An Enhanced Language Model for Predicting Amyloid Beta PET, Tau PET, and Cognitive Decline in Alzheimer's Disease

Authors: Rita Qiuran Lyu[1#], Marcell Borhi[1#], Yanping Li[1], Nanshan Jia[2], Kevin Shen[3], , Susan Landau[5], William J Jagust[5], Pedro Pinheiro-Chagas[6], Katherine Possin[6], Mark A. Espeland[7] , Joseph Giorgio[5,8]*, Jingshen Wang[1]*, for the Alzheimer's Disease Neuroimaging Initiative†, The Health and Aging Brain Study (HABS-HD) Study Team[+], and the U.S. POINTER Study Group

[#] Contributed equally as first authors
* Contributed equally as senior authors

Affiliations:
[1] Division of Biostatistics, School of Public Health, University of California, Berkeley, California, USA
[2] School of Engineering, University of California, Berkeley, California, USA
[3] The Nueva School, San Mateo, California, USA
[4] Department of Neuroscience, University of California, Berkeley, California, USA, 94720
[5] Department of Neurology, University of California, San Francisco, Memory and Aging Center
[6] Department of Internal Medicine, Wake Forest University School of Medicine, Winston-Salem, North Carolin, USA, 27157
[7] School of Psychological Sciences, College of Engineering, Science and the Environment, University of Newcastle, Newcastle, New South Wales, Australia, 2308

†Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found in Appendix: Collaborators. [+] HABS-HD MPIs: Sid E O'Bryant, Kristine Yaffe, Arthur Toga, Robert Rissman, & Leigh Johnson; and the HABS-HD Investigators: Meredith Braskie, Kevin King, James R Hall, Melissa Petersen, Raymond Palmer, Robert Barber, Yonggang Shi, Fan Zhang, Rajesh Nandy, Roderick McColl, David Mason, Bradley Christian, Nicole Phillips, Stephanie Large, Joe Lee, Badri Vardarajan, Monica Rivera Mindt, Amrita Cheema, Lisa Barnes, Mark Mapstone, Annie Cohen, Amy Kind, Ozioma Okonkwo, Raul Vintimilla, Zhengyang Zhou, Michael Donohue, Rema Raman, Matthew Borzage, Michelle Mielke, Beau Ances, Ganesh Babulal, Jorge Llibre-Guerra, Carl Hill and Rocky Vig.

Abstract

Summary
Current machine learning (ML) models for Alzheimer's disease (AD) prediction are often constrained by binary outcome targets, limited integration of multimodal data, and poor generalizability across diverse populations. These issues are compounded by the often black-box nature of ML predictions. In contrast, large language models (LLMs) offer a promising alternative due to their ability to process heterogeneous, longitudinal, and unstructured textual data providing textual justification for supporting clinician's decision-making. In this study, we developed an enhanced LLM fine-tuned on multimodal datasets to predict key AD pathology markers and cognitive decline trajectories.

Methods
We trained ADLLM, an enhanced LLM based on LLaMA 3·1, an open-source pre-trained LLM released by META, fine-tuned using Low-Rank Adaptation (LoRA) on serialized cross-sectional and longitudinal data from three diverse aging and dementia cohorts: ADNI, HABS-HD, and U.S. POINTER, using a fourth external cohort (A4) for validation. Multimodal predictors were used including demographics, neuroimaging and fluid biomarkers, clinical assessments, and additional health-related covariates. Outcomes were PET-derived burden of amyloid-beta (Aβ) and AD tau in all cohorts, and rate of cognitive decline in ADNI and HABS-HD. Inputs were converted to a text prompt, allowing the model to capture relationships across modalities and predict composite outcomes. Performance was evaluated using multi-class AUC, accuracy, and specificity, based on the predicted probabilities generated by ADLLM, with the predicted category assigned to the one with the highest probability. We used multiple methods for LLM interpretation quantifying the importance of variable groups across cohorts.

Findings
ADLLM consistently outperformed both zero-shot LLMs and gradient boosting (GB) models across internal and external validation datasets. For Aβ burden prediction, ADLLM achieved four-class AUCs of 0·825 (ADNI), 0·773 (HABS-HD), 0·702 (U.S. POINTER), and 0·610 (A4). For meta-tau burden, ADLLM exceeded GB in AUC across all datasets and demonstrated superior accuracy and specificity. In predicting cognitive decline, ADLLM achieved AUCs of 0·806 (ADNI) and 0·816 (HABS-HD), significantly outperforming GB and zero-shot baselines. ADLLM's attention scores aligned with variable importance rankings from GB models, highlighting shared influential features, covering genetic, MRI structural variation, and diagnostic variables, with the incorporation of fluid biomarkers further improving predictive performance.

Interpretation
ADLLM represents a promising approach for AI-guided AD prediction, demonstrating strong generalizability and competitive performance compared with traditional ML models. ADLLM can predict composite outcomes by processing multimodal inputs, capturing longitudinal patterns, and providing interpretable attention-based variable importance. The model consistently performed well across cohorts and outcome tasks. Moreover, model-generated explanations and post-hoc analyses provide preliminary insights into model behavior, emphasizing its potential as a scalable and clinically relevant tool for AD diagnosis and progression monitoring.

INTRODUCTION

Alzheimer's disease (AD) is a progressive neurodegenerative disorder that affects millions globally, placing a significant burden on individuals and healthcare systems[1–5]. The canonical AD pathological cascade suggests that amyloid-beta (Aβ) accumulation initiates the disease process, accelerating tau deposition ultimately impacting cognitive function[6]. However, this progression varies significantly across patients, leading to heterogeneity in disease patterns. Early and accurate predictions of clinical progression and AD that account for individual-level heterogeneity could improve diagnosis and precision treatment. [7–10]

Significant efforts have been dedicated to developing supervised machine learning (ML) models for AD prediction tasks, such as Aβ positivity,[11,12] tau PET burden,[13] and MCI-to-dementia conversion.[14–16] Nonetheless, most published models focus on a single binary endpoint, rely on baseline snapshots rather than longitudinal trajectories, lack multimodal integration, and, crucially, seldom undergo rigorous external validation. These limitations hinder generalizability and, consequently, reduce clinical utility.

Large language models (LLMs) provide a unifying framework that can overcome these barriers. By casting heterogeneous inputs into a unified natural language representation, LLMs can capture cross-modal dependencies between structured variables (genetics, demographics, laboratory values), unstructured clinical notes, and free-text neuroimaging reports that conventional pipelines struggle to exploit,[17–20]. Recent studies using zero-shot prompting have explored foundational LLMs for several AD-related tasks, including predicting AD onset, estimating AD risk, and answering AD-related clinical questions based on speech recordings and transcripts, [21–24] electronic health records (EHRs),[25–28] tabular data,[29] and AD literature knowledge graph.[30] However, without fine-tuning, the closed-source nature of the models coupled with opaque dynamics of only relying on prompting hinders interpretability, limits accessibility, and complicates on-premise deployment in protected-health-information environments.

Domain-specific open-weight LLMs alleviate these concerns and confer two additional advantages. *First*, their parameters are inspectable, enabling the use of attention heatmaps[31] and other modern interpretability techniques to reveal which features contribute to the model's predictions. Such transparency is important for regulatory approval, clinician trust, and hypothesis generation. *Second*, open-weight models can be distilled or LoRA-adapted to run on secure, locally-maintained machines, democratizing access for smaller research groups, low-resource clinics, and privacy-sensitive settings where data must remain on-premise.[31–33]

Here we introduce ADLLM, the first enhanced (or fine-tuned) LLM specifically designed to predict AD progression and forecast cognitive decline using multimodal data. ADLLM was built using Low-Rank Adaptation (LoRA)[34] on pre-trained LLaMA 3·1 released by META[19] to efficiently adapt the pre-trained LLM for domain-specific AD predictions while preserving its original pre-training knowledge base, broad reasoning capabilities, and contextual understanding. To fine-tune and validate the model, we used large-scale publicly available datasets from the

Alzheimer's Disease Neuroimaging Initiative*[1] (ADNI), Health and Aging Brain Study: Health Disparities (HABS-HD), U.S. study to Protect Brain Health through Lifestyle Intervention to Reduce Risk (U.S. POINTER), and the initial cross-sectional screening data from the Anti-Amyloid Treatment in Asymptomatic Alzheimer's (A4). We curated a diverse set of subject-level predictors including, genetics, demographics, structural MRI measures, biofluids, and other health measures the ADLLM is trained to predict Aβ and tau PET burden, and cognitive decline measured by MMSE score change rate (Figure 1).

To our knowledge, this is the first study to develop an enhanced LLM specifically designed to support both the prediction of AD biomarkers and forecasting of cognitive decline using multimodal data. Our results demonstrate that ADLLM outperforms zero-shot LLM and traditional ML models in multi-class AUC, accuracy, and specificity across internal hold-out validation datasets and the external validation cohort.

METHODS

Methods for developing the enhanced LLM

Data sources

The model was trained and validated on four multimodal datasets with imaging: ADNI (2601 participants), a longitudinal, multi-center study investigating late onset AD; HABS-HD (3229 participants), a longitudinal study of AD among racially and ethnically diverse communities; and baseline data from the U.S. POINTER imaging study (POINTER, 937 participants[35]), a clinical trial to compare the relative impact of two multidomain lifestyle interventions that differ in intensity and accountability on cognitive function in older adults at increased risk for cognitive decline.[36] The baseline data from the A4 clinical trial served as the external validation dataset for the enhanced model. As tau and Aβ PET were required for validation, the A4 sample primarily consisted of participants with elevated Aβ, with a much smaller proportion from the non-elevated Aβ group. Pooling across the four cohorts, there was a range of clinical diagnoses with the majority of subjects clinically unimpaired (Table 1).

