LLM-MINE: Large Language Model based Alzheimer's Disease and Related Dementias Phenotypes Mining from Clinical Notes

Mingchen Shao, BS[1], Yuzhang Xie, MS[2], Carl Yang, PhD[2], Jiaying Lu, PhD[2*]
[1]Northwestern University, Evanston, IL; [2]Emory University, Atlanta, GA

Abstract

Accurate extraction of Alzheimer's Disease and Related Dementias (ADRD) phenotypes from electronic health records (EHR) is critical for early-stage detection and disease staging. However, this information is usually embedded in unstructured textual data rather than tabular data, making it difficult to be extracted accurately. We therefore propose LLM-MINE, a Large Language Model-based phenotype mining framework for automatic extraction of ADRD phenotypes from clinical notes. Using two expert-defined phenotype lists, we evaluate the extracted phenotypes by examining their statistical significance across cohorts and their utility for unsupervised disease staging. Chi-square analyses confirm statistically significant phenotype differences across cohorts, with memory impairment being the strongest discriminator. Few-shot prompting with the combined phenotype lists achieves the best clustering performance (ARI=0.290, NMI=0.232), substantially outperforming biomedical NER and dictionary-based baselines. Our results demonstrate that LLM-based phenotype extraction is a promising tool for discovering clinically meaningful ADRD signals from unstructured notes.

Introduction

Alzheimer's Disease and Related Dementias (ADRD) encompasses a spectrum of progressive neurocognitive disorders that develop through identifiable stages, typically beginning with normal cognition (CN), progressing to mild cognitive impairment (MCI), and eventually advancing to overt dementia.[1] ADRD affects more than 6 million people in the U.S.[2] and over 55 million people worldwide.[3] Currently, no treatments exist to prevent or halt the progression of ADRD.[4] Throughout the ADRD continuum, individuals exhibit varying phenotypes of cognitive deficits, encompassing domains such as memory, executive function, language, and other cognitive abilities.[5] These phenotypes serve as clinical markers that aid in detecting early cognitive decline, monitoring disease progression, and distinguishing between different stages of impairment.[6] Precise characterization of dementia phenotypes is therefore essential for timely diagnosis, risk stratification, and informing appropriate clinical management strategies.

The phenotypic characterization of ADRD is inherently multidimensional,[7] spanning domains that include demographics, disease manifestations (e.g., comorbidities, symptoms, functional decline), clinical management (e.g., medication), physiological biomarkers (e.g., imaging, labs results, vital signs), and genomic profiles. Electronic health records (EHRs) offer a rich, real-world data source at scale.[8] However, critical ADRD phenotypes are seldom encoded in structured EHR fields, rendering them inaccessible for analytics.[9] For example, cognitive assessments and clinician impressions are often documented in free-text notes or ad hoc tools such as handwritten MMSE or MoCA scores[10] rather than stored in consistent structured tabular formats. Similarly, longitudinal data essential for tracking progression such as cognitive function, neuropsychiatric symptoms, and neurological signs are rarely captured systematically.[11] Consequently, critical phenotype data required for ADRD research remains locked within narrative clinical documentation. Recovering these phenotypes through automated mining is critical for monitoring disease progression, and understanding underlying disease mechanism.

For these reasons, we focus on clinical notes as the primary data source for phenotype mining. Clinical notes contain detailed and nuanced descriptions of cognitive status, symptom trajectories and provider impressions that are essential for characterizing ADRD progression.[11] In addition, structured EHR data can be incomplete or unavailable in many low-resource settings due to limited financial, technological, and organizational infrastructure;[12] in such contexts, documentation may rely heavily on handwritten records or locally stored digital notes. Methods that can robustly extract phenotypes from unstructured text therefore hold broad relevance across diverse healthcare settings. However, extracting meaningful phenotypes from clinical notes presents substantial challenges. Unlike structured data, free-text

*Corresponding Author: Jiaying Lu (jiaying.lu@emory.edu).

requires extensive preprocessing and relies on advanced natural language understanding to capture clinically relevant information.[13] Prior work has commonly relied on predefined vocabularies or traditional natural language processing (NLP) tools.[11,12] While effective in some settings, these approaches are often hard to customize and may not generalize well to complex or institution-specific clinical language.[14]

To address these challenges, we propose LLM-MINE, a large language model-based framework for extracting ADRD-related phenotypes from clinical notes. Unlike traditional NLP approaches that require task-specific training data and extensive feature engineering, LLM-MINE leverages the advanced natural language understanding and domain adaptation capabilities of LLMs. LLM-MINE is both (1) *user-customizable*: allowing researchers to adapt their own phenotype list without retraining the model; and (2) *training-efficient*: requiring minimal task-specific examples rather than large annotated corpus. Specifically, we prompt an LLM to extract ADRD phenotypes predefined by user-specified list from input text. We evaluate LLM-MINE on a large, real-world cohort comprising 8.4K CN, 800 MCI, and 8.4K ADRD patients from the MIMIC-IV database.[15] Given the absence of ground truth annotations, we design a comprehensive, multi-faceted evaluation protocol to indirectly validate the extracted phenotypes. Our contributions are three-fold: (1) *Statistical validation*: we apply chi-square tests to identify significant differences in phenotype prevalence across the CN, MCI, and ADRD groups, demonstrating that LLM-MINE captures clinically meaningful variations aligned with disease progression. (2) *Unsupervised clustering*: we show that the extracted phenotypes cluster patients in a manner consistent with ADRD stages, providing evidence that our method recovers underlying disease structure, even without supervised training. (3) *Superior performance over baselines*: we systematically compare LLM-MINE against dictionary-based and NER-based approaches, demonstrating significant improvements in phenotype extraction accuracy. Our results establish LLM-MINE as a promising tool for unlocking clinically meaningful phenotypes from unstructured narratives.

