dementia using electronic medical records. *J. Am. Geriatr. Soc.* **65**, 658–659, https://doi.org/10.1111/jgs.14744 (2017).
17. Ben Miled, Z. et al. Predicting dementia with routine care EMR data. *Artif. Intell. Med.* **102**, 101771. https://doi.org/10.1016/j.artmed.2019.101771 (2020).
18. Hane, C. A., Nori, V. S., Crown, W. H., Sanghavi, D. M. & Bleicher, P. Predicting onset of dementia using clinical notes and machine learning: case-control study. *JMIR Med. Inform.* **8**, e17819, https://doi.org/10.2196/17819 (2020).
19. Jiang, L. Y. et al. Health system-scale language models are all-purpose prediction engines. *Nature* **619**, 357–362, https://doi.org/10.1038/s41586-023-06160-y (2023).
20. Wang, L. 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* **4**, e2135174, https://doi.org/10.1001/jamanetworkopen.2021.35174 (2021).
21. de Arriba-Pérez, F., García-Méndez, S., Otero-Mosquera, J. & González-Castaño, F. J. Explainable cognitive decline detection in free dialogues with a Machine Learning approach based on pre-trained Large Language Models. *Appl. Intell.* https://doi.org/10.1007/s10489-024-05808-0 (2024).
22. Yan, C. et al. Large language models facilitate the generation of electronic health record phenotyping algorithms. *J. Am. Med. Inf. Assoc.* **31**, 1994–2001, https://doi.org/10.1093/jamia/ocae072 (2024).
23. Hong, Z. et al. Natural language processing to detect cognitive concerns in electronic health records using deep learning. *arXiv:2011.06489 [cs.CL]* (2020).
24. Tyagi, T. et al. Using deep learning to identify patients with cognitive impairment in electronic health records. *arXiv:2111.09115 [cs.CL]* (2021).
25. Du, X. et al. Enhancing early detection of cognitive decline in the elderly: a comparative study utilizing large language models in clinical notes. *medRxiv* https://doi.org/10.1101/2024.04.03.24305298 (2024).
26. Wang, Z. et al. Can LLMs like GPT-4 outperform traditional AI tools in dementia diagnosis? Maybe, but not today. *arXiv:2306.01499 [cs.CL]* (2023).
27. Zou, J. & Topol, E. J. The rise of agentic AI teammates in medicine. *Lancet* **405**, 457, https://doi.org/10.1016/S0140-6736(25)00202-8 (2025).
28. Moura, L. M. V. R. et al. Identifying medicare beneficiaries with dementia. *J. Am. Geriatr. Soc.* jgs.17183 https://doi.org/10.1111/jgs.17183 (2021).
29. Sun, M., Oliwa, T., Peek, M. E. & Tung, E. L. Negative patient descriptors: documenting racial bias in the electronic health record: study examines racial bias in the patient descriptors used in the electronic health record. *Health Aff.* **41**, 203–211, https://doi.org/10.1377/hlthaff.2021.01423 (2022).
30. Gianfrancesco, M. A., Tamang, S., Yazdany, J. & Schmajuk, G. Potential Biases in machine learning algorithms using electronic health record data. *JAMA Intern. Med.* **178**, 1544, https://doi.org/10.1001/jamainternmed.2018.3763 (2018).
31. Goh, E. et al. Large language model influence on diagnostic reasoning: a randomized clinical trial. *JAMA Netw. Open* **7**, e2440969, https://doi.org/10.1001/jamanetworkopen.2024.40969 (2024).
32. Noori, A. et al. Development and evaluation of a natural language processing annotation tool to facilitate phenotyping of cognitive status in electronic health records: diagnostic study. *J. Med. Internet Res.* **24**, e40384, https://doi.org/10.2196/40384 (2022).
33. Hughes, C. P., Berg, L., Danziger, W., Coben, L. A. & Martin, R. L. A new clinical scale for the staging of dementia. *Br. J. Psychiatry* **140**, 566–572, https://doi.org/10.1192/bjp.140.6.566 (1982).
34. Lewis, P. et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Proc. 34th International Conference on Neural Information Processing Systems* Article 793 (Curran Associates Inc., 2020).
35. Besser, L. et al. Version 3 of the National Alzheimer's Coordinating Center's Uniform Data Set. *Alzheimer Dis. Assoc. Disord.* **32**, 351–358, https://doi.org/10.1097/WAD.0000000000000279 (2018).
36. Cer, D. et al. Universal Sentence Encoder. *arXiv:1803.11175 [cs.CL]* (2018).
37. Guyon, I., Weston, J., Barnhill, S. & Vapnik, V. Gene selection for cancer classification using support vector machines. *Mach. Learn.* **46**, 389–422, https://doi.org/10.1023/A:1012487302797 (2002).
38. Cohen, J. Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. *Psychol. Bull.* **70**, 213–220 (1968).
39. Spearman, C. The proof and measurement of association between two things. *Am. J. Psychol.* **15**, 72–101 (1902).
40. Baccianella, S., Esuli, A. & Sebastiani, F. *2009 Ninth International Conference on Intelligent Systems Design and Applications*. 283-287 (IEEE).

Acknowledgements
We acknowledge funding from the following sources NIA P30AG062421 and R01AG082698. We would like to thank MassGeneral Brigham research computing and information technology for the study. We are grateful to Beata-Gabriela Simpson for proofreading the manuscript.

Author contributions
Y.L. and S.D. conceived and designed the study. Y.L. and Y.H. developed the models and performed all the analyses. S.A. and I.P. conceived the USE approach. Y.H. performed data queries and clinical note preprocessing. C.M. provided methodological guidance. S.S.M., L.M.V.R.M., B.W., A.V., C.S.R., and D.B. offered clinical insights, and J.R.D. reviewed errors and offered clinical insights. S.D. supervised the project, oversaw data extraction. Y.L. and S.D. wrote the initial draft; all authors contributed to the final manuscript and approved the final version.

Competing interests
Dr. Westover is a co-founder, scientific advisor, consultant to, and has personal equity interest in Beacon Biosignals.

Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41746-025-01834-5.

Correspondence and requests for materials should be addressed to Sudeshna Das.