Participants in these cohorts have extensive information, such as demographics, neuropsychological assessment, fluid biomarkers (i.e., blood or cerebrospinal fluid), and other health data (e.g. comorbidities). The datasets used for model formulation are well suited for developing ADLLM as they are collected from diverse populations, each contributing unique characteristics. In particular, ADNI consists of an older, predominantly white, and highly educated population with fewer comorbidities, and extensive clinical and PET imaging follow-up. The

---

1*Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at:
http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

HABS-HD study involves a relatively younger population that is mainly cognitively normal, with some comorbidities, racially diverse participants, and some clinical and PET imaging follow-up. The POINTER study includes fewer clinically impaired individuals, many with cardiovascular comorbidities; our analyses are restricted to its baseline data. These datasets provide a wide spectrum of demographic, health, and imaging variations essential for LLM model development. By pooling data from these diverse datasets, we mitigated the sampling biases inherent in a single cohort, leading to an enhanced ADLLM that produces credible outcome predictions across diverse populations and disease contexts. Finally, we use the A4 study as an out-of-sample validation set. The A4 sample we include has tau PET imaging, is cognitively normal, predominantly white, and highly educated, and consists mainly of Aβ positive individuals.

A4 was chosen as the external validation cohort because it is fully independent of model development and includes both amyloid and tau PET, enabling external evaluation of biomarker prediction in a clinically distinct, preclinical AD-enriched sample. Although A4 is narrower demographically and clinically than an ideal real-world validation cohort, it provides a stringent test of transportability to a cognitively normal but amyloid-enriched population.

To further clarify cohort composition and outcome distributions, we additionally examined outcome class balance at the participant level for each prediction task. There was substantial class imbalance in several cohorts. In the HABS-HD cohort the majority of participants fell into the negative Aβ category, whereas the A4 cohort contained a much higher proportion of Aβ-positive individuals by study design. Similarly, the distribution of meta-temporal tau burden and MMSE change rate varied across cohorts, reflecting differences in recruitment criteria and disease stage representation. Table 1 summarizes key covariates and outcomes in each cohort.

Predictors

The predictors (Figure 1) were categorized into thirteen groups: genetic information (APOE 2 and APOE 4 status), family history of AD, demographics, vital signs, cognitive performance, physical performance measures, MRI-derived cortical thickness and volumetric measures, cerebrospinal fluid (CSF) biomarker, blood biomarkers, additional disease burden (e.g., coexisting chronic diseases such as cardiovascular disease, cancer, depression, etc.), medication use, and cognitive diagnosis (Supplementary Table 1). Several predictors were longitudinally measured across visits, including cognitive performance, physical performance, MRI-derived measures, and biomarker data.

Composite outcomes including *Aβ PET, tau PET, and rate of cognitive decline*

PET Imaging: Aβ PET

Each study collected Aβ PET to determine cortical Aβ burden. Within each cohort, Aβ PET imaging was transformed into the centiloid scale (CL) using cohort-specific pipelines (Supplementary Methods: Amyloid-beta (Aβ) PET processing). Aβ CL was discretized into 4 bins (Negative Aβ CL<20; Weakly positive 20≤ Aβ CL<40; Positive 40 ≤ Aβ CL≤ 60; Strongly positive 60< Aβ CL). The Centiloid intervals were chosen as pragmatic ordinal categories intended to distinguish amyloid-negative, early elevated, intermediate elevated, and strongly elevated amyloid burden. The 20-centiloid cutoff is broadly consistent with prior work [37–39] using Centiloids to identify amyloid positivity, whereas the higher intervals were chosen to retain severity gradation beyond a binary classification framework. We therefore view these bins as clinically motivated categorization thresholds rather than universally established disease-stage boundaries.

PET Imaging: tau PET

Each study collected tau PET imaging which was used to define AD tau burden extracted from the meta-temporal region of interest. Within each cohort, tau PET imaging was transformed into a Standardized Uptake Value Ratio (SUVR) image using inferior cerebellar grey matter as a reference region (Supplementary Methods: Tau PET processing). Tau SUVR in meta-temporal ROIs was later normalized by the mean and standard deviation of cognitively normal Aβ negative participants in each cohort (Supplementary Methods: Tau SUVR normalization) and discretized into four categories using cohort-specific thresholds derived from the distribution of normalized SUVR values (Very low normalized SUVR ≤-0·413; Low -0·413<normalized SUVR ≤0·194; High 0·194< normalized SUVR ≤ 1·034; Very high 1·034< normalized SUVR).

Rate of cognitive decline: MMSE change rate

In studies with longitudinal cognitive assessment (ADNI and HABS-HD), we derived the rate of cognitive change for each individual as the least squares fit of MMSE and time from baseline in years (Supplementary Methods: MMSE change rate). We discretized the change rate (the fitted coefficients of the least squares) into four categories based on quartile thresholds: Stable 0<rate (72%-100%); Slow decline -0·148<rate≤0 (50%-72%); Moderate decline -0·767<rate≤-0·148 (25%-50%); Fast decline rate≤-0·767 (0%-25%).

Fixed formatting in prompt instruction

We converted predictors and outcomes into natural language by textual serialization (Supplementary Table 2) and used fixed formatting and task-specific instructions to facilitate ADLLM downstream validation, improve prediction consistency, and avoid hallucinations (Figure 1). Specifically, we used prompt instruction, "You are a specialist in diagnosing and forecasting the disease progression for Alzheimer's disease participants. Your task is to predict the following for a patient based on their medical records: the current amyloid beta (abeta) burden in the brain PET scan, the current tau burden of the temporal meta regions of interest (ROIs) in the brain PET scan, and Mini-Mental State Examination (MMSE) score change rate. Please provide the prediction in this structured format: Abeta: Unknown or This patient abeta burden is [*Negative/Weakly positive/Positive/Clearly positive*]. Meta tau: Unknown or This patient's Meta tau burden is [*Very low/Low/High/Very high*]. MMSE rate: Unknown or The MMSE is [*Slow decline/Moderate decline/Fast decline/Stable*]." By constraining the model's output to a predefined set of categorical options, we transformed the language generation task to a multiple-choice classification problem. This narrowed the token generation space, allowing us to extract token probabilities and interpret them as the model's confidence in each category. The token with the highest probability was treated as the predicted category. The structured output enabled the use of standard multi-class classification metrics such as accuracy, AUC, and specificity. To ensure consistency and reproducibility, which is critical in clinical contexts, we employed deterministic decoding, setting the temperature to 1.0 and disabling sampling. This ensured the model selected the most probable token at each step. In contrast, less structured outputs that allow free-form generation require alternative evaluation metrics such as perplexity.[40] However, perplexity is sensitive to domain-specific vocabulary and tokenization, and often fails to reflect the clinical relevance or quality of generated content. Thus, it is less suitable for healthcare applications. Details on token probability extraction and decoding are in Supplementary Methods: Evaluation Metric Calculation Methods.

Enhance pre-trained LLM with fine-tuning

We used Low-Rank Adaptation (LoRA) to fine-tune a foundational LLM (LLaMA 3·1 with 8 billion parameters pre-trained on a mix of publicly available online data) for accurate prediction of

biomarker profiles and cognitive changes in AD. The model was fine-tuned using a training set composed of records from ADNI (n=1878), HABS-HD (n=1872), and POINTER (n=749), with the fixed prompt instruction as described above applied to LLaMA 3·1. Detailed implementation of LoRA is provided in the Supplementary Material (Supplementary Methods: Model enhancement pipeline).

Methods for evaluating, validating, and interpreting the enhanced LLMs

External and internal validation

In each of the serialized datasets from the ADNI, HABS-HD, and POINTER cohorts, 80% of participants were randomly selected as the training dataset for LLM enhancement, while the remaining 20% were used as the internal hold-out validation dataset. The A4 cohort with both Aβ and tau PET served as the external validation set to assess the model's generalization capability. The A4 sample includes predominantly Aβ positive individuals who are cognitively normal, representing a sample that is distinct from the training cohorts, with a much higher proportion of asymptomatic AD participants.

Classical evaluation metrics

Benefiting from the fixed formatting prompt instruction, we extracted the generation probability of each candidate class token and used these probabilities as prediction scores. Model performance was evaluated using multi-class weighted AUC, accuracy, specificity, and sensitivity[41]. To better characterize performance under class imbalance, we additionally computed precision-recall curves and area under the precision-recall curve (PR-AUC). The multi-class weighted AUC captures a model's ability to distinguish between classes, adjusting for class sample sizes to better reflect performance in imbalanced datasets. To quantify its uncertainty, we added '[]' in the reported results to showcase the bootstrapped[42] (with 2000 bootstrap samples) weighted AUC with 95% confidence interval. Accuracy measures the overall proportion of correct classifications across four discrete categories (i.e., chance predictions are 0·25) for all three interested outcomes. Specificity measures the proportion of true negatives that are correctly identified.

Clinical utility evaluation metric

We investigated whether incorporating the cerebrospinal fluid (CSF) or blood biomarkers related to Aβ burden and tau burden (e.g. Aβ 40, Aβ 42, Phosphorylated tau 181 protein, total tau protein) could improve ADLLM performance. As records in A4 and POINTER do not contain relevant CSF/blood biomarkers information, we restricted the analysis to the subgroup of ADNI that contains CSF biomarkers and HABS-HD, which contains blood biomarkers. We compared the differences in ADLLM prediction results with and without the inclusion of CSF or blood biomarkers. Additional results with only AD-related CSF or blood biomarkers are provided in Supplementary Results: Additional results.

