

**UC Davis**

**UC Davis Electronic Theses and Dissertations**

**Title**

Evaluating Linguistic Pattern Learning in Large Language Models for Dementia Detection

**Permalink**

<https://escholarship.org/uc/item/2vv6z5sk>

**ISBN**

9798189227271

**Author**

Ishihara, Yui

**Publication Date**

2026-06-25

Peer reviewed|Thesis/dissertation

# Evaluating Linguistic Pattern Learning in Large Language Models for Dementia Detection

By

Yui Ishihara  
Thesis

Submitted in partial satisfaction of the requirements for the degree of

Masters of Science

in

Electrical and Computer Engineering

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

---

Chen-Nee Chuah (Chair)

---

Alyssa Weakley

---

Hyoyoung Jeong

Committee in Charge

2026

# ABSTRACT

## Evaluating Linguistic Pattern Learning in Large Language Models for Dementia Detection

Yui Ishihara

Large language models (LLMs) have recently attracted interest for speech-based dementia detection using transcripts from picture description tasks. However, pretrained LLMs are not inherently aligned with specialized clinical tasks, and prior work often evaluates them primarily using accuracy or F1 scores without examining explainability, making it unclear whether performance improvements reflect meaningful clinical features or prompt-specific artifacts. In this thesis, we conduct a systematic study of LLM finetuning for dementia detection, incorporating task-relevant conditioning schemes that introduce linguistic feature guidance (e.g., clinical or linguistic framing) and shared scene context. We find that task-guided finetuning improves accuracy by approximately 20–40 percentage points across evaluated LLM architectures. Our results further show that clinically framed prompting achieves the strongest performance, reaching 86.7% accuracy and 86.6 macro-F1, competitive with prior text and audio-based approaches.

We further conduct attribution analyses across multiple training runs to evaluate explanation reliability. Despite strong predictive performance, attribution patterns exhibit substantial instability, with Jaccard overlap of top-ranked tokens ranging from 0.16–0.24 and Spearman rank correlations of 0.24–0.33 across five finetuning seeds. Aggregation of attribution mass into linguistically interpretable categories further reveals weak alignment with clinically relevant linguistic features. These findings suggest that classification performance alone provides an incomplete picture of model behavior. LLM-based dementia detection tools require more rigorous explainability evaluation to be reliable in real-world deployment.

## ACKNOWLEDGEMENTS

I would like to acknowledge a group of people who were deeply important to the creation of this thesis. Firstly, I would like to thank Professor Chen-Nee Chuah, who has supported not only the research presented in this thesis, but has guided me as a researcher from my time as an undergraduate student. Without her continued mentorship, I would not have had the opportunity to teach undergraduate students, connect with experts in the field, or pursue industry internships as a PhD student—all experiences that have contributed significantly to my growth as an engineer and researcher.

I would also like to sincerely thank Professor Alyssa Weakley and Dr. Michelle Cohn for their continuous support in my exploration of machine learning applications for dementia. I am deeply grateful to them for grounding my research in real-world applications and for allowing me to contribute my engineering skills to a field that feels profoundly meaningful.

I would like to thank my thesis committee— Prof. Chen-Nee Chuah, Prof. Alyssa Weakley, and Prof. Hyoyoung Jeong— for their thoughtful feedback, careful review of this work, and support for this research.

Finally, I would like to thank the members of the RUBINET Lab, as well as my close friends and family, for their continuous support throughout my graduate studies.

# Contents

|                                                       |            |
|-------------------------------------------------------|------------|
| <b>Abstract</b>                                       | <b>ii</b>  |
| <b>Acknowledgements</b>                               | <b>iii</b> |
| <b>List of Figures</b>                                | <b>v</b>   |
| <b>List of Tables</b>                                 | <b>vii</b> |
| <b>1 Introduction</b>                                 | <b>1</b>   |
| <b>2 Background</b>                                   | <b>4</b>   |
| 2.1 Machine Learning for Dementia Detection . . . . . | 4          |
| 2.2 Explainability Methods . . . . .                  | 4          |
| <b>3 Methodology</b>                                  | <b>6</b>   |
| 3.1 LLM Finetuning Strategies . . . . .               | 6          |
| 3.1.1 Linguistic Feature Guided (LFG) . . . . .       | 7          |
| 3.1.2 Visual Context Guided (VCG) . . . . .           | 8          |
| 3.2 Evaluation Protocol . . . . .                     | 8          |
| 3.3 Attribution-Based Model Analysis . . . . .        | 9          |
| 3.3.1 Token Attribution . . . . .                     | 9          |
| 3.3.2 Attribution Stability . . . . .                 | 10         |
| 3.3.3 POS-Aligned Attribution . . . . .               | 10         |
| <b>4 Experiments</b>                                  | <b>11</b>  |
| 4.1 Experimental Setup . . . . .                      | 11         |
| 4.1.1 Datasets . . . . .                              | 11         |
| 4.1.2 Models . . . . .                                | 11         |
| 4.1.3 Baseline . . . . .                              | 12         |
| 4.1.4 Training details . . . . .                      | 12         |
| 4.2 LLM Finetuning Results . . . . .                  | 13         |
| 4.2.1 Main Results . . . . .                          | 13         |
| 4.2.2 Comparison with Prior Work . . . . .            | 14         |
| 4.3 Pairwise Prompt Strategy . . . . .                | 15         |
| 4.4 Ablation Analysis . . . . .                       | 15         |

|          |                                                        |           |
|----------|--------------------------------------------------------|-----------|
| 4.4.1    | Model Size . . . . .                                   | 16        |
| 4.4.2    | Model Architecture . . . . .                           | 16        |
| 4.5      | Explainability and Feature Probing . . . . .           | 16        |
| 4.5.1    | Token Attribution Analysis . . . . .                   | 17        |
| 4.5.2    | Attribution Stability . . . . .                        | 17        |
| 4.5.3    | Linguistic Feature Alignment of Attribution . . . . .  | 19        |
| <b>5</b> | <b>Discussion &amp; Future Work</b>                    | <b>20</b> |
| 5.1      | LLM Adaptation for Dementia Detection . . . . .        | 20        |
| 5.1.1    | Interpretability and Attribution Reliability . . . . . | 20        |
| 5.1.2    | Implications for Clinical NLP . . . . .                | 21        |
| 5.2      | Future Research Directions . . . . .                   | 21        |
| <b>6</b> | <b>Conclusion</b>                                      | <b>22</b> |
| 6.1      | Limitations . . . . .                                  | 22        |
|          | <b>References</b>                                      | <b>24</b> |
| B.1      | Extended Related Works . . . . .                       | 30        |
| B.2      | Implementation and Training Details . . . . .          | 31        |
| B.3      | Linguistic Feature Extraction . . . . .                | 32        |
| B.4      | Prompt Formats . . . . .                               | 34        |
| B.5      | Additional Results . . . . .                           | 35        |
| B.6      | Additional Explainability Results. . . . .             | 39        |
| B.7      | Comparison with Prior Work . . . . .                   | 44        |
|          | <b>Appendices</b>                                      | <b>30</b> |

## LIST OF FIGURES

|     |                                                                        |    |
|-----|------------------------------------------------------------------------|----|
| 3.1 | Finetuning strategies for LLM-based dementia classification. . . . .   | 6  |
| 4.1 | LLaMA model size on classification accuracy. . . . .                   | 15 |
| 4.2 | Comparison with baseline LLM classification. . . . .                   | 16 |
| 4.3 | Top 10 tokens by normalized attribution mass. . . . .                  | 17 |
| 4.4 | Attribution stability across prompts. . . . .                          | 18 |
| 4.5 | Example visualization of patient transcript token heatmap. . . . .     | 18 |
| 4.6 | Token attribution by linguistic categories. . . . .                    | 19 |
| B.1 | Boston Diagnostic Aphasia Examination Cookie Theft stimulus image. . . | 34 |
| B.2 | Accuracy trends across learning rates. . . . .                         | 38 |
| B.3 | Top 10 tokens for AD/NC prediction. . . . .                            | 39 |
| B.4 | Additional token heatmaps for patient transcripts. . . . .             | 43 |

## LIST OF TABLES

|      |                                                                                                                                          |    |
|------|------------------------------------------------------------------------------------------------------------------------------------------|----|
| 3.1  | Representative prompt excerpts used in the visual context guided (VCG) settings. Full prompt templates are provided in App. B.4. . . . . | 7  |
| 4.1  | ADReSS dataset participant demographics. . . . .                                                                                         | 12 |
| 4.2  | Main results: performance across finetuning and conditioning strategies. . .                                                             | 13 |
| 4.3  | Comparison with prior work. . . . .                                                                                                      | 14 |
| 4.4  | Pairwise comparison of VCG prompt strategies. . . . .                                                                                    | 15 |
| B.1  | Linguistic feature extraction setup. . . . .                                                                                             | 33 |
| B.2  | Numeric-to-text conversion for linguistic features. . . . .                                                                              | 33 |
| B.3  | VLM caption generation prompt. . . . .                                                                                                   | 34 |
| B.4  | Zero-shot performance of tested LLMs on dementia classification. . . . .                                                                 | 36 |
| B.5  | Performance on additional dataset. . . . .                                                                                               | 36 |
| B.6  | Performance on captions generated by varied VLMs. . . . .                                                                                | 37 |
| B.7  | Attribution stability across prompts and methods. . . . .                                                                                | 40 |
| B.8  | Short vs. long prompt comparisons within each prompt setting. . . . .                                                                    | 41 |
| B.9  | Pairwise prompt comparisons across short and long prompting settings. . .                                                                | 42 |
| B.10 | Mapping from POS-aligned attribution categories to expert-informed proxy features. . . . .                                               | 42 |
| B.11 | Quantitative alignment of attribution mass with clinically motivated proxy features. . . . .                                             | 42 |
| B.12 | Full comparison with prior works. . . . .                                                                                                | 45 |
| B.13 | Comparison of short and long prompt formulations. . . . .                                                                                | 46 |

## Introduction

Alzheimer’s disease (AD), affecting more than 57 million people worldwide (1), often results in subtle changes in speech, including difficulty retrieving words, reduced lexical diversity, and changes in narrative coherence (2; 3; 4; 5; 6). These linguistic challenges have motivated computational approaches for speech-based dementia detection. Among the elicitation tasks studied in clinical linguistics, we focus on the picture-description task (PDT), where participants describe everything they observe in an image (7; 8). This task requires participants to produce short verbal narratives grounded in a shared visual scene (e.g., people having a picnic) and forms the basis of widely used benchmarks such as the ADReSS dataset (9) for studying transcript-based dementia detection. A widely used example is the Cookie Theft task, in which participants describe a kitchen scene containing multiple interacting objects and events. While transcript-based approaches do not capture acoustic features such as prosody and pause structure, the data is easier to collect, share, and standardize across clinical settings. They also enable direct analysis of linguistic features, which are difficult to extract from acoustic data alone.