Related Work

ADRD Phenotyping Using Structured EHR Data. Recent studies have used phenotypes present within structured EHR data to predict the risk of developing ADRD.[16,17,18] Routinely recorded EHR variables such as demographics, medical diagnoses, vital signs, and medications are commonly used as ADRD phenotypes to build risk prediction models.[9] However, many critical ADRD phenotypes, such as cognitive assessments and clinician impressions are seldom encoded in structured EHR fields.[9] This limitation motivates our focus on unstructured clinical notes as the primary data source for phenotype mining. Other important ADRD prognostic phenotypes that exist in the broader context of EHR include brain imaging data and genetic data. While these modalities offer valuable insights, we restrict our focus to phenotypes documented in clinical notes.

NLP Models for ADRD Phenotype Extraction. Extracting ADRD phenotypes from unstructured clinical notes remains a key challenge in medical informatics. The earliest approaches relied on rule-based NLP pipelines. For instance, Chen et al.[19] developed regular expressions and pattern-matching rules to identify 85 cognitive tests and 11 biomarkers from clinical text. While interpretable, rule-based methods struggle to generalize beyond predefined patterns, often missing nuanced phrasing and complex semantic relationships. To address these limitations, subsequent work adopted deep learning-based models, particularly transformers pre-trained on biomedical corpora. Cheng et al.[20] fine-tuned BioBERT to classify sentences from memory clinic notes across seven cognitive domains (e.g., memory, executive function, language), identifying whether they described impaired or intact symptoms. Most recently, large language models (LLMs) have demonstrated significant advances in understanding complex clinical narratives.[21] For example, Shao et al.[22] used LLMs to extract socioeconomic phenotypes from heart failure patients' notes, finding significant associations with 30-day readmission risks. However, despite these advances, the application of LLMs to ADRD-specific phenotype extraction remains underexplored—a gap this work aims to address.

Method

Preliminaries. We first mathematically define the task of this study. *Definition 1 (Phenotype Extraction from Clinical Notes).* Given a patient's note $\mathbf{T} = \{t_1, \ldots, t_{|\mathbf{T}|}\}$ consisting of a sequence of $|\mathbf{T}|$ tokens, the goal is to utilize a LLM f_θ to extract certain ADRD-related phenotypes $\mathbf{P} = \{p_1, \ldots, p_{|\mathbf{P}|}\}$ that is from a predefined phenotype list. Therefore, the mining process is defined as $\mathbf{P} = f_\theta(\mathbf{T}, \mathbf{I_P})$, where $\mathbf{I_P}$ denotes the phenotype specific mining prompt for LLM.

Data Source and Cohort Construction. This study uses unstructured clinical notes from MIMIC-IV,[15] a large open source healthcare database that contains information for patients admitted to either emergency department or critical care units. To investigate the progression of neurocognitive decline, we construct a cohort reflecting the clinical continuum of ADRD (see Table 1 for details). Following the established phenotype definitions,[23] we stratify patients into three groups based on their diagnostic status: (1) **Cognitively Normal (CN):** Patients with no evidence of cognitive impairment. Following prior work,[16] we define the CN group as patients who were (1) aged above 40 years; (2) had at least one year of medical history in the database; (3) had no ADRD-related diagnoses, determined by excluding patients with any of the ICD codes defined below; and (4) had no exposure to dementia-related medications based on an established list.[24] These criteria yielded an initial CN pool of 91,264 patients. To mitigate class imbalance, we randomly downsampled the CN group at a $2\times$ ratio across three independent draws, resulting in a final CN sample of 11,408 discharge notes. (2) **Mild Cognitive Impairment (MCI):** Patients in the symptomatic pre-dementia stage, characterized by subtle cognitive changes that do not significantly interfere with daily functioning. MCI status was defined using ICD-9 code *331.83* and ICD-10 code *G31.84*. (3) **Alzheimer's Disease and Related Dementias (ADRD):** Patients with major neurocognitive disorder in which symptoms have progressed to impair independent functioning. ADRD status was defined using ICD codes for Alzheimer's disease (*ICD-9: 331.0; ICD-10: G30.0, G30.1, G30.8, G30.9*), Lewy body dementia (*ICD-9: 331.82; ICD-10: G31.83*), frontotemporal dementia (*ICD-9: 331.1, 331.11, 331.19; ICD-10: G31.09*), vascular dementia (*ICD-9: 290.40–290.43; ICD-10: F01.50, F01.51*), and non-specific dementias (*ICD-9: 290.x, 294.x, 797; ICD-10: F02.80, F02.81, F03.90, F03.91*).

Table 1: Characteristics of constructed cohort.

Group	CN	MCI	ADRD
#Patients	8,372	841	8,327
#D/C Notes	11,408	992	13,675
Age			
Mean (SD)	64.0 (12.9)	71.1 (15.2)	80.2 (9.1)
Sex			
Female (%)	4,337 (51.8%)	433 (51.5%)	4,944 (59.4%)
Male (%)	4,035 (48.2%)	408 (48.5%)	3,383 (40.6%)
Ethnicity			
Hispanic (%)	431 (5.2%)	34 (4.0%)	281 (3.4%)
Non-Hisp (%)	7903 (94.4%)	777 (92.4%)	7622 (91.5%)
Unknown Ethnicity (%)	38 (0.5%)	30 (3.6%)	424 (5.1%)
Race			
White (%)	6,003 (71.7%)	625 (74.3%)	6,142 (73.8%)
Black (%)	1,432 (17.1%)	113 (13.4%)	1,036 (12.4%)
Asian (%)	236 (2.8%)	17 (2.0%)	191 (2.3%)
Others (%)	298 (3.6%)	23 (2.7%)	349 (4.2%)
Unknown (%)	38 (0.5%)	30 (3.6%)	424 (5.1%)