Correlation: converting discrete labels to continuous values

Since Aβ CL, normalized meta tau SUVR, and MMSE score change rates were discretized into four bins, we investigated whether the token generation probabilities from ADLLM could meaningfully recover these metrics in their original continuous form using the model-generated probability for each ordinal category to transform the discrete prediction into a continuous variable (Supplementary Methods: Converting discrete labels to continuous values and regression analysis). To evaluate the model's performance, we regressed the recovered values on the original continuous values to derive Mean Square Error (MSE), and computed the correlation (Pearson's) between the recovered values and the original continuous values.

ADLLM interpretation
In LLMs, attention scores from self-attention layers can indicate which components of the input data most strongly influence the model's predictions.[43] Visualizing attention scores could thus potentially offer insights into ADLLM's decision-making. To facilitate this, we grouped predictors into thirteen categories and extracted attention scores for each group in output generation (Supplementary Table 1). Details on calculating group attention scores are provided in the Supplementary Material (Supplementary Methods: Attention scores visualization). Furthermore, we directly prompted ADLLM to justify predictions, using the prompt "Provide your reasons for the prediction." to ask ADLLM to give specific reasons for its prediction.

Model comparison: zero-shot pre-trained LLM and Machine learning (ML)
As a baseline to demonstrate ADLLM's capabilities, we employed zero-shot predictions by instructing LLaMA 3·1 to directly predict outcomes based solely on natural language descriptions of predictors and outcomes without any fine-tuning. Furthermore, we used CatBoost[44,45] from Python package 'catboost' as the baseline ML model for comparison with ADLLM.

RESULTS

Aβ burden
Using ADLLM and GB trained on 161 features taken from 13 non-PET imaging categories (Supplementary Table 2) we performed a four-class classification of Aβ burden (Negative, Weakly positive, Positive, and Clearly positive), observing good performance in hold-out data from the ADNI, HABS-HD, and POINTER training cohorts. ADLLM demonstrated predictive performance comparable to or higher than the zero-shot LLM and GB models across most cohorts (Table 2). The highest performance was observed in the ADNI cohort, where ADLLM achieved the strongest AUC, with similarly competitive performance in HABS-HD and POINTER. Overall, discrimination for ADLLM was consistently higher than the zero-shot LLM and similar to or slightly higher than GB across cohorts. Accuracy was broadly comparable between ADLLM and the GB model, with small cohort-specific differences, but ADLLM demonstrated greater specificity in its predictions (Table 2; Figure 2). Notably, ADLLM performed particularly well at identifying the extreme categories—clearly positive and clearly negative Aβ burden (Figure 2B).

In the external validation A4 dataset, we observed similar predictive performance using ADLLM compared to GB, and slightly higher discrimination than the zero-shot LLM (Table 2). Accuracy remained modest across models, while specificity was similar across methods. Notably though, when employed to screen out the Aβ positive candidates, ADLLM consistently identified more true-positive candidates than GB in all investigated cohorts. Specifically, ADLLM identified additional true positives in ADNI (55 more), HABS-HD (11 more), POINTER (2 more), and A4 (37 more). Transforming the discrete predictions back to original continuous values recovered continuous variability, which correlated strongly with the raw Aβ CL (r= 0·584; Supplementary Figure 1a).

Finally, to further evaluate model performance in biomarker-rich settings, we analyzed subsets of the ADNI and HABS-HD cohorts with available fluid biomarkers. For ADNI, cerebrospinal fluid (CSF) Aβ40, Aβ42, pTau-181, and total tau were available, whereas HABS-HD included plasma measurements of the same markers. The inclusion of fluid biomarker information improved classification of Aβ burden in all metrics in both cohorts (Supplementary Figure 2). In ADNI,

multi-class discrimination of levels of Aβ burden increased substantially (AUC 0.851 vs 0.760), while in HABS-HD, improvements were more modest but consistent (AUC 0.861 vs 0.857) (Supplementary Table 3).

Meta-tau burden

We used the composite-outcome trained ADLLM (i.e. same model that predicted Aβ burden) and a new GB model trained on the same 161 features to perform a four-class classification of meta temporal -tau burden (Very low, Low, High, and Very high). We observed better performance of ADLLM over GB in making hold-out predictions in the training cohorts with higher AUC, accuracies, and specificity (Table 3). Similarly, within the external validation dataset, ADLLM outperformed GB with higher AUC (A4 0·600 [0·564–0·635] vs. 0·505 [0·456 – 0·558] vs. 0·561[0·524–0·599]), accuracy (A4 0·284 vs. 0·173 vs. 0·277) and specificity (A4 0·789 vs. 0·832 vs. 0·752) (Figure 3). Moreover, converting the ADLLM predicted discrete labels back to the original continuous predicted values resulted in positive correlation of r=0·543 with the ground truth Meta-tau burden (Supplementary Figure 1b).

In subgroup analyses of participants with available fluid biomarkers, incorporation of CSF biomarkers in ADNI resulted in modest improvement in four class prediction accuracy (0.467 vs. 0.336), with minimal change in AUC (0.707 vs. 0.700) and slightly lower specificity (0.803 vs. 0.830). In contrast, inclusion of plasma biomarkers in HABS-HD resulted in slightly lower performance, with AUC decreasing from 0.602 to 0.575, accuracy from 0.422 to 0.411, and specificity from 0.710 to 0.703 (Supplementary Table 4). These findings suggest that fluid biomarkers provided limited added value for meta-temporal tau prediction, with modest benefit in ADNI but not in HABS-HD.

MMSE score change rate

We used the composite-outcome trained ADLLM (i.e. same model that predicted Aβ and meta temporal tau burden) and a new GB model trained on the same 161 features to predict four-class classifications of MMSE score change rates (Stable, Slow decline, Moderate decline, and Fast decline). Across all evaluation metrics and in both hold-out validation cohorts, ADLLM consistently outperformed both the zero-shot LLM and the GB model (Table 4Error: Reference source not found). The strongest discrimination was observed in the HABS-HD cohort (AUC 0.883), with similarly strong performance in ADNI (AUC 0.806). Accuracy followed a similar pattern, with ADLLM outperforming the comparison models in both datasets. ADLLM also excelled in identifying individuals with fast MMSE decline, achieving high discriminative ability in the ADNI dataset (AUC=0.90; Figure 4). These results indicate that the composite-outcome trained ADLLM captures features predictive of future cognitive decline more effectively than the comparison models. Transforming the discrete predictions back to the original continuous values recovered continuous variability. Correlating ADLLM predicted continuous values with the raw MMSE change rates yielded a strong positive correlation of r= 0.600 (Supplementary Figure 1c). Finally, With increasing longitudinal data, ADLLM performance improved substantially. When the number of visits exceeded nine, AUC approached 0.90 and accuracy approached 0.70 (Figure 4), indicating strong utilization of longitudinal information.

Attention scores in LLM and variable importance in ML

To determine the most important variables for making the composite prediction of Aβ, meta-temporal tau burden, and MMSE change rate, we extracted attention scores from ADLLM's transformer layers across different cohorts and variable groups. The attention scores in ADLLM

showed consistent patterns and cohort-specific variability. Notably, AD-related fluid biomarkers emerged as highly attended variable groups, with "AD CSF biomarker" receiving high attention scores exclusively in ADNI and "AD blood biomarker" standing out in HABS-HD. Variable groups such as "Genetics" and "Cortical volume" consistently showed high attention scores across all cohorts, whereas "Cognitive tests" and "Questionnaires" received lower attention scores (Figure 5).

To further explore the model's interpretability, we instructed ADLLM to explain its predictions by prompting, "Provide your reasons for the prediction." Examples from ADNI, along with their corresponding predictions and explanations, are presented in Figure 6 briefly and in Supplementary Methods: ADLLM Reasoning Examples From ADNI Cohort in detail.

The variable importance of GB was also computed across different outcome prediction tasks. For Aβ burden prediction, the top five variable groups in feature importance were "Genetics", "AD CSF biomarker", "Cognitive Diagnosis", "Cortical volume", and "Cortical thickness". For Meta tau prediction, the top five were "Cognitive Diagnosis", "Cortical volume", "Cortical thickness", "Other health conditions", and "Demographics". For MMSE score change rate prediction, "Cognitive Diagnosis", "Medication", "Cognitive tests", "Cortical volume", and "Cortical thickness" ranked highest. More detailed results of GB features' importance are provided in the Supplementary Figure 3.

Heterogeneity effects:

Subgroup performance and bias analysis

To evaluate potential heterogeneity in model performance, we conducted subgroup analyses across sex, age group, race, ethnicity, and education level where sufficient sample size was available. Model performance metrics (weighted AUC, macro PR-AUC, accuracy, and macro sensitivity) were computed within each subgroup using participant-level bootstrap confidence intervals.

Across well-powered subgroups, ADLLM demonstrated broadly similar performance across sex and age groups. For example, in the A4 cohort for Aβ prediction, weighted AUC values were approximately 0.52 for female participants and 0.49 for male participants, with overlapping bootstrap confidence intervals. Consistent patterns were observed in the HABS-HD cohort, where Aβ prediction AUC was 0.512 in female participants and 0.498 in male participants, and meta-temporal tau prediction AUC was 0.603 vs 0.587, respectively. For MMSE change rate prediction, performance was similarly consistent across sex groups in HABS-HD (AUC 0.881 in females vs 0.872 in males). Differences across age tertiles were similarly small (ΔAUC typically <0.04).

Analyses across race, ethnicity, and education strata yielded comparable findings for the major groups represented in the cohorts. In HABS-HD, Aβ prediction AUC was 0.605 in White participants and 0.589 in Black/African American participants, with similar patterns observed for meta-temporal tau prediction (AUC 0.598 vs 0.581, respectively). In most cases, differences in AUC and PR-AUC between groups were modest and confidence intervals overlapped substantially. Several smaller demographic categories contained limited numbers of participants and were therefore flagged as underpowered, resulting in wider uncertainty intervals.