Earlier work on speech-based dementia detection primarily relied on supervised models trained on handcrafted linguistic and acoustic features associated with cognitive decline (10; 11). More recently, large language models (LLMs) have been applied directly to transcripts of these recordings due to their strong language understanding capabilities, with several studies reporting accuracies ranging from approximately 84–92% on the ADReSS dataset (12; 13; 14). However, evaluation in this setting remains challenging. Most studies rely on small PDT datasets and report results from a single random seed, typically measured only by accuracy or F1 score. In low-data settings, these evaluation practices make it difficult to determine whether models learn clinically meaningful linguistic patterns or simply exploit dataset artifacts or prompt-specific cues. Moreover, existing work rarely examines whether LLM predictions are

grounded in clinically meaningful linguistic features from the transcripts, making it difficult to understand what information drives model decisions. Such analysis is important for understanding model behavior and supporting trust in clinical settings.

In this work, we conduct a multifaceted study of LLM finetuning for transcript-based dementia detection, examining how different conditioning strategies and contextual cues affect model performance in this task. Our study design is motivated by both the structure of the PDT and findings from clinical linguistics. Clinical studies link changes in features such as filler usage, sentence length, and pronoun use to cognitive decline, motivating the incorporation of linguistically relevant features to guide model learning. Additionally, since the PDT requires describing a shared visual scene, incorporating this context may help ground transcript understanding. We design task-guided conditioning strategies that integrate transcript information with clinically motivated linguistic features and shared scene context, allowing us to examine how these signals affect transcript-based dementia detection. To obtain a more reliable picture of model behavior in this low-data setting, we also emphasize rigorous evaluation practices, including reporting results across multiple training runs, clinically-motivated metrics, and controlled ablations across model architectures and scales.

Beyond performance metrics, it is equally important to understand how these models arrive at their decisions, particularly in clinical settings where reliability and transparency are essential. Because pretrained LLMs operate largely as black-box systems (15; 16), reporting accuracy alone provides limited insight into their predictions. To take a step toward addressing this gap, we analyze token-level attribution patterns across multiple training runs to examine whether model explanations align with interpretable linguistic structures and whether these patterns remain stable across runs. This analysis provides a lens for assessing whether improved classification performance reflects meaningful linguistic reasoning or unstable model behavior.

Our results show that incorporating linguistic and contextual grounding improves LLM performance for dementia detection compared to cognitively normal (CN) older adults. Clinically framed prompts and task-specific context from the Cookie-Theft PDT yield the

strongest results, with the best model achieving 86.7% accuracy and 86.6% F1. However, deeper analysis reveals instability in the linguistic signals underlying model predictions across training runs. These findings suggest that accuracy alone provides an incomplete picture of model behavior in low-data clinical settings and highlight the need for more rigorous evaluation.

To summarize, our main contributions are as follows: **(1)** We present a systematic evaluation of LLM conditioning strategies for transcript-based dementia detection on the ADReSS dataset, comparing transcript-only inputs, linguistic feature augmentation, and task-oriented contextual grounding. **(2)** We show that incorporating contextual information from the PDT improves performance, achieving results comparable to current state-of-the-art approaches. **(3)** Using attribution-based probing, we analyze the linguistic signals underlying LLM predictions and reveal instability across training runs. **(4)** We adopt more rigorous evaluation practices for LLM-based dementia detection through multi-seed experiments, mean and standard deviation reporting, and systematic ablations.

## Background

### 2.1 Machine Learning for Dementia Detection

Early works on dementia detection have used acoustic signal processing (2; 17; 18) and natural language processing (NLP)-based linguistic feature extraction (19; 8) to train traditional ML classifiers or shallow neural networks (17; 20; 18). More recent work has shifted toward pretrained speech and language models, which are used either for feature extraction in downstream classifiers or directly finetuned for dementia classification (21; 22). LLMs extend this through prompting, finetuning, and formulations based on paired perplexity differences rather than direct classification (14). (12) explores zero-shot prompting strategies for this task, though such methods may rely on spurious correlations rather than dementia-relevant linguistic markers. (13) apply chain-of-thought (CoT) prompting to generate diagnostic rationales. However, like most prior LLM-based work, it does not evaluate model explainability or examine which linguistic cues drive predictions.

### 2.2 Explainability Methods

Earlier works focused on expert-defined linguistic features, which can provide some inherent explainability (23; 3). However, many approaches also incorporate automatically extracted feature representations that are not directly expert-defined (21; 24). Recent work has instead focused on handcrafted or LLM-derived features (25; 26), or on analyzing task behavior through visualization and image-text alignment (27; 28). While informative, these approaches do not examine how LLMs internally encode or prioritize expert-defined cues. For recent language-based LLM works (13; 14), explainability is rarely considered. An extended related works is in Appendix (App.) B.1.

Our work is a study of finetuning LLMs for dementia detection across prompting

strategies and task-specific context. To our knowledge, we are the first to consider such settings and to explore how model predictions relate to linguistically and clinically relevant cues with token-level attributions.

## Methodology

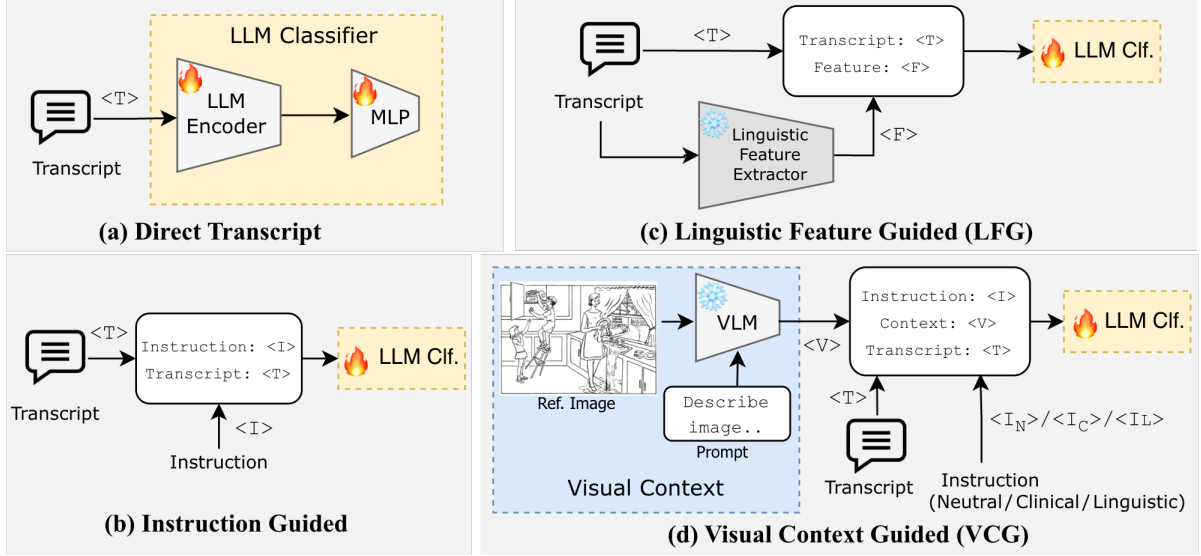


Figure 3.1: Our finetuning strategies for LLM-based dementia classification. LFG augments transcripts with extracted numeric features or discretized linguistic markers (e.g., “used some filler words”; see App. B.3). VCG uses three instructions with visual context: transcript-only, neutral, and clinically framed instruction (see Table 3.1).

We investigate finetuning strategies with progressively richer conditioning, inspired by linguistic indicators of cognitive decline and the visual grounding inherent in picture-description tasks (Sec. 3.1). We then outline evaluation protocols emphasizing robust testing and statistical significance (Sec. 3.2). Finally, we analyze token-level attribution to assess whether model predictions align with clinically meaningful linguistic cues (Sec. 4.5.3).

### 3.1 LLM Finetuning Strategies

We study finetuning strategies with progressively richer task conditioning for dementia detection. Our baseline finetunes a pretrained LLM directly on patient transcripts from the picture-description task. We then add task guidance through instruction-based finetuning, where the transcript is paired with an explicit instruction to classify NC or AD.

Table 3.1: Representative prompt excerpts used in the visual context guided (VCG) settings. Full prompt templates are provided in App. B.4.

| Inst.                 | Task Framing                                                                                                                                                                                                 | Example                                                                                                                                                          |
|-----------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| $\langle I \rangle$   | Instruction to classify into AD or NC without auxiliary context or reasoning guidance.                                                                                                                       | "Classify the patient as AD or NC."                                                                                                                              |
| $\langle I_N \rangle$ | Neutral task context. Provides background on the Cookie Theft picture-description task but does not emphasize specific indicators.                                                                           | "The Cookie Theft task is commonly used in dementia detection, where participants describe everything happening in a picture."                                   |
| $\langle I_L \rangle$ | Linguistically framed reasoning. Emphasizes expert-defined speech features including disfluencies, syntactic complexity, lexical frequency, type-token ratio, discourse structure, and lexical concreteness. | "Focus on linguistic features commonly observed in dementia detection, including disfluencies, syntactic complexity, lexical frequency, and discourse features." |
| $\langle I_C \rangle$ | Clinically guided feature prompting. Highlights expert-defined indicators including word-finding difficulty, repetition, sentence length, topic repetition, idea density, and topic completeness.            | "Focus on clinical features commonly observed in dementia detection, including word-finding difficulties, repetition, idea density, and topic completeness."     |

Finally, we introduce two conditioning strategies: Linguistic Feature Guided (LFG), which incorporates clinically motivated linguistic markers, and Visual Context Guided (VCG), which provides scene grounding through textual descriptions of the image being narrated. All methods use an LLM encoder followed by a single-layer multilayer perceptron (MLP) classifier. Fig. 3.1 outlines our strategies.