Phenotype List. We develop two phenotype lists and compare their results in experiments. Following prior work,[11] we first define *Phenotype List 1*, which includes ten phenotypes organized into six broad categories: Memory Indicators, Comorbidities, Family History, Neurobehavioral Tests/Ratings, Neuroimaging Findings, and Biomarker Test Results. After consultation with ADRD experts, we then develop *Phenotype List 2*, which contains 27 additional phenotypes spanning five categories: Memory, Executive Functions, Language, Visuospatial Skills, and Behavior. By comparing phenotype lists derived from different perspectives, we aim to provide a more comprehensive view of the domain. This strategy helps identify overlapping features while also highlighting differences between structured, data-driven phenotypes and those grounded in clinical expertise.

LLM-based Dementia Phenotype Mining Framework. We propose LLM-MINE (**LLM**-based de**M**ent**I**a phe**N**otyp**E** mining), a modular framework for extracting ADRD-related phenotypes from unstructured clinical notes. Figure 1 illustrates the overall pipeline. We use gemma-3-12b-it,[25] an open-source instruction-tuned LLM whose parameters have been pre-trained on a massive amount of textual corpus including biomedical literature, public clinical databases, and general web data. LLM-MINE directly takes in raw clinical notes without any domain-specific preprocessing, tokenization, or concept normalization. Unlike traditional NLP pipelines that depend on handcrafted rules or tailored vocabulary mappings, LLM-MINE leverages the broad medical knowledge encoded during pre-training to generalize across diverse note formats, abbreviations, and institution-specific clinical language. As the average length of the clinical notes in our dataset often exceeds the model's maximum context window, we partition each note into chunks at sentence boundaries, prompt the model to extract phenotypes from each chunk independently, and take the union of all extracted phenotypes across chunks to produce the final phenotype profile for that note. This chunking strategy ensures that no sentence is truncated mid-span, preserving local clinical context within each extraction call.

We frame phenotype extraction as a user-defined classification-style task rather than open-ended generation. For each phenotype category $\mathbf{P}_i$, the model is instructed to select only from a user-customizable candidate phenotypes defined in the phenotype lists, outputting a comma-separated list of present phenotypes or `'none'` if none are mentioned. This design choice reflects a deliberate trade-off between flexibility and controllability: while open generation would allow the model to capture phenotypic descriptions in its own terms, constraining the output space to a fixed vocabulary

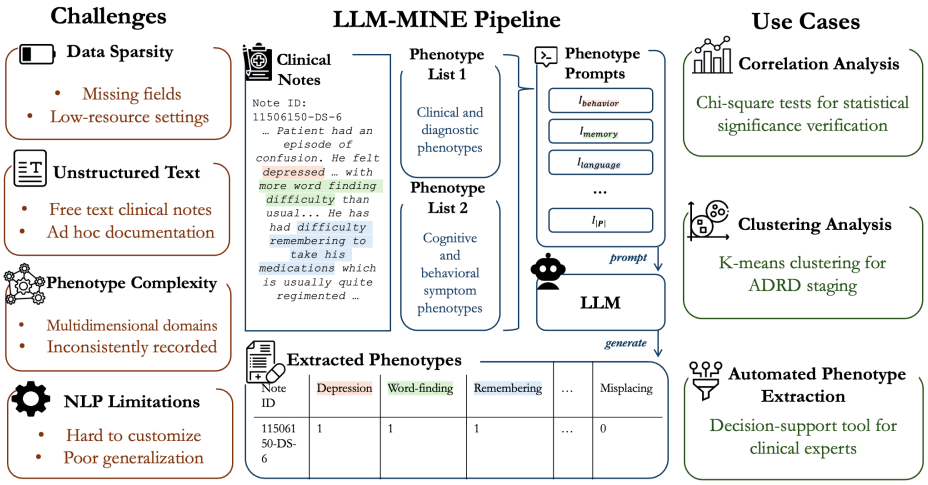


Figure 1: Overview of the LLM-MINE framework. *(Left)* Key challenges motivating the use of LLMs for ADRD phenotype extraction from clinical notes. *(Center)* The LLM-MINE pipeline: unstructured clinical notes are processed through expert-defined phenotype lists using prompting with a LLM to produce binary phenotype feature vectors. *(Right)* Downstream use cases including statistical correlation analysis (chi-square tests), unsupervised disease staging (k-means clustering), and baseline comparison for automated phenotype extraction.

ensures consistent, structured results that can be directly mapped to binary feature vectors without post-processing. This approach also suppresses the model's tendency to produce explanatory reasoning that, while potentially useful for interpretability, introduces parsing complexity and increases the risk of hallucinations. While this comes at the cost of output interpretability, the constrained output form yields more accurate extractions, reduces output token cost, and facilitates downstream aggregation.

LLM-MINE is a training-efficient pipeline and we use both zero-shot[26] and few-shot prompting methods.[27] Zero-shot prompting instructs the model to perform extraction without providing task-specific examples, relying entirely on the model's pre-trained knowledge to interpret clinical language and map it to the target phenotype features. Few-shot prompting is used by appending three examples to each prompt. These demonstrations provide the model with concrete input-output pairs illustrating the expected extraction behavior, including cases where the correct output is `'none'`. This prompting method has been shown to increase task performance that is comparable to fine-tuning without the training costs.[27] Specifically, we use the following prompt templates:

Zero-shot Prompt for Phenotype Extraction

You are analyzing a segment of a clinical nursing note. Extract the patient's $[name(\mathbf{P}_i)]$ from the given discharge note. Please choose from $candidate([\mathbf{P}_i])$. Return only the combination of the above outputs or `'none'` if none are mentioned in the note. ##Note##: $[\mathbf{T}]$.