Overall, these analyses did not identify strong or systematic disparities in ADLLM predictive performance across major demographic subgroups in the available datasets, although further

validation in larger and more diverse populations will be important for confirming fairness and generalizability.

DISCUSSION

Here we present a novel modeling framework, ADLLM, a fine-tuned LLM that is specifically trained to make AD-related predictions based on a broad range of clinical, demographic, and biomarker predictors. Using state-of-the-art LLM enhancement methods and extensive training data from a diverse set of participants, we show the ADLLM can utilize non-PET predictors to capture subject-level variability in Aβ and tau burden measured with PET, while also forecasting cognitive decline. Critically, we probed ADLLM to generate explanations for its predictions and observed that ADLLM integrates semantic information captured across multimodal predictors to make an accurate and clinically reasonable assessment of a patient's underlying AD pathology and incipient cognitive decline.

Across internal hold-out cohorts and external validation, ADLLM demonstrated competitive performance relative to GB, with its clearest advantage observed for predicting MMSE change rate, whereas performance gains for Aβ and meta-temporal tau burden were more modest and varied across cohorts and metrics. These findings suggest that ADLLM may be particularly well suited for integrating heterogeneous longitudinal information relevant to cognitive decline, while multi-class biomarker prediction remains challenging because of class imbalance, cohort differences, and the difficulty of inferring PET-derived pathology from non-PET predictors. Subgroup analyses across sex, age, race, ethnicity, and education did not identify strong systematic disparities in predictive performance, although several smaller demographic strata were underpowered and require further validation. Beyond model discrimination, ADLLM provides a flexible framework for AD prediction because structured clinical, demographic, biomarker, and imaging-derived predictors can be serialized into natural language prompts, enabling integration of new variables, partially missing inputs, and potentially unstructured clinical information without redesigning task-specific feature engineering pipelines. This prompt-based structure is consistent with the broader clinical LLM literature suggesting that language-model interfaces may lower implementation barriers for heterogeneous clinical data relative to conventional ML workflows that often require extensive preprocessing, feature engineering, and task-specific retraining.[46] Thus, ADLLM's main contribution is not only improved performance for selected outcomes, but also a scalable and adaptable prediction framework for multimodal longitudinal AD monitoring.

When assessing predictor relevance between ADLLM and GB, we saw broad alignment, with both approaches identifying key predictive features, such as "Cognitive Diagnosis," "Genetics," "Cortical volume," and "Cortical thickness," as highly influential across all prediction tasks. This indicates that ADLLM accurately integrates the numerical values of the predictors with semantic information already present in the pre-trained model. Furthermore, in line with more traditional supervised learning models[47–52] the addition of fluid biomarkers can improve ADLLM performance in certain tasks, highlighting the compatibility of ADLLM with emerging biomarkers for AD.

The vast majority of previous modeling approaches to predict similar variables utilize more traditional supervised ML algorithms with similar predictive performance to ADLLM in predicting Aβ positivity[11,53,54], tau PET positivity[13,55], and future cognitive decline[56–59]. However, these previous approaches only integrate a select number of predictors, and none attempt to predict all three outcomes simultaneously as we do with ADLLM. Furthermore, there is limited robust external validation and cross-cohort generalizability in these prior works, a particular concern with

approaches that use heavily parameterized models such as deep neural networks[56–59] , or require heavily engineered features. In contrast, our ADLLM generalizes across datasets, achieving stable performance even when applied to external cohorts. Furthermore, our ADLLM naturally accommodates feature complexity by processing serialized textual representations of clinical data, bridging structured and unstructured sources within a uniform serialization pipeline.

ADLLM enhances interpretability compared to traditional prediction architectures. Traditional deep learning techniques are often regarded as "black boxes," with limited interpretability of understanding of how predictions are made. However, post-hoc analyses using attention scores and generated reasoning provide exploratory insight into which input features may influence model predictions. Though these approaches do not establish causal interpretability, this is useful in clinical settings, allowing clinicians and patients to understand the rationale behind predictions and build trust in the model's outputs.

Despite its strengths, ADLLM has some potential limitations. First, deeper probing of feature importance in ADLLM's decision-making process may be required, as it still remains unclear if model attention scores are directly proportional to predictor importance.[43,60] Second, there are substantial computational resources required to fine-tune the pre-trained LLM (~10 hours), far greater than the resources to train classical ML models (~2 minutes). However, this is a one-time cost, and once fine-tuned, ADLLM can produce predictions within seconds for a wide range of input scenarios, which may help with first-pass screening for patients with suspected cognitive decline, early signs of Alzheimer's disease, or incomplete clinical profiles. Third, serialized tabulated predictors could potentially be aggregated with unstructured electronic health record information to improve classification performance. Additionally, expanding the genetic feature set may enhance the model's ability to predict tau burden. Finally, predictive information from MRI may also be underutilized in our current setup due to reliance on summary measures which may carry information loss from the original 3D image.

Conclusion:
Here we present ADLLM as a novel tool in AI-guided AD prediction. ADLLM demonstrated promising generalizability, interpretability, and performance compared to traditional ML models when predicting composite outcomes using multimodal longitudinal predictors in four large and diverse cohorts (ADNI, HABS-HD, POINTER, and A4). To our knowledge, ADLLM is the first study to develop and fine-tune a domain-specific language model for the prediction of AD pathology and cognitive decline. Through the efficient embedding of AD predictors, an accurate and reliable semantic association can be made by ADLLM with the model justifications providing a reasonable summary and synthesis of a patient's underlying Aβ and tau pathology while also accurately forecasting cognitive decline.

Our findings demonstrate the strong potential of fine-tuned LLMs to advance precision medicine for Alzheimer's disease, providing the groundwork for LLMs to serve as valuable and trustworthy decision-support tools in neurodegenerative disease care.

DATA AVAILABILITY

ADNI, HABS-HD, POINTER, and A4 studies data are available upon access approval through the following website.
ADNI website: https://adni.loni.usc.edu/.
HABS-HD website: https://apps.unthsc.edu/itr/.
POINTER website: https://www.alz.org/us-pointer/home.asp.

A4 website: https://atri.usc.edu/study/a4-study/.

CODE AVAILABILITY

The codes for LLM fine-tuning, LLM inference, GB training, and GB inference in this study are available at LQRrrrr/ADLLM (github.com). Our fine-tuned LoRA adaptor is also available through GitHub.

ACKNOWLEDGEMENTS

J.G. is supported by the Alzheimer's Association (23AARF-1026883). Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California. Research reported in this publication was supported by the National Institute on Aging of the National Institutes of Health under Award Numbers R01AG054073 and R01AG058533, R01AG070862, P41EB015922 and U19AG078109. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. U.S. POINTER is supported through funding provided by the Alzheimer's Association (POINTER-19-611541, led by Dr Laura Baker) and the NIA (R01AG062689). The A4 study is a secondary prevention trial in preclinical AD, aiming to slow cognitive decline associated with brain amyloid accumulation in clinically normal older individuals. The A4 study is funded by a public-private-philanthropic partnership, including funding from the NIH-National Institute on Aging, Eli Lilly and Co, Alzheimer's Association, Accelerating Medicines Partnership, GHR Foundation, an anonymous foundation, and additional private donors, with in-kind support from Avid and Cogstate. The companion observational LEARN study is funded by the Alzheimer's Association and GHR Foundation. The A4 and LEARN studies are led by Dr. Reisa Sperling at Brigham and Women's Hospital, Harvard Medical School, and Dr. Paul Aisen at the Alzheimer's Therapeutic Research Institute (ATRI), University of Southern California. The A4 and LEARN Studies are coordinated by ATRI at the

University of Southern California, and the data are made available through the Laboratory for Neuro Imaging at the University of Southern California. The participants screened for the A4 study provided permission to share their deidentified data to advance the quest to find a successful treatment for AD. The authors acknowledge the dedication of all the participants, the site personnel, and all of the partnership team members who continue to make the A4 and LEARN studies possible. The complete A4 Study Team list is available at a4study.org/ a4-study-team.

References

1. Chowdhary, N. *et al.* Reducing the Risk of Cognitive Decline and Dementia: WHO Recommendations. *Front. Neurol.* **12**, 765584 (2022).
2. Klyucherev, T. O. *et al.* Advances in the development of new biomarkers for Alzheimer's disease. *Transl. Neurodegener.* **11**, 25 (2022).
3. Li, X. *et al.* Global, regional, and national burden of Alzheimer's disease and other dementias, 1990–2019. *Front. Aging Neurosci.* **14**, 937486 (2022).
4. Porsteinsson, A. P., Isaacson, R. S., Knox, S., Sabbagh, M. N. & Rubino, I. Diagnosis of Early Alzheimer's Disease: Clinical Practice in 2021. *J. Prev. Alzheimers Dis.* **8**, 371–386 (2021).
5. World Health Organization. *Risk Reduction of Cognitive Decline and Dementia: WHO Guidelines*. (World Health Organization, Geneva, 2019).
6. Jack, C. R. *et al.* Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade. *Lancet Neurol.* **9**, 119–128 (2010).
7. De La Torre, J. C. Alzheimer's Disease is Incurable but Preventable. *J. Alzheimers Dis.* **20**, 861–870 (2010).
8. De Strooper, B. & Karran, E. The Cellular Phase of Alzheimer's Disease. *Cell* **164**, 603–615 (2016).
9. for the Alzheimer's Disease Neuroimaging Initiative *et al.* Predicting the course of Alzheimer's progression. *Brain Inform.* **6**, 6 (2019).
10. Van Der Flier, W. M., De Vugt, M. E., Smets, E. M. A., Blom, M. & Teunissen, C. E. Towards a future where Alzheimer's disease pathology is stopped before the onset of dementia. *Nat. Aging* **3**, 494–505 (2023).
11. Moradi, E. *et al.* Machine learning prediction of future amyloid beta positivity in amyloid-negative individuals. *Alzheimers Res. Ther.* **16**, 46 (2024).
12. Shan, G., Bernick, C., Caldwell, J. Z. K. & Ritter, A. Machine learning methods to predict amyloid positivity using domain scores from cognitive tests. *Sci. Rep.* **11**, 4822 (2021).
13. Karlsson, L. *et al.* A machine learning-based prediction of tau load and distribution in Alzheimer's disease using plasma, MRI and clinical variables. Preprint at https://doi.org/10.1101/2024.05.31.24308264 (2024).
14. Cai, Y. *et al.* Comparing machine learning-derived MRI-based and blood-based neurodegeneration biomarkers in predicting syndromal conversion in early AD. *Alzheimers Dement.* **19**, 4987–4998 (2023).
15. Franciotti, R., Nardini, D., Russo, M., Onofrj, M. & Sensi, S. L. Comparison of Machine Learning-based Approaches to Predict the Conversion to Alzheimer's Disease from Mild Cognitive Impairment. *Neuroscience* **514**, 143–152 (2023).
16. Rossini, P. M., Miraglia, F. & Vecchio, F. Early dementia diagnosis, MCI-to-dementia risk prediction, and the role of machine learning methods for feature extraction from integrated