### 3.1.1 Linguistic Feature Guided (LFG)

Prior work on dementia detection has shown that expert-derived linguistic features (e.g., lexical diversity, fluency disruptions, and syntactic complexity) can indicate cognitive decline (2; 3; 4; 5; 6; 29). Before the adoption of LLM-based models, many approaches relied on classifiers trained directly on such extracted features. We therefore investigate whether incorporating these clinically motivated signals can improve LLM finetuning beyond transcript-only inputs.

We evaluate two feature representations. In *LFG: Numeric*, extracted linguistic features are appended to the model input as scalar values. In *LFG: Language*, normalized ( $z$ -scored) features are converted into short natural-language summaries (e.g., "the speaker used moderately varied vocabulary"). This formulation tests whether LLMs benefit more from features expressed in language form, which are potentially better aligned with pre-

trained language representations, than from direct numerical inputs. More details are in App. B.3.

### 3.1.2 Visual Context Guided (VCG)

Picture-description tasks evaluate how well a speaker’s narrative aligns with the visual scene being described. To approximate this setting, we provide the model with a textual description of the stimulus image generated by a vision-language model (VLM). This scene description is appended to the patient transcript, allowing the model to reason about the narrative relative to the underlying visual context. We further examine how explicit task guidance influences model behavior through prompt instructions. Prompts were designed with input from domain expert clinicians, including a neuropsychologist specializing in dementia assessment and a linguistics researcher studying language impairment in AD. To control for prompt length effects, all prompts were constrained to 50 tokens.

We evaluate four configurations corresponding to the instruction variants summarized in Table 3.1. In (1) *VCG: Transcript*, the scene description is appended without additional task guidance. In (2) *VCG: Neutral* ( $\langle I_N \rangle$ ), the prompt provides general context about the picture-description task. In (3) *VCG: Linguistic* ( $\langle I_L \rangle$ ), the prompt emphasizes linguistically grounded indicators of dementia-related speech patterns. In (4) *VCG: Clinical* ( $\langle I_C \rangle$ ), the prompt highlights clinically motivated markers of cognitive impairment.

## 3.2 Evaluation Protocol

Evaluation practices for dementia detection from speech often rely on single-run accuracy or F1 scores (13; 12), which can obscure model variability in small clinical datasets. We therefore adopt a more rigorous evaluation protocol emphasizing clinically relevant metrics and statistical robustness. In addition to accuracy and F1, we report sensitivity, specificity, and area under the ROC curve (AUC), commonly used in clinical evaluations (15; 30). Sensitivity measures the ability to correctly detect dementia cases (true positive rate), while specificity measures the ability to avoid false diagnoses among healthy individuals, both important metrics for clinicians. Because dementia speech datasets are

typically small and performance can vary across random initializations, we report mean and standard deviation across multiple runs. To ensure reliable comparisons between prompting and feature-guided strategies, we perform paired bootstrap resampling over test examples to estimate 95% confidence intervals for metric differences. Finally, we complement quantitative results with token-level attribution analyses (Sec. 4.5.3) to examine whether model predictions align with clinically meaningful linguistic cues.

### 3.3 Attribution-Based Model Analysis

High predictive accuracy does not necessarily imply that models rely on clinically meaningful signals. We therefore use attribution-based probing to estimate the influence of individual tokens on model predictions, as attribution methods provide a natural way to analyze token-level contributions in language models. This allows us to examine whether explanations are stable across training runs and aligned with linguistically meaningful structures in patient speech. To the best of our knowledge, such systematic evaluation of attribution stability and linguistic alignment has not been explored in prior dementia language analysis.

#### 3.3.1 Token Attribution

We analyze token-level model behavior using gradient-based attribution methods, including Integrated Gradients (IG) (31), DeepLIFT (DL) (32), and DeepLIFT-SHAP (DLS) (33). Using multiple attribution techniques allows us to assess whether observed attribution patterns are consistent across methods. Attribution scores are computed with respect to the output logit and summed across embedding dimensions to obtain a scalar importance value for each subword token. Scores are normalized within each transcript and aggregated across correctly classified samples for each class, allowing inspection of which tokens most strongly influence predictions.

### 3.3.2 Attribution Stability

While attribution methods highlight influential tokens, their reliability across training runs remains unclear. To evaluate explanation stability, we compare influential tokens across models trained with different random seeds. For each run, we extract the top- $k$  tokens ranked by attribution score and compare them using two metrics. Jaccard similarity (34) measures overlap between top- $k$  token sets, while Spearman rank correlation ( $\rho$ ) (35) evaluates consistency of attribution score rankings across runs. This analysis provides a quantitative measure of explanation robustness.

### 3.3.3 POS-Aligned Attribution

To examine whether attribution patterns correspond to meaningful features, we align token-level attribution scores with part-of-speech (POS) annotations from CHAT transcripts (36), a widely used transcription format in clinical language datasets such as DementiaBank. Subword tokens are mapped to CHAT word units using character-span alignment, and attribution mass is aggregated across all subwords belonging to each token. Tokens are then grouped into linguistic categories (e.g., nouns, verbs, pronouns), and attribution mass is pooled by category and class. This analysis allows us to assess whether influential tokens correspond to linguistically meaningful POS patterns in patient speech, providing an interpretable view of model reasoning.

The experiments are organized into two parts. Sections 4.2–4.4 evaluate predictive performance under different finetuning and conditioning strategies. Section 4.5 then applies the described attribution framework to investigate whether these models rely on meaningful linguistic signals and whether explanations remain stable across training runs.

## Experiments

### 4.1 Experimental Setup

#### 4.1.1 Datasets

Following prior work (13; 12), our main evaluations use the ADReSS IS-2020 dataset (9), which contains English-speaking individuals with AD and CN older adults. Participants in ADReSS were drawn from the Pitt Corpus of DementiaBank (37). In the original Pitt cohort, individuals with AD were older ( $71.4 \pm 8.3$  years vs.  $63.8 \pm 8.3$  years), had fewer years of education ( $12.1 \pm 2.9$  vs.  $14.3 \pm 2.9$ ), and were more frequently female than cognitively normal controls. Screening criteria included a minimum age of 45 years of age alongside the ability to read English and excluded individuals with major neurological disorders (e.g., stroke, epilepsy, traumatic brain injury), major psychiatric illness, substance abuse, or other conditions that could affect dementia diagnosis. The ADReSS challenge subsequently selected a balanced subset of these recordings to mitigate age and gender imbalance. The dataset includes 156 samples (78 AD, 78 CN) with a fixed 108/48 training/test split (detailed information on the dataset can be found in Table 4.1). Transcripts of these of patients participating in the Cookie Theft PDT were human-annotated using the CHAT coding system (36). Additional experiments on the Lu corpus (38) are reported in Table B.5.

#### 4.1.2 Models

Following prior LLM-based dementia studies (13; 12), we use LLAMA-3 (39) for comparison. We also include several additional LLM families for architecture ablation. For VCG experiments, visual context is obtained using LLAVA-1.5-7B (40). Experiments with additional captioning VLMs are in Table B.6.

Table 4.1: Participant demographics for the ADReSS dataset training set (top) and test set (bottom), separated by age range and diagnosis group. MMSE values are reported as mean (standard deviation).

| Age     | AD |    |            | non-AD |    |            |
|---------|----|----|------------|--------|----|------------|
|         | M  | F  | MMSE (sd)  | M      | F  | MMSE (sd)  |
| [50,55) | 1  | 0  | 30.0 (n/a) | 1      | 0  | 29.0 (n/a) |
| [55,60) | 5  | 4  | 16.3 (4.9) | 5      | 4  | 29.0 (1.3) |
| [60,65) | 3  | 6  | 18.3 (6.1) | 3      | 6  | 29.3 (1.3) |
| [65,70) | 6  | 10 | 16.9 (5.8) | 6      | 10 | 29.1 (0.9) |
| [70,75) | 6  | 8  | 15.8 (4.5) | 6      | 8  | 29.1 (0.8) |
| [75,80) | 3  | 2  | 17.2 (5.4) | 3      | 2  | 28.8 (0.4) |
| Total   | 24 | 30 | 17.0 (5.5) | 24     | 30 | 29.1 (1.0) |

| Age     | AD |    |            | non-AD |    |            |
|---------|----|----|------------|--------|----|------------|
|         | M  | F  | MMSE (sd)  | M      | F  | MMSE (sd)  |
| [50,55) | 1  | 0  | 23.0 (n/a) | 1      | 0  | 28.0 (n/a) |
| [55,60) | 2  | 2  | 18.7 (1.0) | 2      | 2  | 28.5 (1.2) |
| [60,65) | 1  | 3  | 14.7 (3.7) | 1      | 3  | 28.7 (0.9) |
| [65,70) | 3  | 4  | 23.2 (4.0) | 3      | 4  | 29.4 (0.7) |
| [70,75) | 3  | 3  | 17.3 (6.9) | 3      | 3  | 28.0 (2.4) |
| [75,80) | 1  | 1  | 21.5 (6.3) | 1      | 1  | 30.0 (0.0) |
| Total   | 11 | 13 | 19.5 (5.3) | 11     | 13 | 28.8 (1.5) |

### 4.1.3 Baseline

We evaluate zero-shot inference as the performance baseline. All LLMs use the same prompt templates as in finetuning. Predictions are deterministic under fixed decoding settings.