Few-shot Prompt for Phenotype Extraction

You are analyzing a segment of a clinical nursing note. Extract the patient's $[name(\mathbf{P}_i)]$ from the given discharge note. Please choose from $candidate([\mathbf{P}_i])$. Return only the combination of the above outputs or `'none'` if none are mentioned in the note. Examples: $[(\mathbf{E}_i)]$ ##Note##: $[\mathbf{T}]$.

In the prompt, text in "[]" are placeholders to be filled with specific values. For instance, the filled prompt for comorbidities is "*You are analyzing a segment of a clinical nursing note. Extract the patient's comorbidities of ADRD from*

the given discharge note. Please choose from 'hypertension' and 'depression'. Return only the combination of the above outputs or `'none'` *if none are mentioned in the note.*"

Experimental Results

In this section, we present a series of experiments designed to evaluate the effectiveness and clinical utility of our proposed LLM-MINE. We organize our evaluation around three research questions addressing phenotype extraction quality, downstream discriminative power, and comparison against established baselines.

RQ1: What is the clinical utility for LLM-based automated phenotyping of patients on clinical notes? We begin by examining the accuracy of our LLM-based phenotype extraction approach. We extract phenotypes following the definitions in both phenotype lists from clinical notes. Because the MCI cohort contains only 992 notes, we randomly sample 1,000 notes from each of the CN and ADRD groups to maintain cohort balance. The extraction results, indicated by phenotype presence per cohort, are presented in Table 2 (Phenotype List 1) and Table 3 (Phenotype List 2). For each phenotype, we calculate the number of occurrences; instances containing none of the phenotypes in a given category are marked as "None". From the extraction result, we observe several noteworthy findings. First, most phenotypes **show a monotonic increase in prevalence from the CN group, through MCI, to the ADRD group**, particularly in Phenotype List 2. A small number of categories deviate from this pattern, including memory indicators and family history in Phenotype List 1, and behavior in Phenotype List 2. Phenotype List 2 yields a **higher overall phenotype prevalence than Phenotype List 1**, largely because List 2 targets cognitive-domain symptoms that are near-universal among impaired patients. For instance, in the Memory category, fewer than 4% of MCI and ADRD patients have "None" recorded (MCI: 32/992, ADRD: 17/1,000), whereas in List 1, 27–34% of impaired patients show no memory indicators. This suggests that List 2's cognitive-domain phenotypes are more consistently documented in clinical notes and offer broader coverage of disease-relevant features. Memory symptoms are near-universal among MCI and ADRD patients but substantially less common in CNs. For instance, "Recent Events" memory problems appear in 926 of 992 MCI and 973 of 1,000 ADRD patients, compared with only 648 of 1,000 CNs. Similarly, "Missing Appointments" is documented in 941 MCI and 976 ADRD patients versus 799 CNs. These findings confirm that **memory impairment is the strongest discriminator between cohorts**. Some phenotypes display **clinically interpretable non-monotonic patterns**. "Driving problem" decreases across CN (222), MCI (193), and ADRD (113). Although counterintuitive at first glance, this likely reflects the fact that patients with severe dementia are no longer driving, causing driving-related concerns to disappear from their clinical notes entirely. Additionally, executive function and language problems sharply separate MCI and ADRD from CNs but **offer limited discrimination** between MCI and ADRD. For example, "Judgement" is documented in 609 CNs but jumps to 943 in MCI and 967 in ADRD—a large increase from CN to MCI with negligible further change. This suggests that executive and language features are effective for detecting cognitive impairment but less useful for staging severity.

Table 2: LLM phenotype mining result for Phenotype List 1.

Category	Phenotype	CN	MCI	ADRD
Memory Indicators	None	514	275	336
	Repeating	476	698	652
	Misplacing	311	582	548
Comorbidities	None	39	27	16
	Hypertension	912	896	955
	Depression	665	785	860
Family history	None	490	494	589
	Family history	510	498	411
Neurobehavioral tests	None	986	947	952
	MMSE	13	42	34
	CDR	1	10	7
Neuroimaging findings	None	677	498	399
	Atrophy	238	413	521
	Infarct	303	457	564
Biomarker test results	None	997	980	978
	tau	3	12	22

We further perform chi-square tests of independence[28] to measure the statistical significance of differences between the cohorts, which is presented in Table 4a and Table 4b. For each phenotype, we construct a 2×3 contingency table comparing the presence/absence of the phenotype across the three cohorts. Pairwise 2×2 tests are subsequently conducted between each cohort pair (CN vs. MCI, CN vs. ADRD, *etc*). Statistical significance was defined as $p < 0.05$.