biomarkers, in particular for EEG signal analysis. *Alzheimers Dement.* **18**, 2699–2706 (2022).
17. Anil, R. *et al.* PaLM 2 Technical Report. Preprint at https://doi.org/10.48550/arXiv.2305.10403 (2023).
18. Brown, T. B. *et al.* Language Models are Few-Shot Learners. Preprint at https://doi.org/10.48550/arXiv.2005.14165 (2020).
19. Grattafiori, A. *et al.* The Llama 3 Herd of Models. Preprint at https://doi.org/10.48550/arXiv.2407.21783 (2024).
20. Zhang, S. *et al.* OPT: Open Pre-trained Transformer Language Models. Preprint at https://doi.org/10.48550/arXiv.2205.01068 (2022).
21. Agbavor, F. & Liang, H. Predicting dementia from spontaneous speech using large language models. *PLOS Digit. Health* **1**, e0000168 (2022).
22. B.T, B. & Chen, J.-M. Performance Assessment of ChatGPT versus Bard in Detecting Alzheimer's Dementia. *Diagnostics* **14**, 817 (2024).
23. Casu, F., Grosso, E., Lagorio, A. & Trunfio, G. A. Optimizing and Evaluating Pre- Trained Large Language Models for Alzheimer's Disease Detection. in *2024 32nd Euromicro International Conference on Parallel, Distributed and Network-Based Processing (PDP)* 277–284 (2024). doi:10.1109/PDP62718.2024.00046.
24. Chen, C.-P. & Li, J.-L. Profiling Patient Transcript Using Large Language Model Reasoning Augmentation for Alzheimer's Disease Detection. Preprint at https://doi.org/10.48550/arXiv.2409.12541 (2024).
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* 2024.04.03.24305298 (2024) doi:10.1101/2024.04.03.24305298.
26. Mao, C. *et al.* AD-BERT: Using pre-trained language model to predict the progression from mild cognitive impairment to Alzheimer's disease. *J. Biomed. Inform.* **144**, 104442 (2023).
27. Wang, J. *et al.* Augmented Risk Prediction for the Onset of Alzheimer's Disease from Electronic Health Records with Large Language Models. Preprint at https://doi.org/10.48550/arXiv.2405.16413 (2024).
28. Yan, C. *et al.* Leveraging generative AI to prioritize drug repurposing candidates for Alzheimer's disease with real-world clinical validation. *Npj Digit. Med.* **7**, 46 (2024).
29. Feng, Y., Xu, X., Zhuang, Y. & Zhang, M. Large Language Models Improve Alzheimer's Disease Diagnosis Using Multi-Modality Data. in *2023 IEEE International Conference on Medical Artificial Intelligence (MedAI)* 61–66 (2023). doi:10.1109/MedAI59581.2023.00016.
30. Li, D. *et al.* DALK: Dynamic Co-Augmentation of LLMs and KG to answer Alzheimer's Disease Questions with Scientific Literature. Preprint at https://doi.org/10.48550/arXiv.2405.04819 (2024).
31. Wolfe, R. *et al.* Laboratory-Scale AI: Open-Weight Models are Competitive with ChatGPT Even in Low-Resource Settings. in *The 2024 ACM Conference on Fairness, Accountability, and Transparency* 1199–1210 (2024). doi:10.1145/3630106.3658966.
32. Wang, Y., Lin, Y., Zeng, X. & Zhang, G. PrivateLoRA For Efficient Privacy Preserving LLM. Preprint at https://doi.org/10.48550/arXiv.2311.14030 (2023).
33. Borzunov, A. *et al.* Distributed Inference and Fine-tuning of Large Language Models Over The Internet. Preprint at https://doi.org/10.48550/arXiv.2312.08361 (2023).
34. Hu, E. J. *et al.* LoRA: Low-Rank Adaptation of Large Language Models. Preprint at https://doi.org/10.48550/arXiv.2106.09685 (2021).

35. Harrison, T. M. *et al.* The POINTER Imaging baseline cohort: Associations between multimodal neuroimaging biomarkers, cardiovascular health, and cognition. *Alzheimers Dement.* **21**, e14399 (2025).
36. Baker, L. D. *et al.* Study design and methods: U.S. study to protect brain health through lifestyle intervention to reduce risk (U.S. POINTER). *Alzheimers Dement.* **20**, 769–782 (2024).
37. Wang, G. *et al.* The CentiMarker Project: Standardizing Quantitative Alzheimer's disease Fluid Biomarkers for Biologic Interpretation. Preprint at https://doi.org/10.1101/2024.07.25.24311002 (2024).
38. Amadoru, S. *et al.* Comparison of amyloid PET measured in Centiloid units with neuropathological findings in Alzheimer's disease. *Alzheimers Res. Ther.* **12**, 22 (2020).
39. Iaccarino, L. *et al.* A practical overview of the use of amyloid-PET Centiloid values in clinical trials and research. *NeuroImage Clin.* **46**, 103765 (2025).
40. Meister, C. & Cotterell, R. Language Model Evaluation Beyond Perplexity. Preprint at https://doi.org/10.48550/arXiv.2106.00085 (2021).
41. Huang, Y. *et al.* Look Before You Leap: An Exploratory Study of Uncertainty Measurement for Large Language Models. *IEEE Trans. Softw. Eng.* 1–18 (2025) doi:10.1109/TSE.2024.3519464.
42. Efron, B. & Tibshirani, R. J. *An Introduction to the Bootstrap*. (Chapman and Hall/CRC, 1994). doi:10.1201/9780429246593.
43. Wiegreffe, S. & Pinter, Y. Attention is not not Explanation. Preprint at https://doi.org/10.48550/arXiv.1908.04626 (2019).
44. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V. & Gulin, A. CatBoost: unbiased boosting with categorical features. Preprint at https://doi.org/10.48550/arXiv.1706.09516 (2019).
45. Dorogush, A. V., Ershov, V. & Gulin, A. CatBoost: gradient boosting with categorical features support.
46. Sivarajkumar, S., Kelley, M., Samolyk-Mazzanti, A., Visweswaran, S. & Wang, Y. An Empirical Evaluation of Prompting Strategies for Large Language Models in Zero-Shot Clinical Natural Language Processing: Algorithm Development and Validation Study. *JMIR Med. Inform.* **12**, e55318 (2024).
47. Chang, C.-H., Lin, C.-H. & Lane, H.-Y. Machine Learning and Novel Biomarkers for the Diagnosis of Alzheimer's Disease. *Int. J. Mol. Sci.* **22**, 2761 (2021).
48. Forlenza, O. V. *et al.* Cerebrospinal fluid biomarkers in Alzheimer's disease: Diagnostic accuracy and prediction of dementia. *Alzheimers Dement. Diagn. Assess. Dis. Monit.* **1**, 455–463 (2015).
49. Graff-Radford, J. *et al.* Cerebrospinal fluid dynamics disorders: Relationship to Alzheimer biomarkers and cognition. *Neurology* **93**, (2019).
50. Lehallier, B. *et al.* Combined Plasma and Cerebrospinal Fluid Signature for the Prediction of Midterm Progression From Mild Cognitive Impairment to Alzheimer Disease. *JAMA Neurol.* **73**, 203 (2016).
51. Perrin, R. J. *et al.* Identification and Validation of Novel Cerebrospinal Fluid Biomarkers for Staging Early Alzheimer's Disease. *PLoS ONE* **6**, e16032 (2011).
52. Sutphen, C. L. *et al.* Longitudinal Cerebrospinal Fluid Biomarker Changes in Preclinical Alzheimer Disease During Middle Age. *JAMA Neurol.* **72**, 1029 (2015).
53. Jung, Y. H. *et al.* Prediction of amyloid β PET positivity using machine learning in patients with suspected cerebral amyloid angiopathy markers. *Sci. Rep.* **10**, 18806 (2020).