### 4.1.4 Training details

Models are trained on stratified 80:20 train:val splits, with hold-out test set for evaluation. Models are trained on stratified 80:20 train:validation splits, with the official hold-out test set used for evaluation. Inputs are tokenized using the LLAMA tokenizer with a maximum sequence length of 512 tokens. Across all 156 transcripts, sequence lengths ranged from 178 to 954 tokens (mean = 319.5, median = 297), with only 8 samples exceeding the length limit and therefore subject to truncation. Finetuning uses cross-entropy loss with LoRA. Training runs for up to 15 epochs with early stopping, and results are reported as

the mean and std. dev. over five fixed random seeds. More details in Table. B.7.

|                         | Acc. (%)                | F1 (%)                  | Prec. (%)               | Sens. (%)               | Spec. (%)               | AUC                    |
|-------------------------|-------------------------|-------------------------|-------------------------|-------------------------|-------------------------|------------------------|
| Zero-Shot               | 54.17                   | 63.33                   | 52.78                   | 79.17                   | 29.17                   | 0.45                   |
| Transcript Only         | 77.92 $\pm$ 3.78        | 75.86 $\pm$ 7.32        | 82.56 $\pm$ 5.36        | 72.50 $\pm$ 16.82       | 83.33 $\pm$ 9.77        | 0.89 $\pm$ 0.02        |
| Instr. Guided           | 76.67 $\pm$ 6.49        | 73.57 $\pm$ 10.04       | 82.28 $\pm$ 3.04        | 68.33 $\pm$ 17.82       | 85.00 $\pm$ 5.59        | 0.87 $\pm$ 0.02        |
| <i>LFG</i> : Numeric    | 82.08 $\pm$ 7.88        | 80.35 $\pm$ 13.18       | 83.88 $\pm$ 4.83        | 80.83 $\pm$ 21.97       | 83.33 $\pm$ 7.80        | 0.89 $\pm$ 0.01        |
| <i>LFG</i> : Language   | 85.00 $\pm$ 0.93        | 85.69 $\pm$ 1.32        | 81.85 $\pm$ 0.83        | <b>90.00</b> $\pm$ 3.73 | 80.00 $\pm$ 1.86        | 0.88 $\pm$ 0.01        |
| <i>VCG</i> : Transcript | 84.58 $\pm$ 4.32        | 83.78 $\pm$ 6.06        | <b>87.04</b> $\pm$ 3.34 | 81.67 $\pm$ 11.64       | 87.50 $\pm$ 5.10        | 0.89 $\pm$ 0.02        |
| <i>VCG</i> : Neutral    | 84.58 $\pm$ 2.38        | 84.79 $\pm$ 2.10        | 84.02 $\pm$ 4.72        | 85.83 $\pm$ 3.73        | 83.33 $\pm$ 5.89        | 0.88 $\pm$ 0.01        |
| <i>VCG</i> : Clinical   | <b>86.67</b> $\pm$ 2.80 | <b>86.58</b> $\pm$ 3.21 | 86.86 $\pm$ 3.35        | 86.67 $\pm$ 6.85        | <b>86.67</b> $\pm$ 4.56 | <b>0.90</b> $\pm$ 0.01 |
| <i>VCG</i> : Linguistic | 81.67 $\pm$ 5.39        | 81.47 $\pm$ 6.22        | 81.66 $\pm$ 4.10        | 81.67 $\pm$ 9.59        | 81.67 $\pm$ 4.75        | 0.89 $\pm$ 0.01        |

Table 4.2: Performance across finetuning and conditioning strategies for LLaMA-3.2-1B-Instruct. Results are reported as mean  $\pm$  std. dev. over five random seeds. Full results in Appendix B.7

## 4.2 LLM Finetuning Results

### 4.2.1 Main Results

Table 4.2 reports the main results (mean  $\pm$  std. dev. across seeds). Statistical reliability is analyzed further in Section 4.3. We highlight three empirical observations.

*Observation 1: Task finetuning is necessary* Zero-shot inference performs poorly ( $\sim$ 54% acc.), whereas finetuning on transcripts improves performance substantially ( $\sim$ 78% acc.).

*Observation 2: Linguistic feature guidance improves performance* Augmenting transcripts with linguistic statistics improves performance over transcript-only finetuning ( $\sim$ 82% acc. for *LFG: Numeric*). Expressing the same features as short natural-language summaries yields stronger results ( $\sim$ 85% acc. for *LFG: Language*). We hypothesize that linguistic cues expressed in natural language may align more naturally with pretrained LLM representations than equivalent numeric features.

*Observation 3: Contextual grounding with clinically framed instructions yields the strongest performance* Providing models with contextual grounding through scene descriptions of the picture being narrated further improves performance over transcript-only finetuning. Even without additional task guidance, appending scene descriptions alone (*VCG: Transcript*) improves accuracy to 84.58%. The strongest performance is obtained when contextual grounding is combined with clinically framed instructions (*VCG: Clinical*),

achieving 86.67% accuracy and 0.90 AUC. As summarized in Table 3.1, the VCG conditions differ only in the task guidance provided, with the clinical variant explicitly directing attention toward indicators of impairment such as word-finding difficulty, repetition, idea density, and topic completeness. Similar trends are observed across model sizes (Fig. 4.1).

Overall, these results indicate that while LLMs can achieve strong performance on dementia detection, performance depends on how task context and linguistic cues are incorporated during finetuning.

| Method                       | Audio    | Text     | Acc.        | F1          |
|------------------------------|----------|----------|-------------|-------------|
| Luz et al. (9)               | ✓        | ✓        | 76.8        | 74.5        |
| Syed et al. (41)             | ✓        | ✓        | 81.3        | –           |
| Chen et al. (42)             | ✓        | ✗        | 73.7        | 72.8        |
| Chen et al. (42)             | ✓        | ✗        | 74.2        | 73.6        |
| Cummins et al. (24)          | ✗        | ✓        | 81.3        | 81.2        |
| Zhu et al. (43)              | ✓        | ✓        | 82.9        | 82.9        |
| Zhu et al. (12)*             | ✗        | ✓        | 83.3        | –           |
| Botelho et al. (44)*         | ✗        | ✓        | 81.3        | –           |
| Xiao et al. (14)*            | ✗        | ✓        | <b>91.7</b> | –           |
| Park et al. (13)*            | ✓        | ✓        | 83.3        | 83.2        |
| Park et al. (13)*            | ✗        | ✓        | 87.5        | <b>87.5</b> |
| <b>VCG: Clinical (Ours)*</b> | <b>✗</b> | <b>✓</b> | <b>86.7</b> | <b>86.6</b> |

Table 4.3: Comparison with prior work on the ADReSS dataset. Full table in App. B.7.

\*LLaMA-1B-Inst. model.

## 4.2.2 Comparison with Prior Work

Table B.12 compares our best configuration (*VCG: Clinical*) with prior methods. Our method achieves competitive performance (86.7% accuracy, 86.6% F1) relative to recent transcript-based LLM models, while outperforming several classical text-only and hybrid audio-text baselines. Direct comparisons across studies should be interpreted with caution, as reporting practices differ across prior work. Many studies report only accuracy or single-seed results, whereas our evaluation reports mean and standard deviation across multiple seeds. Prior work also often does not report metrics such as AUC, sensitivity, and specificity. While paired-perplexity models such as (14) report higher accuracy, our approach achieves comparable performance using a simpler finetuning method based on contextual grounding and linguistic guidance. Our study also examines how these conditioning strategies influence model behavior through additional probing and attribution analyses (Sec. 4.5).

### 4.3 Pairwise Prompt Strategy

We analyze how different prompts influence LLM finetuning performance. We consider our prompt strategies: clinical instructions (C), linguistic feature guidance (L), and neutral prompts (N). To reduce variance, each configuration is trained using five random seeds. We then compute pairwise differences in balanced accuracy between prompt strategies and estimate uncertainty using bootstrap confidence intervals. As shown in Table 4.4, clinical prompts consistently outperform both neutral and linguistic prompts, while linguistic prompts show mixed performance relative to neutral prompts. These results reinforce that explicitly incorporating clinical task framing provides stronger supervision.

| Comparison Prompts  | Mean $\Delta$        | Bootstrap 95% CI                  | W/T/L        |
|---------------------|----------------------|-----------------------------------|--------------|
| Clinical–Neutral    | 0.0208 $\pm$ 0.0255  | [ <b>0.0000</b> , <b>0.0375</b> ] | <b>4/1/0</b> |
| Linguistic–Neutral  | -0.0292 $\pm$ 0.0480 | [-0.0667, 0.0083]                 | 2/0/3        |
| Clinical–Lingusitic | 0.0500 $\pm$ 0.0432  | [ <b>0.0167</b> , <b>0.0833</b> ] | <b>4/1/0</b> |

Table 4.4: Pairwise comparison of prompt strategies used during finetuning: clinical (C), linguistic (L), and neutral (N). We report the mean difference ( $\Delta$ ) in balanced accuracy across five seeds ( $\pm$  SD), bootstrap 95% confidence intervals, and win/tie/loss counts. Positive  $\Delta$  favors the first prompt.

### 4.4 Ablation Analysis

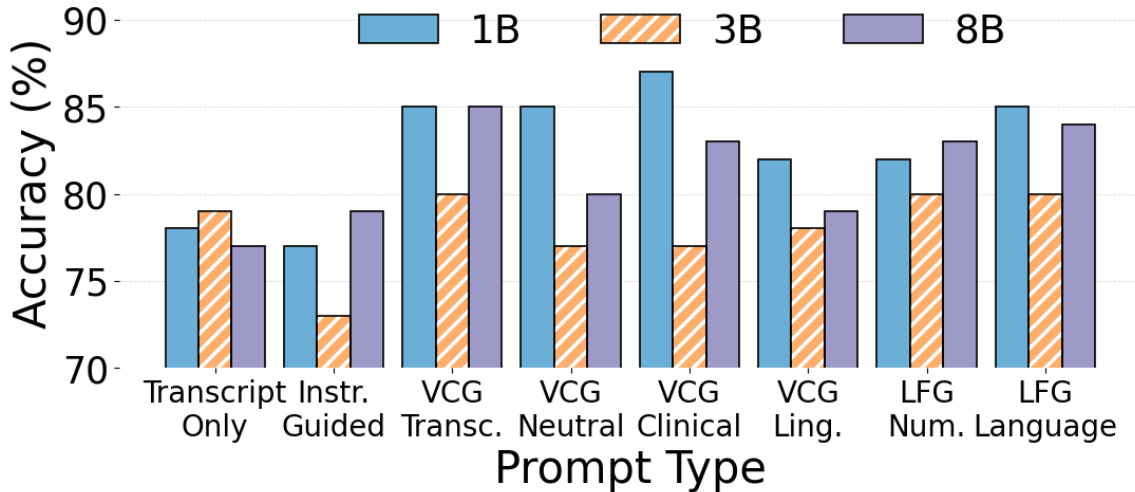


Figure 4.1: Accuracy across prompting strategies for LLaMA models of different sizes (1B, 3B, 8B)

#### 4.4.1 Model Size

We evaluate model scaling using LLaMA models of sizes 1B, 3B, and 8B under different prompt strategies (Fig. 4.1). The 1B model shows the consistent high performance across prompt variants. Increasing model size does not consistently improve accuracy, suggesting that additional parameters do not necessarily translate to better adaptation for small clinical datasets.