Several key findings emerge from the chi-square analysis. **First, the MCI vs. ADRD comparison is consistently the weakest across both phenotype lists**. In List 2, only Memory reaches statistical significance for this pairwise comparison ($p < 0.05$), while Behavior is essentially non-significant ($p = 0.494$) and Visuospatial Skills ($p = 0.056$) and Language ($p = 0.084$) remain at borderline significance. This confirms that while LLM-extracted phenotypes are effective at detecting cognitive impairment, they have limited power for distinguishing disease stage. **Second, several individual patterns merit attention**. Family History in List 1 exhibits an unusual profile: it is highly significant overall and for CN vs. ADRD and MCI vs. ADRD ($p < 0.001$), yet non-significant for CN vs. MCI ($p = 0.755$). This suggests that family history documentation does not differ between CNs and MCI patients but becomes more thoroughly recorded as disease progresses and patients undergo more comprehensive clinical workups. Comorbidities is the weakest discriminator in List 1, achieving only $p = 0.007$ overall, with a non-significant CN vs. MCI comparison ($p = 0.179$). This is consistent with the high baseline prevalence of conditions such as hypertension and depression across all groups, which diminishes their discriminative utility. By contrast, neuroimaging findings is the only List 1 category that achieves $p < 0.001$ across all comparisons, including MCI vs. ADRD, making it the single strongest discriminator for three-way group separation. **Finally, List 2 phenotypes are more uniformly significant than those in List 1**. All five List 2 categories achieve $p < 0.001$ for the overall, CN vs. MCI, and CN vs. ADRD comparisons, whereas List 1 contains categories with weaker or non-significant pairwise results. This suggests that cognitive-domain phenotypes (memory, executive function, language) are more consistently informative than the clinical and demographic features captured in List 1.

Table 3: LLM phenotype mining result for Phenotype List 2.

Category	Phenotype	CN	MCI	ADRD
Memory	None	189	32	17
	Recent Events	648	926	973
	Remote Events	116	254	460
	Misplacing	654	918	959
	Missing Appointments	799	941	976
Executive Functions	None	233	43	31
	Planning and organization	206	618	611
	Multi-tasking	17	41	22
	Concentration	163	529	388
	Judgement	609	943	967
	Problem-solving	519	908	933
Language	None	679	318	284
	Word-finding	55	330	230
	Slurred speech	77	150	161
	Halting or stuttering	29	87	103
	Impairment	60	259	231
	Abnormal speech	313	664	706
Visuospatial Skills	None	216	41	25
	Finding way around	154	287	280
	Lost in familiar places	82	255	320
	Driving problem	222	193	113
	Seeing things	713	894	912
	Recognizing objects or faces	536	894	957
Behavior	None	134	45	53
	Emotional expression	254	566	607
	Personality or behavior	277	693	742
	Agitation and aggression	107	242	290
	Apathy or decreased motivation	265	636	762
	Hygiene and eating	481	539	585
	Depression or anxiety	606	771	791
	Hallucination	224	582	707
	Weight change	846	920	915

Table 4: Chi-square Test P-values Results. Significance levels: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

(a) Phenotype List 1.

Phenotypes	Overall	CN vs. MCI	CN vs. ADRD	MCI vs. ADRD
Memory Indicators	***	***	***	**
Comorbidities	**	0.179	**	0.117
Family History	***	0.755	***	***
Neurobehavioral tests	***	***	***	0.863
Neuroimaging findings	***	***	***	***
Biomarker test results	***	*	***	0.125

(b) Phenotype List 2.

Phenotypes	Overall	CN vs. MCI	CN vs. ADRD	MCI vs. ADRD
Memory	***	***	***	*
Executive Functions	***	***	***	0.181
Language	***	***	***	0.084
Visuospatial skills	***	***	***	0.056
Behavior	***	***	***	0.494

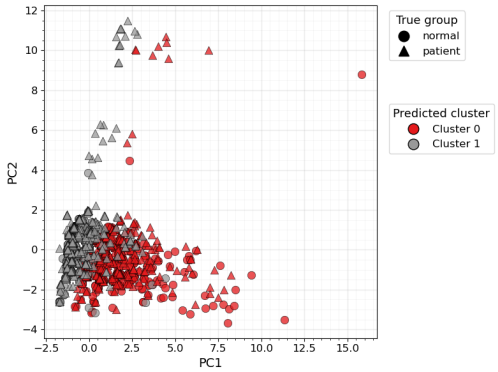
RQ2: What diagnostic discriminative power do the resulting phenotypic profiles hold when used to stratify patients? With the ADRD-centric phenotypes extracted from clinical notes, we perform K-means clustering with $K = 2$ and $K = 3$. Each patient is represented as a feature vector consisting of binary phenotype indicators (1 = present, 0 = not present) across the extracted phenotype categories. We use Euclidean distance as the distance metric. For $K = 2$, we organize the cohort into CN and patient, with the patient group consisting of both MCI and ADRD patients from the original cohorts. We do this because of the earlier observation from the chi-square tests that MCI against ADRD patients in general has higher p-values across phenotypes, suggesting their distributions are not statistically significantly different. Since treating two highly similar groups as distinct clusters may hinder clustering performance, we collapse them into a single patient group and evaluate whether reducing K from 3 to 2 yields better clustering results. For $K = 3$, the group labels remain the same as the original cohort.

By this we explore more fine-grained clustering and see if our extracted phenotype features are distinct enough to correctly cluster the patients. Clustering performance was evaluated by comparing the resulting cluster assignments with cohort labels using external validation metrics, including Adjusted Rand Index (ARI), Normalized Mutual Information (NMI), and Fowlkes–Mallows Index (FMI). The clustering results for both phenotype lists, using both zero-shot and few-shot prompting methods, are presented in Table 5. From the result, we observe a **lower performance across Phenotype List 1 compared to Phenotype List 2**, with the highest performing set being $K = 2$ clustering using features extracted with few-shot prompting. **Few-shot prompting consistently improves performance over zero-shot across both phenotype lists**. For Phenotype List 2, few-shot prompting improves ARI by 46% (0.190 $\rightarrow$ 0.277) and NMI by 88% (0.120 $\rightarrow$ 0.226) under $K = 2$, indicating that more accurate phenotype extraction directly translates to stronger clustering signal. We also observe **consistently higher performance for $K = 2$ over $K = 3$** for Phenotype List 2, consistent with the chi-square finding that MCI and ADRD phenotype distributions are not statistically significantly different, making the two-group partition easier for the clustering algorithm to recover.