54. Youn, Y. C. *et al.* Prediction of amyloid PET positivity via machine learning algorithms trained with EDTA-based blood amyloid-β oligomerization data. *BMC Med. Inform. Decis. Mak.* **22**, 286 (2022).
55. Kim, J. *et al.* Prediction of tau accumulation in prodromal Alzheimer's disease using an ensemble machine learning approach. *Sci. Rep.* **11**, 5706 (2021).
56. Lee, L. Y. *et al.* Robust and interpretable AI-guided marker for early dementia prediction in real-world clinical settings. *eClinicalMedicine* **74**, 102725 (2024).
57. Valizadeh, G., Elahi, R., Hasankhani, Z., Rad, H. S. & Shalbaf, A. Deep Learning Approaches for Early Prediction of Conversion from MCI to AD using MRI and Clinical Data: A Systematic Review. *Arch. Comput. Methods Eng.* https://doi.org/10.1007/s11831-024-10176-6 (2024) doi:10.1007/s11831-024-10176-6.
58. Venugopalan, J., Tong, L., Hassanzadeh, H. R. & Wang, M. D. Multimodal deep learning models for early detection of Alzheimer's disease stage. *Sci. Rep.* **11**, 3254 (2021).
59. Bron, E. E. *et al.* Cross-cohort generalizability of deep and conventional machine learning for MRI-based diagnosis and prediction of Alzheimer's disease. *NeuroImage Clin.* **31**, 102712 (2021).
60. Jain, S. & Wallace, B. C. Attention is not Explanation. Preprint at https://doi.org/10.48550/arXiv.1902.10186 (2019).

Table 1. Demographic variables are summarized per participant, ensuring that individuals with multiple visits contribute only once to demographic counts. The number of visits per participant is reported separately to describe longitudinal follow-up. Outcome class counts represent the distribution of available measurements across visits, reflecting class balance in the prediction tasks.

Characteristic	**A4 Training**	**A4 Validation**	**ADNI Training**	**ADNI Validation**	**HABS-HD Training**	**HABS-HD Validation**	**POINTER Training**	**POINTER Validation**
Participants, n	352	337	1878	456	1872	1198	749	184
Records, n	352	541	9906	2464	3129	742	749	188
Visits per participant, mean (SD)	1.00 (0.00)	1.61 (0.73)	5.27 (3.17)	7.61 (4.96)	1.67 (1.62)	5.08 (8.27)	1.00 (0.00)	2.00 (0.33)
Age, mean (SD)	71.53 (4.77)	70.78 (4.59)	72.93 (7.28)	72.81 (7.64)	64.85 (8.46)	64.27 (8.28)	68.48 (5.20)	68.11 (5.36)
Education	16.28	16.63	16.06	16.06	14.71	13.66	15.97	15.91

Characteristic	A4 Training	A4 Validation	ADNI Training	ADNI Validation	HABS-HD Training	HABS-HD Validation	POINTER Training	POINTER Validation
years, mean (SD)	(2.76)	(2.61)	(2.72)	(2.76)	(2.10)	(4.07)	(2.23)	(2.32)
Sex								
Male	147	133	972	239	674	469	284	79
Female	205	204	874	217	1198	729	465	105
Race								
White	320	316	1638	400	1312	721	515	118
Black / African American	8	6	131	34	544	459	110	25
Asian	18	10	43	13	0	—	22	8
Multiple race	0	—	24	4	13	15	0	—
American Indian / Alaska Native	1	1	2	2	2	2	0	—
Other / Missing	5	4	40	3	1	1	102	33
Ethnicity								

Characteristic	A4 Training	A4 Validation	ADNI Training	ADNI Validation	HABS-HD Training	HABS-HD Validation	POINTER Training	POINTER Validation
Not Hispanic / Latino	338	327	1758	429	1247	814	0	—
Hispanic / Latino	6	9	78	25	625	384	54	—
Amyloid PET (Centiloids)								
Negative	—	66	1653	372	1439	364	543	140
Weakly positive	—	84	280	67	103	30	102	21
Positive	—	104	256	69	78	13	47	11
Clearly positive	—	186	859	237	153	35	57	16
Total measurements	—	440	3048	745	1773	442	749	188
Meta-temporal tau SUVR								
Very low	—	72	242	57	306	98	219	49

Characteristic	A4 Training	A4 Validation	ADNI Training	ADNI Validation	HABS-HD Training	HABS-HD Validation	POINTER Training	POINTER Validation
Low	—	85	321	59	308	50	212	48
High	—	103	297	62	294	74	201	55
Very high	—	180	458	131	221	46	99	28
Total measurements	—	440	1318	309	1129	268	731	180
MMSE change rate								
Stable	—	—	2612	681	731	147	—	—
Slow decline	—	—	2088	428	615	118	—	—
Moderate decline	—	—	2318	611	171	603*	—	—
Fast decline	—	—	2705	694	249	60	—	—
Total measurements	—	—	9723	2414	2198	496	—	—

*Table 2 Four-class Aβ classification performance **across validation cohorts.** Performance of ADLLM, zero-shot LLM, and gradient boosting models for predicting amyloid-beta PET burden across ADNI, HABS-HD, POINTER, and A4 validation data. Amyloid-beta burden was classified into four ordered categories: negative, weakly positive, positive, and clearly positive. Model performance is reported using multi-class weighted AUC with 95% confidence intervals, accuracy, and specificity. Bold values indicate the best-performing model within each cohort and metric.*

Evaluation cohort	Model	AUC (95% CI)	Accuracy	Specificity
ADNI	ADLLM	**0.825 (0.803–0.843)**	0.685	**0.770**
ADNI	Zero-shot LLM	0.713 (0.683–0.741)	0.554	0.741
ADNI	GB	**0.825 (0.804–0.845)**	**0.689**	0.729
HABS-HD	ADLLM	**0.773 (0.721–0.818)**	0.812	0.290
HABS-HD	Zero-shot LLM	0.658 (0.601–0.714)	0.735	**0.346**
HABS-HD	GB	0.746 (0.691–0.797)	**0.826**	0.174
POINTER	ADLLM	**0.702 (0.625–0.770)**	0.734	0.285
POINTER	Zero-shot LLM	0.631 (0.552–0.708)	0.644	**0.423**
POINTER	GB	0.664 (0.587–0.737)	**0.750**	0.250
A4	ADLLM	**0.610 (0.587–0.632)**	**0.256**	0.837
A4	Zero-shot LLM	0.591 (0.565–0.616)	0.173	0.843
A4	GB	0.609 (0.587–0.632)	0.204	**0.852**

*Table 3 **Four-class meta-temporal tau burden classification performance across validation cohorts.** Performance of ADLLM, zero-shot LLM, and gradient boosting models for predicting meta-temporal tau PET burden across ADNI, HABS-HD, POINTER, and A4 validation data. Meta-temporal tau burden was classified into four ordered categories: very low, low, high, and very high. Model performance is reported using multi-class weighted AUC with 95% confidence intervals, accuracy, and specificity. Bold values indicate the best-performing model within each cohort and metric.*

Cohort	Model	AUC (95% CI)	Accuracy	Specificity
ADNI	ADLLM	**0.730 (0.692–0.767)**	**0.460**	**0.806**
ADNI	Zero-shot LLM	0.551 (0.509–0.592)	0.194	0.803
ADNI	GB	0.707 (0.668–0.745)	0.414	0.761
HABS-HD	ADLLM	**0.597 (0.551–0.643)**	**0.399**	**0.708**
HABS-HD	Zero-shot LLM	0.530 (0.486–0.573)	0.381	0.694
HABS-HD	GB	0.580 (0.538–0.622)	0.354	0.695
POINTER	ADLLM	**0.516 (0.456–0.574)**	**0.278**	0.711
POINTER	Zero-shot LLM	0.445 (0.392–0.498)	0.261	**0.724**
POINTER	GB	0.494 (0.440–0.545)	0.276	0.693
A4	ADLLM	**0.600 (0.564–0.635)**	**0.284**	0.789
A4	Zero-shot LLM	0.505 (0.456–0.558)	0.173	**0.832**
A4	GB	0.561 (0.524–0.599)	0.277	0.752

Table 4 ***Four-class MMSE change-rate classification performance in longitudinal validation cohorts.*** *Performance of ADLLM, zero-shot LLM, and gradient boosting models for predicting longitudinal MMSE change-rate category in ADNI and HABS-HD validation data. MMSE change rate was classified into stable, slow decline, moderate decline, and fast decline categories. Model performance is reported using multi-class weighted AUC with 95% confidence intervals, accuracy, and specificity. Bold values indicate the best-performing model within each cohort and metric.*

Cohort	Model	AUC (95% CI)	Accuracy	Specificity
ADNI	ADLLM	**0.806 (0.794–0.817)**	**0.573**	**0.851**
ADNI	Zero-shot LLM	0.591 (0.576–0.605)	0.274	0.775
ADNI	GB	0.771 (0.759–0.782)	0.516	0.824
HABS-HD	ADLLM	**0.883 (0.862–0.904)**	**0.677**	**0.872**
HABS-HD	Zero-shot LLM	0.576 (0.542–0.608)	0.597	0.742
HABS-HD	GB	0.610 (0.577–0.642)	0.337	0.741