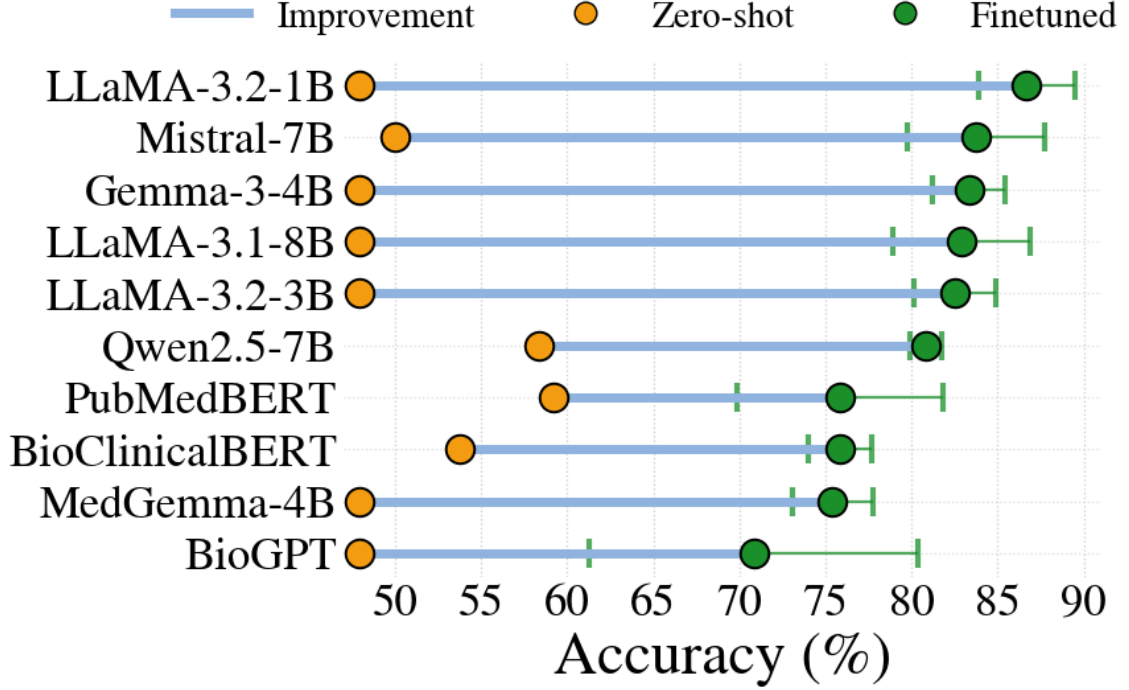


Figure 4.2: Accuracy of LLM architectures under zero-shot prompting and VCG:Clinical finetuning

#### 4.4.2 Model Architecture

We further evaluate whether the observed trends generalize across different LLM architectures (App. B.5). While performance varies across models, the relative benefits of contextual grounding and instruction framing remain consistent across architectures.

### 4.5 Explainability and Feature Probing

High accuracy may not necessarily imply that models rely on clinically meaningful signals. We therefore analyze token-level attribution patterns to examine (1) which tokens influence model predictions, (2) whether attribution explanations are stable across training

runs, and (3) whether attribution signals align with interpretable linguistic categories in patient speech.

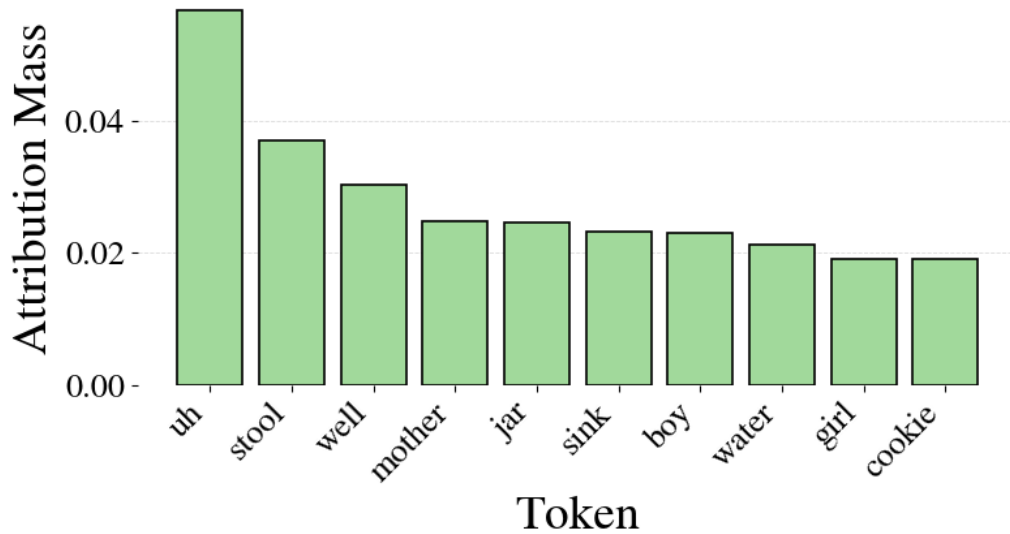


Figure 4.3: Top 10 tokens ranked by normalized attribution mass contributing to AD predictions

#### 4.5.1 Token Attribution Analysis

We first examine which tokens contribute most strongly to model predictions. For each transcript, we compute token-level attribution scores and extract the top- $k$  tokens with highest attribution mass (Fig. 4.3). Tokens were filtered to remove artifacts such as punctuation and isolated characters. Among the most highly attributed tokens, discourse fillers such as “*uh*” and “*well*” appear prominently. Such disfluencies are commonly reported in clinical studies of dementia-related speech (2; 45; 46). However, attribution alone does not establish whether the model consistently relies on these cues across training runs or whether the observed patterns are stable. We therefore perform additional analyses to assess the reliability of attribution explanations.

#### 4.5.2 Attribution Stability

To evaluate explanation reliability, we measure whether influential tokens remain consistent across models trained with different random seeds. Fig. 4.4 reports Jaccard similarity for overlap of top- $k$  tokens and Spearman correlation for attribution score rankings (see

|          |                    |                    |                    |                     |
|----------|--------------------|--------------------|--------------------|---------------------|
| Clinical | 0.16<br>$\pm 0.09$ | 0.23<br>$\pm 0.12$ | 0.23<br>$\pm 0.12$ | Jaccard (top-k)     |
|          | 0.16<br>$\pm 0.10$ | 0.24<br>$\pm 0.12$ | 0.24<br>$\pm 0.11$ |                     |
|          | 0.16<br>$\pm 0.10$ | 0.23<br>$\pm 0.11$ | 0.23<br>$\pm 0.11$ |                     |
| Ling.    |                    |                    |                    | Spearman ( $\rho$ ) |
|          | 0.29<br>$\pm 0.13$ | 0.32<br>$\pm 0.14$ | 0.25<br>$\pm 0.26$ |                     |
|          | 0.31<br>$\pm 0.16$ | 0.31<br>$\pm 0.13$ | 0.24<br>$\pm 0.22$ |                     |
| Neutral  |                    |                    |                    |                     |
|          | 0.33<br>$\pm 0.19$ | 0.30<br>$\pm 0.12$ | 0.25<br>$\pm 0.19$ |                     |
|          |                    |                    |                    |                     |
|          | IG                 | DL                 | DLS                | Attribution method  |

Figure 4.4: Attribution stability across prompts. IG: Integrated Gradients. DL: DeepLIFT. DLS: DeepLIFT-Shap. Top: Jaccard similarity (top- $k$  tokens). Bottom: Spearman correlation ( $\rho$ ). Mean  $\pm$  std. over five seeds

Sec. 3.3). Across prompt settings and attribution methods, Jaccard overlap remains low ( $\sim 0.16$ ), showing that the specific tokens identified as most influential vary substantially across training runs. Spearman correlations (0.29–0.33) show low agreement in attribution rankings. These results show that token-level attribution explanations are not stable across training runs, raising concerns about the reliability of such explanations for clinical interpretation.

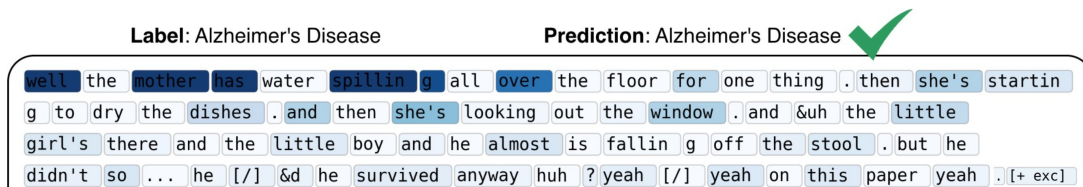


Figure 4.5: Token heatmap for sample patient transcripts using VCG:Neutral model. More in Fig. B.4 (Appendix)

### 4.5.3 Linguistic Feature Alignment of Attribution

To examine whether attribution patterns correspond to interpretable linguistic categories, we aggregate token-level attribution scores by part-of-speech (POS) labels derived from CHAT annotations. Subword tokens are aligned to CHAT word units and attribution mass is pooled across linguistic categories. Fig. 4.6 shows the distribution of attribution mass across POS categories. A large portion of attribution is concentrated in the residual *Other* category, while nouns and function words receive moderate attribution. Categories more directly associated with dementia-related speech patterns such as verbs, discourse markers, and disfluencies receive comparatively less attribution mass. These results indicate that token-level attribution explanations do not strongly align with interpretable linguistic categories in patient speech, limiting their usefulness for explaining model behavior in clinically meaningful terms. One possible explanation is that the clinically framed prompts encourage the model to attend to higher-level discourse and narrative structure, such as idea density, repetition, and topic completeness, which are difficult to capture through attribution scores assigned to individual tokens. As a result, improvements from clinical prompting may arise from representations distributed across larger linguistic structures rather than from a small set of highly attributed words.