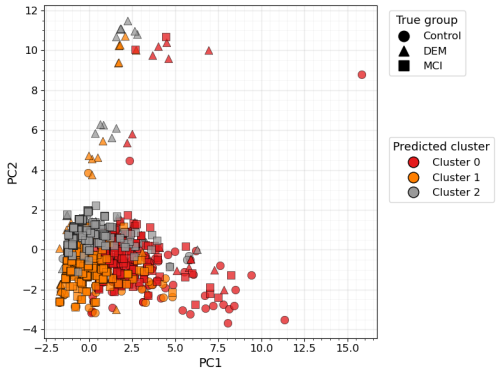
Table 5: Clustering results for compared methods.

Phenotype	Setting	ARI	NMI	FMI
QuickUMLS	n = 2	0.004	0.000	0.657
	n = 3	0.003	0.003	0.378
BioNER	n = 2	0.003	0.006	0.530
	n = 3	0.011	0.011	0.341
Phenotype List 1	n = 2 w/ zero-shot	-0.001	0.000	0.546
	n = 3 w/ zero-shot	0.034	0.032	0.358
	n = 2 w/ few-shot	0.052	0.023	0.568
	n = 3 w/ few-shot	0.030	0.029	0.363
Phenotype List 2	n = 2 w/ zero-shot	0.190	0.120	0.628
	n = 3 w/ zero-shot	0.119	0.105	0.483
	n = 2 w/ few-shot	0.277	0.226	0.660
	n = 3 w/ few-shot	0.172	0.166	0.448
Phenotype List 1 & 2	n = 2 w/ few-shot	**0.290**	**0.232**	**0.666**
	n = 3 w/ few-shot	0.168	0.165	0.446

To further understand the structure of the extracted features, we perform principal component analysis (PCA) and visualize the patient representations in two-dimensional space. The plots are presented in Figure 2 and Figure 3. The PCA plots further verify the clustering performance. We observe that the true groups overlap substantially in 2-dimensional PCA space, suggesting the clusters are not well-separated with compact, spherical geometry in the feature space, which is consistent with the low performance of k-means clustering. Overall, the PCA results confirm that the patient stratification task is inherently difficult: the phenotype distributions do not form compact, well-separated clusters, which limits the performance ceiling of distance-based methods such as K-means.

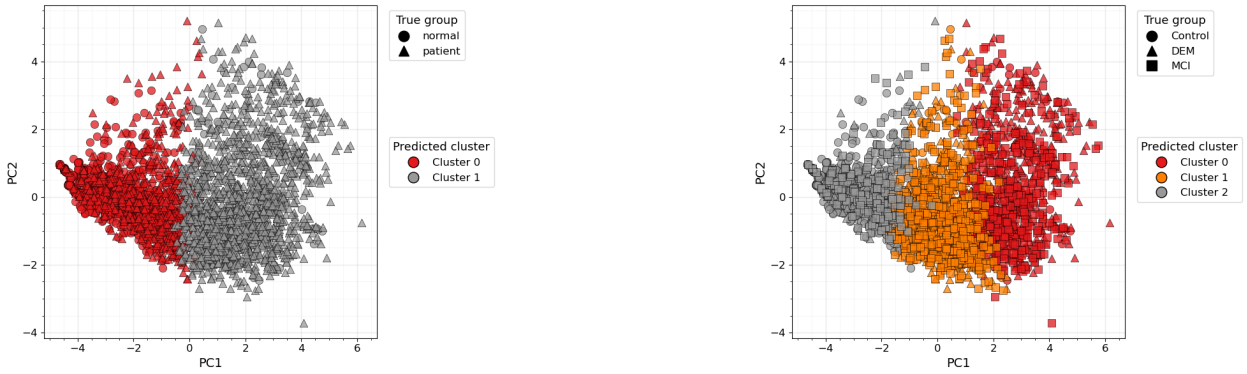


(a) PCA for n=2 clustering.



(b) PCA for n=3 clustering.

Figure 2: PCA for Phenotype List 1. Phenotypes are extracted using few-shot prompting method.



(a) PCA for n=2 clustering. **(b)** PCA for n=3 clustering.

Figure 3: PCA for Phenotype List 2. Phenotypes are extracted using few-shot prompting method.

RQ3: Is our LLM-based phenotype extraction approach better than established approaches? We compare the performance of our proposed method with established methods. One baseline method is to utilize biomedical-ner-all,[29] a transformer model trained on annotated biomedical data, to extract all comorbidities and symptoms as phenotypes. We also use a dictionary based method quickUMLS[30] that matches similar biomedical patterns in the Unified Medical Language System (UMLS) metathesaurus. For biomedical-ner-all, we keep only the concepts with a confidence score greater than or equal to 0.8 to ensure extraction accuracy. For quickUMLS, we restrict the extraction to words longer than 4 characters, and ones that appear in at least 50 notes to filter out rare and noisy concepts. We then perform $K = 2$ and $K = 3$ clustering based on the one hot encoded concepts. The baseline clustering results, alongside our LLM-based extraction clustering results, are presented in Table 5.