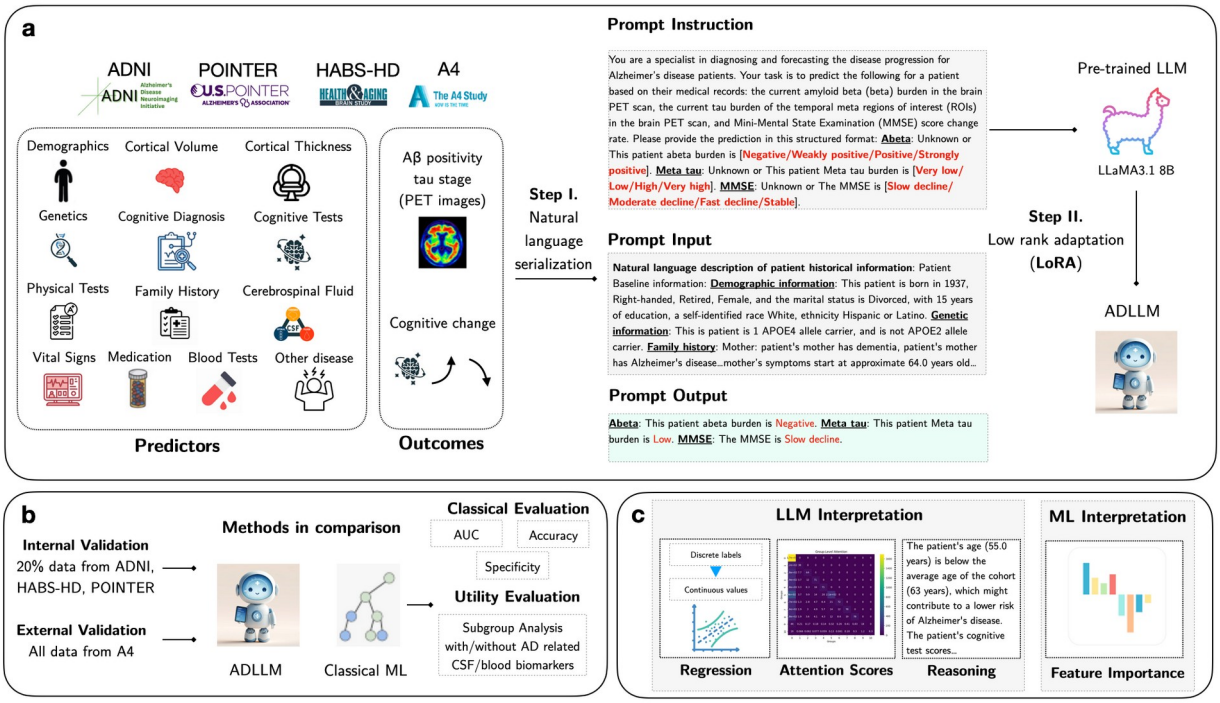

Figure 1 ***Overview of the study design. a.*** *Pre-trained model enhancement: the pre-trained model LLaMA 3.1 with 8 billion parameters was fine-tuned with Low-rank adaptation (LoRA) by serialized data from ADNI, HABS-HD, and POINTER study cohorts. The enhanced model is named "ADLLM".* ***b.*** *Internal and external validation: ADLLM was internally validated by the left 20% of serialized data from ADNI, HABS-HD, and POINTER study cohorts. ADLLM was externally validated by all the serialized data from the A4 study cohort. The classical evaluation metrics included multi-class weighted AUC, Accuracy, and multi-class weighted Specificity. The utility evaluation was the performance of ADLLM on identifying participants with Aβ positivity as eligible candidates for anti-Aβ drug treatment and the subgroup analysis of participants with CSF/blood biomarkers.* ***c.*** *Model interpretation: we first converted the discrete labels back to the continuous scales and did the regression analysis with respect to the original values; we extracted the attention scores across all transformers' layers and generated reasoning for model predictions to interpret ADLLM. For ML interpretation, we extracted the feature importance of the trained GB model.*

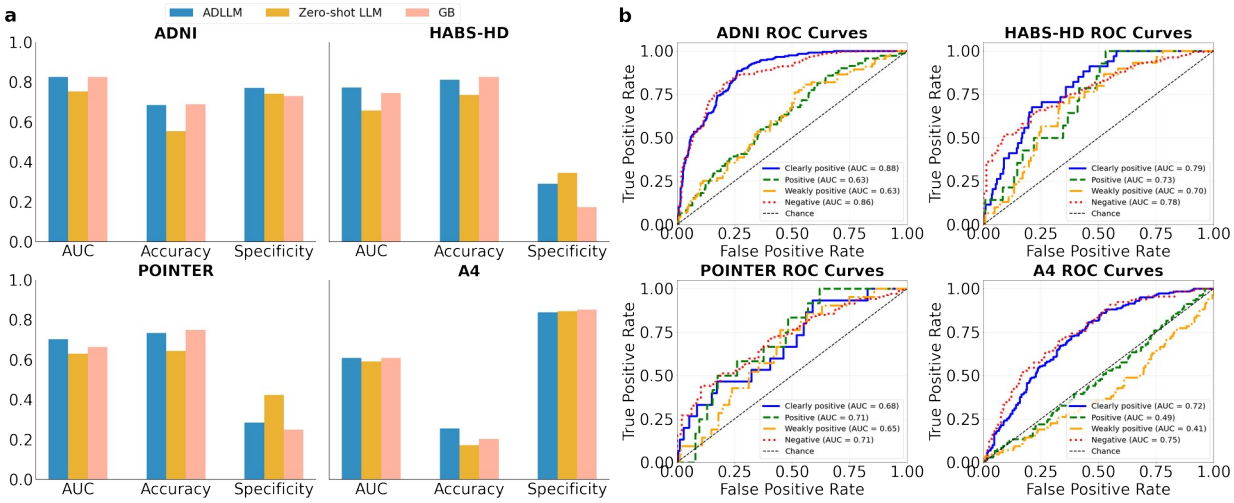

Figure 2 ADLLM and GB performance in Aβ burden prediction. a. *Aβ was categorized into four classes: Clearly positive, Positive, Weakly positive, and Negative. We showed ADLLM, Zero-shot LLM, and GB performance on evaluation metrics: AUC, Accuracy, and Specificity.* ***b.*** *ADLLM ROC curves for different Aβ burden categories across different cohorts.*

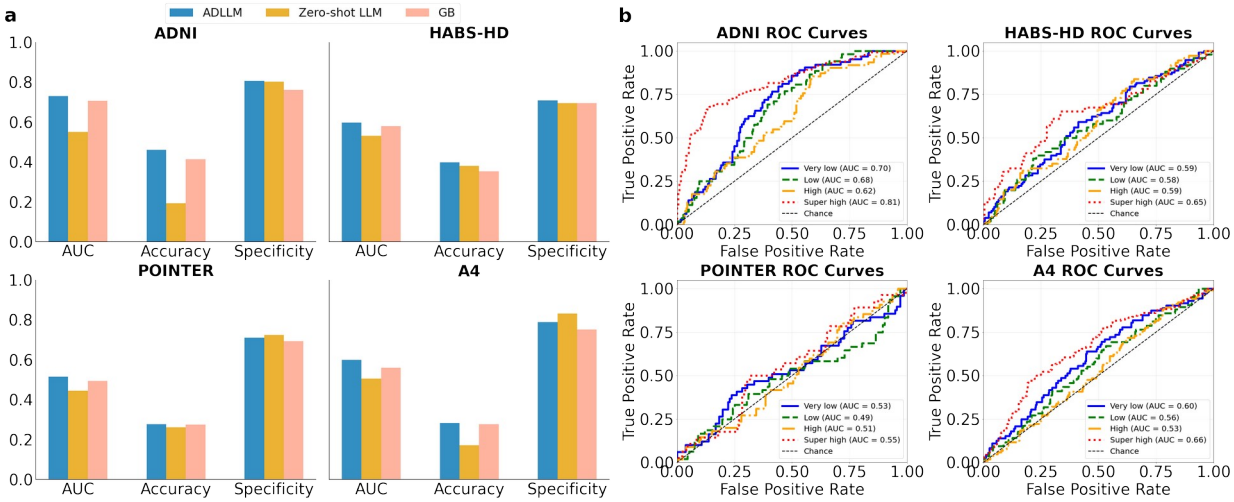

Figure 3 ADLLM and GB performance in Meta tau burden prediction. a. *Meta tau burden was categorized into four classes: Very low, Low, and Very high. We showed ADLLM and GB performance on evaluation metrics: AUC, Accuracy, and Specificity.* ***b.*** *ADLLM ROC curves for different Meta tau burden categories across different cohorts.*