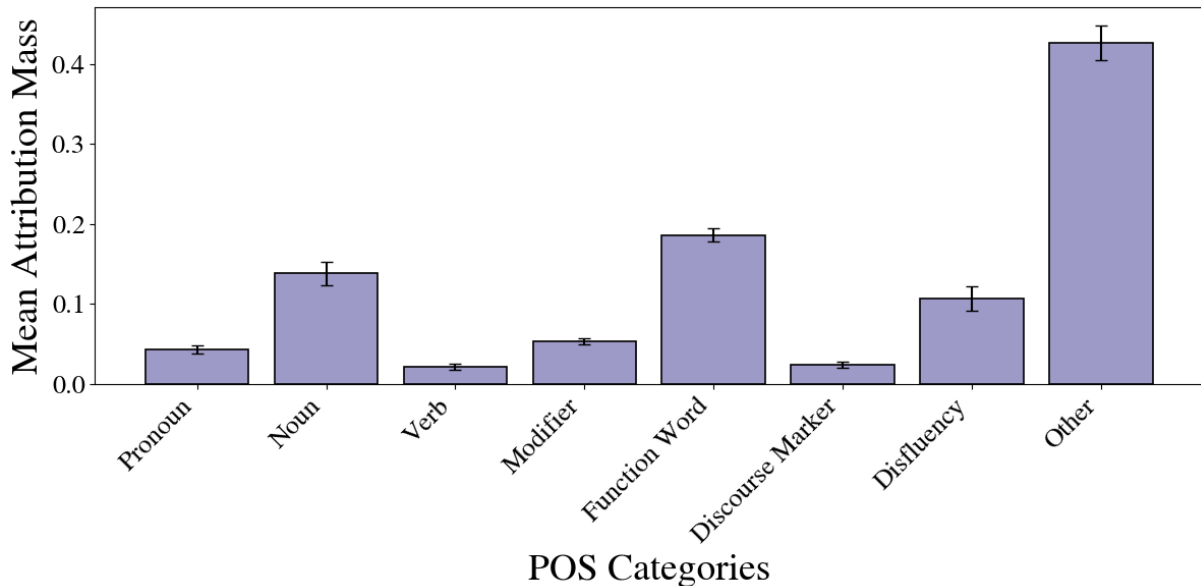


Figure 4.6: Token attributions aggregated onto CHAT-aligned linguistic categories to compare POS categories contributing to AD predictions

## Discussion & Future Work

### 5.1 LLM Adaptation for Dementia Detection

Pretrained LLMs achieve strong performance for transcript-based dementia classification when adapted with PEFT methods. Our results suggest that performance depends strongly on how clinical context and task information are incorporated: prompts including linguistic cues in natural language outperform numeric or minimally framed baselines. This indicates that LLMs integrate domain knowledge more effectively when it aligns with their pretrained language representations. Future work should explore more structured ways of incorporating clinical context, potentially through prompts or representations that reflect how clinicians interpret speech during cognitive assessments.

#### 5.1.1 Interpretability and Attribution Reliability

Despite strong predictive performance, interpreting model behavior remains difficult. Attribution analysis shows variability across training runs, with low agreement in which tokens are identified as most influential. POS-level aggregation further indicates limited alignment between attribution mass and interpretable POS categories. These findings suggest that token-level attribution explanations do not consistently uncover linguistically meaningful signals in this task. This mismatch may reflect limitations of attribution methods, LLM reliance on irrelevant cues, or constraints of the dataset and task format. Future work should therefore evaluate explanation methods with the same rigor as predictive performance, including stability analysis and alignment with domain-relevant linguistic features.

### 5.1.2 Implications for Clinical NLP

The results highlight an important distinction between predictive performance and clinically meaningful reasoning. While LLM-based models can capture statistical differences in language associated with cognitive impairment, reliable clinical interpretation requires explanations that are stable and aligned with clinically relevant linguistic markers. In practice, transcript-based LLM systems may still provide useful screening signals, especially when compared with acoustic-based systems as transcripts are easier to store, share, and standardize across clinical settings, while also enabling direct analysis of linguistic features associated with cognitive impairment. However, given the instability and limited interpretability observed in this study, such systems should be viewed as decision-support tools rather than standalone diagnostic instruments. A broader challenge is that commonly used dementia detection benchmarks were developed before LLM interpretability became a focus and provide limited supervision for evaluating explanations. Future datasets with clinician-informed linguistic annotations may help enable more rigorous interpretability evaluation.

## 5.2 Future Research Directions

Our findings suggest several directions for future work. First, interpretability evaluation should incorporate clinician-informed alignment metrics that assess whether model explanations correspond to relevant linguistic markers of cognitive decline. Second, future studies should compare multiple interpretability approaches beyond token attribution, such as representation probing or concept-based explanations, to better understand how models encode linguistic patterns relevant to dementia. Third, broader datasets are needed to determine whether the observed limitations arise from the small-data setting of current benchmarks or reflect more fundamental challenges in modeling clinical language. Finally, evaluating explanation stability across seeds and model configurations should become standard practice when assessing interpretability in clinical NLP systems.

## Conclusion

We study finetuning of LLMs for dementia classification using speech transcripts from the ADRess-IS2020 dataset. Finetuned models achieve strong performance with text-only inputs, and clinically guided prompting further improves accuracy across model sizes. However, interpretability analyses show that token attribution patterns are unstable across runs and weakly aligned with clinically interpretable linguistic features. These findings highlight the need for more rigorous and standardized evaluation of interpretability methods when applying LLMs to clinical language tasks.

### 6.1 Limitations

This study has several limitations that should be considered when interpreting the results.

**Dataset size and task scope.** Our experiments are conducted on the ADRess Cookie Theft picture-description task, which provides a relatively small and controlled dataset. While this dataset is widely used for dementia detection research, its limited size and single-task format may restrict the generalizability of the findings. The picture-description task differs from more naturalistic forms of speech, such as spontaneous conversation, autobiographical recall, or clinician-patient interactions, where linguistic behavior may be expressed differently. In addition, the dataset contains only single-session recordings and does not support evaluation of longitudinal changes in language over time. As a result, the performance and explainability patterns observed here may not directly transfer to real-world clinical monitoring or in conversational settings beyond structured picture descriptions.

**Interpretability methodology.** Our interpretability analysis relies primarily on token-level attribution methods. While these techniques provide a useful interface for visualizing model influence, attribution scores are known to exhibit variability across model runs and may not fully capture the reasoning processes of LLMs. As shown in our stabil-

ity analysis, token-level explanations can vary across training seeds, and their alignment with interpretable linguistic categories is limited. Other interpretability approaches may provide complementary insights.

**Evaluation variability.** Although we report multi-seed experiments and bootstrap confidence intervals for key comparisons, performance estimates remain subject to variability due to the limited dataset size and the stochastic nature of model training. Additional runs and larger evaluation datasets would provide more precise estimates of performance differences between prompting strategies.

## REFERENCES

- [1] E. Nichols, J. D. Steinmetz, S. E. Vollset, K. Fukutaki, J. Chalek, F. Abd-Allah, A. Abdoli, A. Abualhasan, E. Abu-Gharbieh, T. T. Akram *et al.*, “Estimation of the global prevalence of dementia in 2019 and forecasted prevalence in 2050: an analysis for the global burden of disease study 2019,” *The Lancet Public Health*, vol. 7, no. 2, pp. e105–e125, 2022.
- [2] J. Yuan, Y. Bian, X. Cai, J. Huang, Z. Ye, and K. Church, “Disfluencies and fine-tuning pre-trained language models for detection of alzheimer’s disease.” in *Inter-speech*, vol. 2020, 2020, pp. 2162–6.
- [3] A. Khodabakhsh, S. Kuşxuoğlu, and C. Demiroğlu, “Natural language features for detection of alzheimer’s disease in conversational speech,” in *IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*. IEEE, 2014, pp. 581–584.
- [4] E. Burke, J. Gunstad, and P. Hamrick, “Comparing global and local semantic coherence of spontaneous speech in persons with alzheimer’s disease and healthy controls,” *Applied Corpus Linguistics*, vol. 3, no. 3, p. 100064, 2023.
- [5] O. Ivanova, I. Martínez-Nicolás, E. García-Piñuela, and J. J. G. Meilán, “Defying syntactic preservation in alzheimer’s disease: what type of impairment predicts syntactic change in dementia (if it does) and why?” *Frontiers in Language Sciences*, vol. 2, p. 1199107, 2023.
- [6] M. Pattison, A. Begde, and T. D. Wilcockson, “Detecting dementia using lexical analysis: Terry pratchett’s discworld tells a more personal story,” *Brain Sciences*, vol. 16, no. 1, p. 94, 2026.
- [7] H. Goodglass, E. Kaplan, and S. Weintraub, *BDAE: The Boston diagnostic aphasia examination*. Lippincott Williams & Wilkins Philadelphia, PA, 2001.
- [8] R. Ostrand and J. Gunstad, “Using automatic assessment of speech production to predict current and future cognitive function in older adults,” *Journal of Geriatric Psychiatry and Neurology*, vol. 34, no. 5, p. 357–369, Sep. 2021. [Online]. Available: <https://doi.org/10.1177/0891988720933358>
- [9] S. Luz, F. Haider, S. de la Fuente Garcia, D. Fromm, and B. MacWhinney, “Alzheimer’s dementia recognition through spontaneous speech,” p. 780169, 2021.
- [10] A. Javeed, A. L. Dallora, J. S. Berglund, A. Ali, L. Ali, and P. Anderberg, “Machine learning for dementia prediction: a systematic review and future research directions,” *Journal of medical systems*, vol. 47, no. 1, p. 17, 2023.
- [11] E. Edwards, C. Dognin, B. Bollepalli, M. K. Singh, and V. Analytics, “Multiscale