The comparison result reflects the advantageous performance of our framework compared to the two baseline methods. Overall performance remains low in absolute terms across all methods, reflecting the difficulty of the stratification task. Among evaluation metrics, ARI and NMI are the most informative, as FMI is heavily influenced by cluster size imbalance and therefore less sensitive to meaningful structural differences between cluster assignments; we report it for completeness but focus our analysis on ARI and NMI. Under the same $K = 2$ setting, our best method, combining Phenotype List 1 and Phenotype List 2 with few-shot prompting, improves ARI from 0.011 to 0.290 ($\Delta = +0.279$), NMI from 0.011 to 0.232 ($\Delta = +0.221$), and FMI from 0.657 to 0.666 ($\Delta = +0.009$). Our results suggest that the current baselines are not competitive in this clinical stratification setting, while our approach captures substantially more structured and disease-relevant information.

We also conduct a ***human evaluation*** to verify the accuracy of the LLM-extracted phenotypes. Three human annotators independently review 20 sampled notes, and the results are shown in Table 6. We measure inter-annotator agreement using Fleiss' kappa[31] for each binary phenotype category across the three annotators. For categories with no label variation, including comorbidities, neurobehavioral tests/ratings, and biomarker test results, Fleiss' kappa is undefined; therefore, we report percent agreement instead. Overall, the LLM-extracted results show high consistency with human annotations for memory indicators, comorbidities, neurobehavioral tests/ratings, and biomarker test results. However, agreement is lower for family history and neuroimaging findings. Error analysis shows that these lower scores are mainly caused by overly broad interpretations by the LLM. For family history, the LLM sometimes extracts family history of any disease, even though the prompt defines this category as ADRD-related family history only. This leads to several false positive cases. Similarly, for neuroimaging findings, the LLM sometimes identifies atrophy or infarct in any organ system, rather than restricting extraction to brain-related atrophy or cerebral infarct.

Table 6: Human evaluation of LLM-extracted phenotypes for Phenotype List 1. Each annotator column reports the consistency between that annotator's labels and the LLM output.

Phenotype	Ann. 1	Ann. 2	Ann. 3	Fleiss' kappa
Memory Indicators	1.00	1.00	1.00	1.00
Comorbidities	1.00	1.00	1.00	1.00[†]
Family History	0.70	0.60	0.60	0.64
Neurobehavioral Tests	0.90	1.00	1.00	0.93[†]
Neuroimaging Findings	0.75	0.60	0.70	0.66
Biomarker Test Results	1.00	1.00	1.00	1.00[†]

[†]Percent agreement is reported because Fleiss' kappa was undefined due to no label variation.

Conclusion

We present LLM-MINE, a modular, user-customizable, and training-free framework that leverages an LLM to extract ADRD-related phenotypes directly from clinical notes without domain-specific preprocessing or model fine-tuning. Through systematic evaluation on a real-world cohort, we verify that the extracted phenotypes exhibit statistically significant differences across the CN, MCI, and ADRD groups, and demonstrate substantial improvements over dictionary-based and biomedical NER baselines in unsupervised disease staging. Our analyses further reveal that cognitive and behavioral symptom phenotypes provide more consistently informative signals than clinical and diagnostic markers, and that distinguishing MCI from ADRD remains a fundamental challenge. These findings highlight the potential practical value of our LLM-based framework for scalable phenotype extraction, early risk stratification, and data-driven dementia research using unstructured clinical documentation. At the same time, the performance of our framework is inherently shaped by the general-purpose nature of the underlying language model. With the recent emergence of domain-adapted medical LLMs, such as Google's MedGemma,[32] our future plan is to evaluate whether models with clinical pretraining yield more precise phenotype extraction.

Limitations. There are several limitations of our work that warrant discussion. First, we used ICD diagnostic codes as a proxy for clinical cognitive status. Because coding practices vary across providers and institutions, and because codes assigned during acute encounters may not reflect a patient's underlying cognitive trajectory, this proxy introduces potential misclassification that could attenuate or distort group-level differences. Second, although LLMs provide a scalable approach for extracting phenotypes from unstructured clinical notes, their performance may be affected by limited interpretability, potential bias, and sensitivity to prompt design as reflected in the human evaluation.

Ethical Considerations. This research involved secondary analysis of a publicly available, de-identified dataset. As the study did not involve interaction with human subjects or access to identifiable private information, it was deemed Not Human Subject Research by the Institutional Review Board (IRB).

Acknowledgments. We thank Dr. Andrew Breithaupt for providing clinical guidance on ADRD phenotype definitions and validation. This research was partially supported by the internal funds and GPU servers provided by School of Nursing's Center for Data Science of Emory University.

References

1. Sperling RA, Aisen PS, Beckett LA, Bennett DA, Craft S, Fagan AM, et al. Toward defining the preclinical stages of Alzheimer's disease: recommendations from the National Institute on Aging–Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimer's & Dementia. 2011;7(3):280-92.
2. Association A. 2025 Alzheimer's disease facts and figures. Alzheimer's & Dementia. 2025;21(5).
3. World Health Organization. Dementia; 2025. Accessed: 2026-01-25. https://www.who.int/en/news-room/fact-sheets/detail/dementia.
4. National Institute of Neurological Disorders and Stroke. Focus on Alzheimer's Disease and Related Dementias; 2025. Accessed: 2026-01-25. Available from: https://www.ninds.nih.gov/current-research/focus-disorders/focus-alzheimers-disease-and-related-dementias.
5. Edition F, et al. Diagnostic and statistical manual of mental disorders. Am Psychiatric Assoc. 2013;21(21):591-643.
6. Grand JH, Caspar S, Macdonald SWS. Clinical features and multidisciplinary approaches to dementia care. Journal of Multidisciplinary Healthcare. 2011;4:125-47.
7. Lyketsos C, Roberts S, Swift EK, Quina A, Moon G, Kremer I, et al. Standardizing electronic health record data on AD/ADRD to accelerate health equity in prevention, detection, and treatment. The journal of prevention of Alzheimer's disease. 2022;9(3):556-60.
8. Xie Y, Han X, Xu R, Hu X, Lu J, Yang C. HypKG: Hypergraph-Based Knowledge Graph Contextualization for Precision Healthcare. In: International Semantic Web Conference. Springer; 2025. p. 328-48.
9. Walling AM, Pevnick J, Bennett AV, Vydiswaran VGV, Ritchie CS. Dementia and electronic health record phenotypes: a scoping review of available phenotypes and opportunities for future research. Journal of the American Medical Informatics Association. 2023;30(7):1333-48.