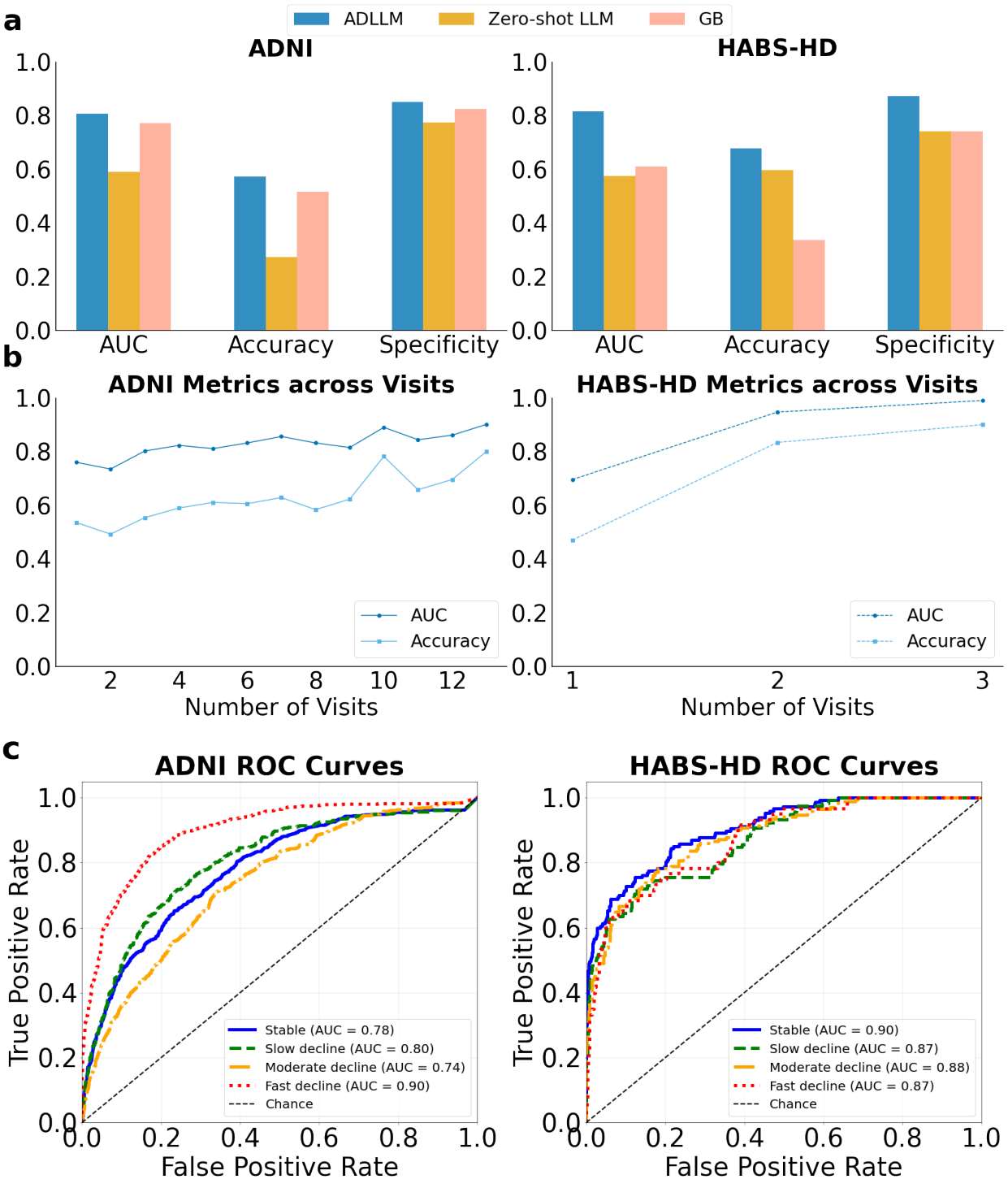

Figure 4 ADLLM and GB performance in MMSE score change rate prediction. a. *MMSE score change rate was categorized into four classes: Stable, Slow decline, Moderate decline, and Fast decline, and two classes: declining and not declining. We showed ADLLM and GB performance on evaluation metrics: AUC, Accuracy, and Specificity.* ***b.*** *The metrics change trends with the increase in participants' visit times.* ***c.*** *ADLLM ROC curves for different MMSE score change rate categories across different cohorts.*

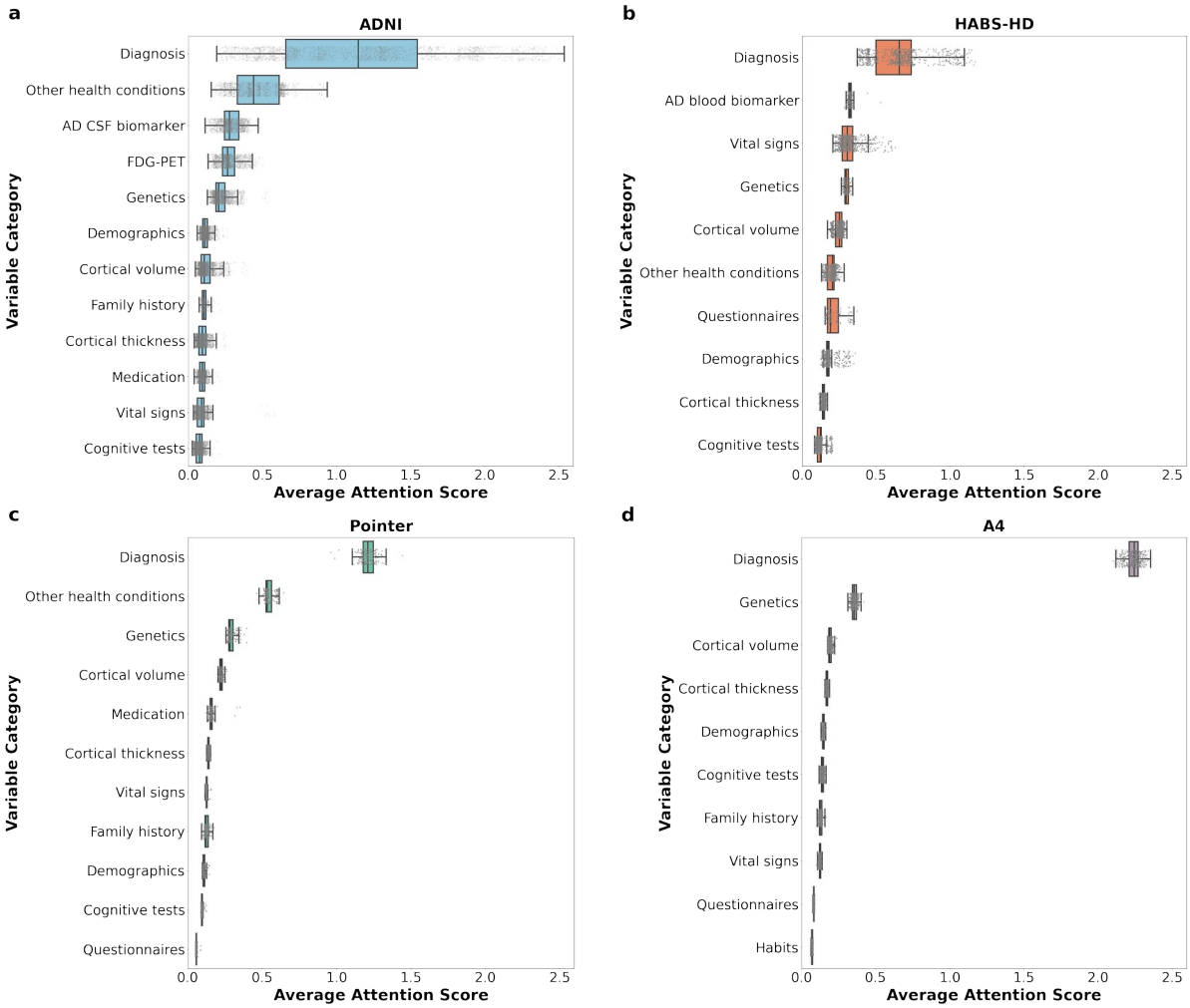

Figure 5 ADLLM attention scores for different variable groups in four cohorts. *The complete list of our defined variable groups and their respective variables can be found in Supplementary Table 1. Our attention score was calculated as the average across multiple heads, multiple layers within the transformer layers, and the token count for each variable category in the text input. See the Methods section and Supplementary Material for detailed calculations.*

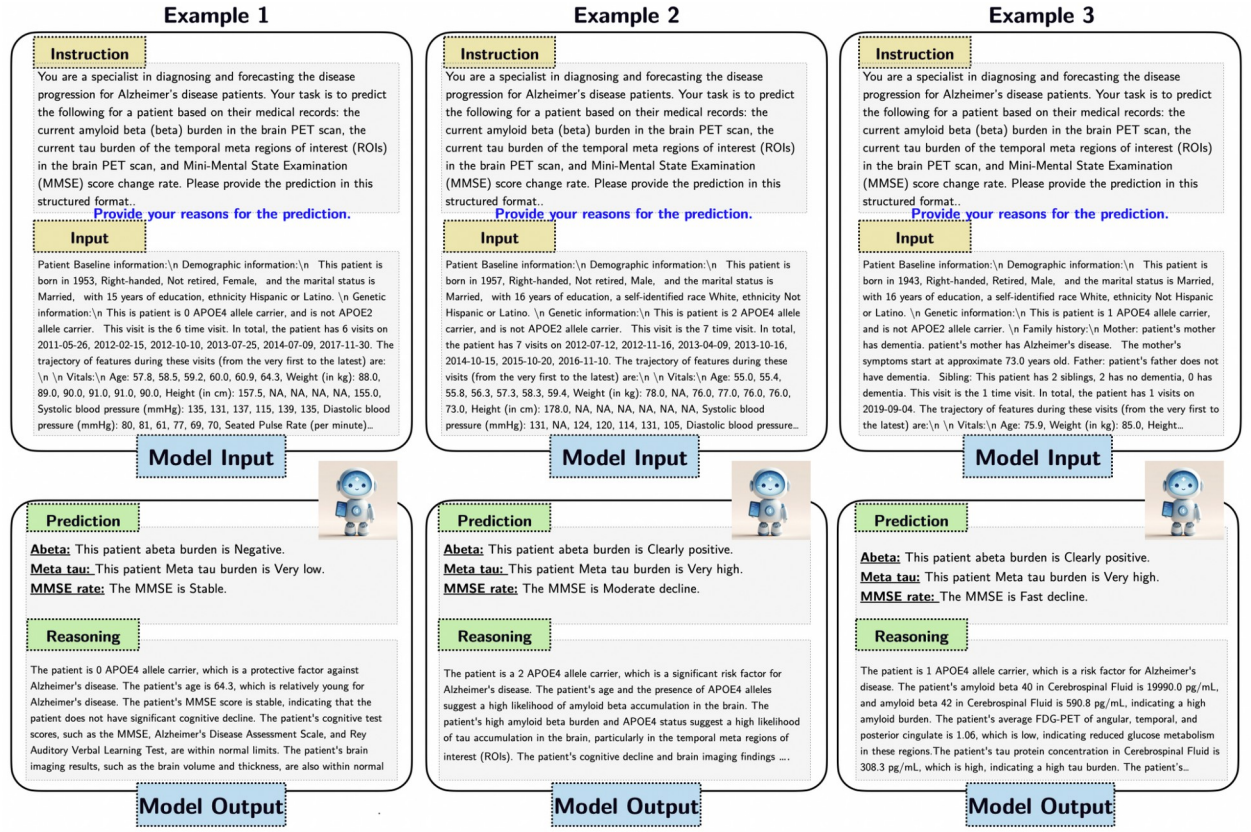

Figure 6 ADLLM prediction reasoning example. *Besides the original instruction, we added the prompt "Provide your reasons for the prediction." to ask ADLLM to give its reasoning for the prediction. From left to right, we provided three examples from ADNI with the predicted outcome and the reasons ADLLM gave for each prediction.*