- system for alzheimer’s dementia recognition through spontaneous speech.” in *Inter-speech*, 2020, pp. 2197–2201.
- [12] Y. Zhu, N. Lin, K. Balivada, D. Haehn, and X. Liang, “Adversarial text generation using large language models for dementia detection,” in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024, pp. 21 918–21 933.
  - [13] C. Park, A. S. G. Choi, S. Cho, and C. Kim, “Reasoning-based approach with chain-of-thought for alzheimer’s detection using speech and large language models,” *arXiv preprint arXiv:2506.01683*, 2025.
  - [14] Y. Xiao, H. Christensen, and S. Goetze, “Alzheimer’s dementia detection using perplexity from paired large language models,” *arXiv preprint arXiv:2506.09315*, 2025.
  - [15] S. S. Band, A. Yarahmadi, C.-C. Hsu, M. Biyari, M. Sookhak, R. Ameri, I. Dehzangi, A. T. Chronopoulos, and H.-W. Liang, “Application of explainable artificial intelligence in medical health: A systematic review of interpretability methods,” *Informatics in Medicine Unlocked*, vol. 40, p. 101286, 2023.
  - [16] S. U. Amin, M. Guizani, and M. S. Hossain, “Advances, evaluation, and explainability of large language models in healthcare: A systematic review,” *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 22, no. 2, pp. 1–32, 2026.
  - [17] M. R. Kumar, S. Vekkot, S. Lalitha, D. Gupta, V. J. Govindraj, K. Shaukat, Y. A. Alotaibi, and M. Zakariah, “Dementia detection from speech using machine learning and deep learning architectures,” *Sensors*, vol. 22, no. 23, p. 9311, 2022.
  - [18] L. Hason and S. Krishnan, “Spontaneous speech feature analysis for alzheimer’s disease screening using a random forest classifier,” *Frontiers in Digital Health*, vol. 4, p. 901419, 2022.
  - [19] M. Gumus, M. Koo, C. M. Studzinski, A. Bhan, J. Robin, and S. E. Black, “Linguistic changes in neurodegenerative diseases relate to clinical symptoms,” *Frontiers in Neurology*, vol. 15, Mar. 2024. [Online]. Available: <https://www.frontiersin.org/journals/neurology/articles/10.3389/fneur.2024.1373341/full>
  - [20] K. Chlasta and K. Wolk, “Towards computer-based automated screening of dementia through spontaneous speech,” *Frontiers in Psychology*, vol. 11, p. 623237, 2021.
  - [21] Y. Santander-Cruz, S. Salazar-Colores, W. J. Paredes-García, H. Guendulain-Arenas, and S. Tovar-Arriaga, “Semantic feature extraction using sbert for dementia detection,” *Brain sciences*, vol. 12, no. 2, p. 270, 2022.
  - [22] Z. Jahan, S. B. Khan, and M. Saraee, “Early dementia detection with speech analysis and machine learning techniques,” *Discover Sustainability*, vol. 5, no. 1, p. 65, 2024.
  - [23] K. C. Fraser, J. A. Meltzer, and F. Rudzicz, “Linguistic features identify alzheimer’s disease in narrative speech,” *Journal of Alzheimer’s disease*, vol. 49, no. 2, pp. 407–422, 2015.

- [24] N. Cummins, Y. Pan, Z. Ren, J. Fritsch, V. S. Nallanthighal, H. Christensen, D. Blackburn, B. W. Schuller, M. Magimai-Doss, H. Strik *et al.*, “A comparison of acoustic and linguistics methodologies for alzheimer’s dementia recognition,” in *Interspeech 2020*. ISCA-International Speech Communication Association, 2020, pp. 2182–2186.
- [25] A. Balagopalan and J. Novikova, “Comparing acoustic-based approaches for alzheimer’s disease detection,” *arXiv preprint arXiv:2106.01555*, 2021.
- [26] N. Wang, B. Wen, M. Wu, Y. Sun, Z. Shao, H. Zhou, and K. Subbalakshmi, “Decoding alzheimer’s: Interpretable visual and logical attention in picture description tasks,” in *Interspeech 2025 Conference-17-21 August’25*, 2025.
- [27] Y. Zhu, N. Lin, X. Liang, J. A. Batsis, R. M. Roth, and B. MacWhinney, “Evaluating picture description speech for dementia detection using image-text alignment,” *ACM Transactions on Computing for Healthcare*, 2023.
- [28] B. Lee, J.-U. Bang, H. J. Song, and B. O. Kang, “Alzheimer’s disease recognition using graph neural network by leveraging image-text similarity from vision language model,” *Scientific Reports*, vol. 15, no. 1, p. 997, 2025.
- [29] P. Hoffman, R. W. Jones, and M. A. L. Ralph, “Be concrete to be comprehended: Consistent imageability effects in semantic dementia for nouns, verbs, synonyms and associates,” *Cortex*, vol. 49, no. 5, pp. 1206–1218, 2013.
- [30] K. Singhal, T. Tu, J. Gottweis, R. Sayres, E. Wulczyn, M. Amin, L. Hou, K. Clark, S. R. Pfohl, H. Cole-Lewis *et al.*, “Toward expert-level medical question answering with large language models,” *Nature Medicine*, vol. 31, no. 3, pp. 943–950, 2025.
- [31] M. Sundararajan, A. Taly, and Q. Yan, “Axiomatic attribution for deep networks,” in *International conference on machine learning*. PMLR, 2017, pp. 3319–3328.
- [32] A. Shrikumar, P. Greenside, and A. Kundaje, “Learning important features through propagating activation differences,” in *International conference on machine learning*. PMIR, 2017, pp. 3145–3153.
- [33] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” *Advances in neural information processing systems*, vol. 30, 2017.
- [34] B. P. Gond, M. Shahnawaz, D. P. Mohapatra *et al.*, “Nlp-driven malware classification: A jaccard similarity approach,” in *2024 IEEE International Conference on Information Technology, Electronics and Intelligent Communication Systems (ICITE-ICS)*. IEEE, 2024, pp. 1–8.
- [35] C. Spearman, “The proof and measurement of association between two things.” 1961.
- [36] C. Pye, “The chldes project: Tools for analyzing talk,” 1994.
- [37] J. Becker, F. Boller, O. Lopez, J. Saxton, and K. McGonigle, “Dementia bank dataset. english pitt corpus,” *The natural history of Alzheimer’s disease. Description of study cohort and accuracy of diagnosis. Arch Neurol*, vol. 51, pp. 585–94, 1994.

- [38] A. M. Lanzi, A. K. Saylor, D. Fromm, H. Liu, B. MacWhinney, and M. L. Cohen, “Dementiabank: Theoretical rationale, protocol, and illustrative analyses,” *American Journal of Speech-Language Pathology*, vol. 32, no. 2, pp. 426–438, 2023.
- [39] Meta AI, “Introducing llama 3,” <https://ai.meta.com/blog/meta-llama-3>, 2024, accessed: 2024-10-05.
- [40] H. Liu, C. Li, Q. Wu, and Y. J. Lee, “Visual instruction tuning,” *Advances in neural information processing systems*, vol. 36, pp. 34 892–34 916, 2023.
- [41] M. S. S. Syed, Z. S. Syed, M. Lech, and E. Pirogova, “Automated screening for alzheimer’s dementia through spontaneous speech.” in *Interspeech*, vol. 2020, 2020, pp. 2222–6.
- [42] M. Chen, C. Miao, J. Ma, S. Wang, and J. Xiao, “Exploring multi-task learning and data augmentation in dementia detection with self-supervised pretrained models,” in *Proc. INTERSPEECH*, vol. 2023, 2023, pp. 5037–5041.
- [43] Y. Zhu, X. Liang, J. A. Batsis, and R. M. Roth, “Exploring deep transfer learning techniques for alzheimer’s dementia detection,” *Frontiers in computer science*, vol. 3, p. 624683, 2021.
- [44] C. Botelho, J. Mendonça, A. Pompili, T. Schultz, A. Abad, and I. Trancoso, “Macro-descriptors for alzheimer’s disease detection using large language models,” in *Proc. Interspeech*, vol. 2024, 2024.
- [45] S. Farzana, A. Deshpande, and N. Parde, “How you say it matters: Measuring the impact of verbal disfluency tags on automated dementia detection,” in *Proceedings of the 21st workshop on biomedical language processing*, 2022, pp. 37–48.
- [46] S. Nasreen, M. Rohanian, J. Hough, and M. Purver, “Alzheimer’s dementia recognition from spontaneous speech using disfluency and interactional features,” *Frontiers in Computer Science*, vol. 3, p. 640669, 2021.
- [47] A. Pistono, M. Jucla, E. J. Barbeau, L. Saint-Aubert, B. Lemesle, B. Calvet, B. Köpke, M. Puel, and J. Pariente, “Pauses during autobiographical discourse reflect episodic memory processes in early alzheimer’s disease,” *Journal of Alzheimer’s disease: JAD*, vol. 50, no. 3, p. 687–698, 2016.
- [48] Y. Zhu, A. Obyat, X. Liang, J. A. Batsis, and R. M. Roth, “Wavbert: Exploiting semantic and non-semantic speech using wav2vec and bert for dementia detection,” in *Interspeech*, vol. 2021, 2021, p. 3790.
- [49] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung, “Survey of hallucination in natural language generation,” *ACM computing surveys*, vol. 55, no. 12, pp. 1–38, 2023.
- [50] C. Dou, Y. Zhang, Y. Chen, Z. Jin, W. Jiao, H. Zhao, and Y. Huang, “Detection, diagnosis, and explanation: A benchmark for chinese medial hallucination evaluation,” in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 2024, pp.

4784–4794.

- [51] Q. Xie, Q. Chen, A. Chen, C. Peng, Y. Hu, F. Lin, X. Peng, J. Huang, J. Zhang, V. Keloth *et al.*, “Me-llama: Foundation large language models for medical applications,” *Research square*, pp. rs–3, 2024.
- [52] X. Wang, N. Chen, J. Chen, Y. Wang, G. Zhen, C. Zhang, X. Wu, Y. Hu, A. Gao, X. Wan *et al.*, “Apollo: A lightweight multilingual medical llm towards democratizing medical ai to 6b people,” *arXiv preprint arXiv:2403.03640*, 2024.
- [53] Y. Labrak, A. Bazoge, E. Morin, P.-A. Gourraud, M. Rouvier, and R. Dufour, “Biomistral: A collection of open-source pretrained large language models for medical domains,” *arXiv preprint arXiv:2402.10373*, 2024.
- [54] C. Li, C. Wong, S. Zhang, N. Usuyama, H. Liu, J. Yang, T. Naumann, H. Poon, and J. Gao, “Llava-med: Training a large language-and-vision assistant for biomedicine in one day,” *Advances in Neural Information Processing Systems*, vol. 36, pp. 28 541–28 564, 2023.
- [55] A. Sellergren, S. Kazemzadeh, T. Jaroensri, A. Kiraly, M. Traverse, T. Kohlberger, S. Xu, F. Jamil, C. Hughes, C. Lau *et al.*, “Medgemma technical report,” *arXiv preprint arXiv:2507.05201*, 2025.
- [56] M. Bouazizi, C. Zheng, S. Yang, and T. Ohtsuki, “Dementia detection from speech: What if language models are not the answer?” *Information*, vol. 15, no. 1, p. 2, 2023.
- [57] R. K. Garg, V. L. Urs, A. A. Agarwal, S. K. Chaudhary, V. Paliwal, and S. K. Kar, “Exploring the role of chatgpt in patient care (diagnosis and treatment) and medical research: A systematic review,” *Health Promotion Perspectives*, vol. 13, no. 3, p. 183, 2023.
- [58] H. L. Walker, S. Ghani, C. Kuemmerli, C. A. Nebiker, B. P. Müller, D. A. Raptis, and S. M. Staubli, “Reliability of medical information provided by chatgpt: assessment against clinical guidelines and patient information quality instrument,” *Journal of Medical Internet Research*, vol. 25, p. e47479, 2023.
- [59] M. Mesinovic, P. Watkinson, and T. Zhu, “Explainability in the age of large language models for healthcare,” *Communications Engineering*, vol. 4, no. 1, p. 128, 2025.
- [60] M. Honnibal, “spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing,” (*No Title*), 2017.
- [61] C. Hutto and E. Gilbert, “Vader: A parsimonious rule-based model for sentiment analysis of social media text,” in *Proceedings of the international AAAI conference on web and social media*, vol. 8, no. 1, 2014, pp. 216–225.
- [62] S. Bird, E. Klein, and E. Loper, *Natural language processing with Python: analyzing text with the natural language toolkit.* ” O’Reilly Media, Inc.”, 2009.
- [63] N. Reimers and I. Gurevych, “Sentence-bert: Sentence embeddings using siamese bert-networks,” *arXiv preprint arXiv:1908.10084*, 2019.
- [64] B. Hui, J. Yang, Z. Cui, J. Yang, D. Liu, L. Zhang, T. Liu, J. Zhang, B. Yu, K. Lu