10. Wang KY, Mahesri M, Novoa-Laurentiev J, Bessette LG, York C, Zakoul H, et al. Extracting Cognitive Impairment Assessment Information From Unstructured Notes in Electronic Health Records Using Natural Language Processing Tools: Validation with Clinical Assessment Data. Clinical Epidemiology. 2025;17:353-65.
11. Oh IY, Schindler SE, Ghoshal N, Lai AM, Payne PRO, Gupta A. Extraction of clinical phenotypes for Alzheimer's disease dementia from clinical notes using natural language processing. JAMIA Open. 2023;6(1):ooad014.
12. Ruckdeschel J, Riley M, Parsatharathy S, Chamarthi R, Rajagopal C, Hsu H, et al. Unstructured Data Are Superior to Structured Data for Eliciting Quantitative Smoking History From the Electronic Health Record. JCO Clinical Cancer Informatics. 2023;7.
13. Huckvale K, Venkatesh S, Christensen H. Toward clinical digital phenotyping: a timely opportunity to consider purpose, quality, and safety. npj Digital Medicine. 2019;2:88.
14. Bilal M, Hamza A, Malik N. NLP for Analyzing Electronic Health Records and Clinical Notes in Cancer Research: A Review. Journal of Pain and Symptom Management. 2025;69(5):e374-94.
15. Johnson AEW, Bulgarelli L, Shen L, Gayles A, Shammout A, Horng S, et al. MIMIC-IV, a freely accessible electronic health record dataset. Scientific Data. 2023;10(1).
16. Li Q, Yang X, Xu J, Guo Y, He X, Hu H, et al. Early prediction of Alzheimer's disease and related dementias using real-world electronic health records. Alzheimer's & Dementia. 2023;19(8):3506-18.
17. Tang AS, Rankin KP, Cerono G, Miramontes S, Mills H, Roger J, et al. Leveraging electronic health records and knowledge networks for Alzheimer's disease prediction and sex-specific biological insights. Nature Aging. 2024;4(3):379-95.
18. Xie Y, Nahab F, Ge Y, Wu Y, Saurman J, Yang C, et al. Abstract wp175: predicting post-stroke cognitive impairment (PSCI) using multiple machine learning approaches. Stroke. 2025;56(Suppl_1):AWP175-5.
19. Chen Z, Zhang H, Yang X, Wu S, He X, Xu J, et al. Assess the documentation of cognitive tests and biomarkers in electronic health records via natural language processing for Alzheimer's disease and related dementias. International journal of medical informatics. 2023;170:104973.
20. Cheng Y, Malekar M, He Y, Bommareddy A, Magdamo C, Singh A, et al. High-Throughput Phenotyping of the Symptoms of Alzheimer Disease and Related Dementias Using Large Language Models: Cross-Sectional Study. JMIR AI. 2025;4:e66926.
21. Xie Y, Lu J, Ho J, Nahab F, Hu X, Yang C. Promptlink: leveraging large language models for cross-source biomedical concept linking. In: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval; 2024. p. 2589-93.
22. Shao M, Xie Y, Yang C, Lu J. LLM-MINE: Large Language Model based Alzheimer's Disease and Related Dementias Phenotypes Mining from Clinical Notes; 2026. Available from: https://arxiv.org/abs/2603.13673.
23. Zhang C, An L, Wulan N, Nguyen KN, Orban C, Chen P, et al. Cross-Dataset Evaluation of Dementia Longitudinal Progression Prediction Models. Human Brain Mapping. 2025;46(11):e70280.
24. Albrecht JS, Hanna M, Randall RL, Kim D, Perfetto EM. An Algorithm to Characterize a Dementia Population by Disease Subtype. Alzheimer Disease and Associated Disorders. 2019;33(2):118-23.
25. Team G, Kamath A, Ferret J, Pathak S, Vieillard N, Merhej R, et al.. Gemma 3 Technical Report; 2025. Available from: https://arxiv.org/abs/2503.19786.
26. Kojima T, Gu SS, Reid M, Matsuo Y, Iwasawa Y. Large Language Models are Zero-Shot Reasoners; 2023. Available from: https://arxiv.org/abs/2205.11916.
27. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al.. Language Models are Few-Shot Learners; 2020. Available from: https://arxiv.org/abs/2005.14165.
28. McHugh ML. The chi-square test of independence. Biochemia Medica (Zagreb). 2013;23(2):143-9.
29. Raza S, Reji DJ, Shajan F, Bashir SR. Large-scale application of named entity recognition to biomedicine and epidemiology. PLOS Digital Health. 2022;1(12):e0000152.
30. Soldaini L. QuickUMLS: a fast, unsupervised approach for medical concept extraction; 2016. Available from: https://api.semanticscholar.org/CorpusID:2990304.
31. Fleiss JL. Measuring nominal scale agreement among many raters. Psychological Bulletin. 1971;76(5):378-82.
32. Sellergren A, Kazemzadeh S, Jaroensri T, Kiraly A, Traverse M, Kohlberger T, et al. MedGemma Technical Report. arXiv preprint arXiv:250705201. 2025.