- et al.*, “Qwen2. 5-coder technical report,” *arXiv preprint arXiv:2409.12186*, 2024.
- [65] G. Team, A. Kamath, J. Ferret, S. Pathak, N. Vieillard, R. Merhej, S. Perrin, T. Matejovicova, A. Ramé, M. Rivière *et al.*, “Gemma 3 technical report,” *arXiv preprint arXiv:2503.19786*, 2025.
- [66] D. Jiang, Y. Liu, S. Liu, J. Zhao, H. Zhang, Z. Gao, X. Zhang, J. Li, and H. Xiong, “From clip to dino: Visual encoders shout in multi-modal large language models,” *arXiv preprint arXiv:2310.08825*, 2023.
- [67] E. Alsentzer, J. Murphy, W. Boag, W.-H. Weng, D. Jindi, T. Naumann, and M. McDermott, “Publicly available clinical bert embeddings,” in *Proceedings of the 2nd clinical natural language processing workshop*, 2019, pp. 72–78.
- [68] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, and H. Poon, “Domain-specific language model pretraining for biomedical natural language processing,” *ACM Transactions on Computing for Healthcare (HEALTH)*, vol. 3, no. 1, pp. 1–23, 2021.
- [69] R. Luo, L. Sun, Y. Xia, T. Qin, S. Zhang, H. Poon, and T.-Y. Liu, “Biogpt: generative pre-trained transformer for biomedical text generation and mining,” *Briefings in bioinformatics*, vol. 23, no. 6, p. bbac409, 2022.

## APPENDICES

### A Paper Submitted and Under Preparation

Yui Ishihara, Kartik Patwari, Michelle Cohn, Alyssa Weakley, Chen-Nee Chuah, “Do LLMs Learn Meaningful Linguistic Patterns in Dementia Detection?”, *Submitted to ACL Rolling Review*.

### B Appendix

#### B.1 Extended Related Works

**Traditional Methods.** Traditional ML and NLP methods have been applied to language-based dementia detection (17; 20). Signal processing extracts acoustic cues such as pitch, pauses, speech rate, and harmonic-to-noise ratio from speech (2; 17; 18), while NLP approaches analyze lexical and syntactic patterns in transcripts using part-of-speech tagging and parsing (19; 8; 47). Classifiers such as Random Forest (18), Linear Discriminant Analysis (LDA) (9), and Support Vector Machines (SVMs) (41), along with deep models like Convolutional Neural Networks (CNNs) and Bidirectional Long Short-Term Memory (Bi-LSTM) networks (24), capture complex acoustic-linguistic patterns linked to cognitive decline (17; 20).

**LLM-Based Methods.** LLMs have recently been explored for their potential in dementia detection. Early studies extracted embeddings from transcribed speech for downstream classification (48; 21; 22), while later work finetuned LLMs directly for demen-

tia inference (2; 43; 14). Following works examined self-supervised speech encoders such as Wav2Vec2 and HuBERT (42), multimodal extensions like SpeechBERT (43), and probabilistic formulations based on paired perplexity differences rather than direct classification (14). Structured prompting with CoT reasoning further aims to emulate clinical-style inference (13). Despite promising accuracy, LLM-based approaches remain vulnerable to bias and hallucination (49; 50; 12), motivating alternative uses of LLMs such as annotating interpretable macro-descriptors for downstream modeling (44). Further, recent medical-domain LLMs demonstrate improved recognition of clinical terminology (51; 52; 53; 54; 30; 55), yet they lack specialization for cognitive assessment and often fail to capture subtle linguistic markers of dementia without diagnostic grounding (12; 56; 57; 58). Overall, while LLMs excel at pattern recognition, generating clinically reliable and explainable reasoning remains challenging (15; 59), motivating our progressive fine-tuning framework that directly aligns internal model representations with the dementia classification objective while systematically varying task framing and contextual grounding.

## B.2 Implementation and Training Details

All preprocessing and feature computation were implemented using reproducible open-source packages. Table B.1 reports the model versions and parameter settings used for each library.

**Training.** Finetuning is performed as supervised binary classification optimized with cross-entropy loss. Parameter-efficient finetuning is performed using Low-Rank Adaptation (LoRA; rank  $r = 8$ , scaling  $\alpha = 16$ , dropout 0.1) applied to attention and projection layers. Data are split 80:20 into train-validation subsets with label stratification, and the test sets are reserved for evaluation. Inputs are tokenized using the LLAMA tokenizer (max length 512). We train for up to 15 epochs with learning rate  $5 \times 10^{-5}$  (see Fig B.2), and batch size 1 with early stopping after 6 non-improving validation epochs, typically converging around 10 epochs. The best validation checkpoint is used for final testing. All experiments were run on an NVIDIA A4500 GPU using PyTorch 2.2 with PEFT and

Transformers (v4.42). finetuning took approximately 8 minutes for the 1B models and 20 minutes for the 8B models. Each experiment was conducted using five random seeds, and results are reported as the mean and standard deviation across runs to account for variability in optimization and initialization.

**Evaluation Protocol.** We report accuracy (Acc.), F1-score, precision (Prec.), sensitivity (Sens.), specificity (Spec.), and AUC, with mean and standard deviation computed across independent training runs (12; 13). For statistical comparison between prompt conditions, we perform paired nonparametric bootstrap resampling over matched random seeds. For each seed, we compute metric differences (e.g., balanced accuracy, macro-F1) between conditions on the same test set, and estimate 95% confidence intervals over the mean paired difference using bootstrap resampling (5,000 iterations) with replacement across seeds. This paired design isolates variability due to initialization while controlling for dataset-level variation. Attribution stability is evaluated across seeds using token-level attribution scores. For each sample, we extract the top- $k$  tokens ranked by absolute attribution magnitude and compute pairwise Jaccard similarity across seed pairs to measure overlap in salient tokens. In parallel, we compute Spearman’s rank correlation over full attribution vectors to quantify consistency in attribution ordering. Stability metrics are aggregated across all examples and seed pairs, and 95% confidence intervals are estimated via bootstrap resampling over example-level scores.

### B.3 Linguistic Feature Extraction

This section provides detailed descriptions of the linguistic and sentiment-based features used to augment and examine our fine-tuned LLM models. All features were derived from open-source NLP toolkits to ensure reproducibility and interpretability. Table B.1 summarizes the libraries and corresponding feature categories.

Table B.1: **Linguistic feature extraction setup.** Open-source NLP libraries used for feature computation and their implementation details.

| Library / Model                   | Extracted Features                                                                  | Implementation Details                                                                                                                         |
|-----------------------------------|-------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------|
| spaCy (60)                        | Lexical diversity (TTR), POS ratios (nouns, verbs, pronouns), sentence segmentation | spaCy v3.7 with <code>en_core_web_sm</code> ; NER disabled for tokenization and POS tagging                                                    |
| VADER (61)                        | Sentiment polarity (compound score)                                                 | <code>vaderSentiment</code> v3.3.2 with default compound scoring thresholds                                                                    |
| NLTK (62)                         | Repetition ratio (lexical bigram frequency)                                         | NLTK v3.8 using frequency counts via <code>Counter</code>                                                                                      |
| Sentence-BERT (MiniLM-L6-v2) (63) | Semantic coherence (adjacent-sentence cosine similarity)                            | <code>sentence-transformers</code> v2.2.2 using model <code>all-MiniLM-L6-v2</code> ; cosine similarity between normalized sentence embeddings |

Table B.2: **Numeric-to-text conversion for linguistic features.** Normalized ( $z$ -score) values are mapped into qualitative levels (*low*, *moderate*, *high*) and converted into short natural-language summaries.

| Feature             | Level                   | Description                                                                                      |
|---------------------|-------------------------|--------------------------------------------------------------------------------------------------|
| Type-Token Ratio    | Low<br>Moderate<br>High | used limited vocabulary<br>used moderately varied vocabulary<br>used rich and diverse vocabulary |
| Filler Ratio        | Low<br>Moderate<br>High | rarely used filler words<br>used some filler words<br>frequently used filler words (“uh”, “um”)  |
| Avg Sentence Length | Low<br>High             | spoke in short, simple sentences<br>spoke in long and complex sentences                          |
| Pronoun Ratio       | High                    | used many pronouns (“he”, “she”, “they”)                                                         |
| Sentiment Mean      | Low<br>High             | spoke with a negative or neutral tone<br>spoke with a positive emotional tone                    |
| Coherence           | Low<br>High             | spoke incoherently or jumped topics<br>strong coherence and logical flow                         |

Each transcript was processed through these tools to extract nine interpretable linguistic indicators spanning lexical diversity, syntactic structure, and affective tone. To evaluate whether LLMs can benefit from explicit language priors, these numeric features were either appended directly or converted into natural-language descriptions. Table B.2 illustrates the conversion templates used for qualitative verbalization of normalized ( $z$ -score) features.

## B.4 Prompt Formats

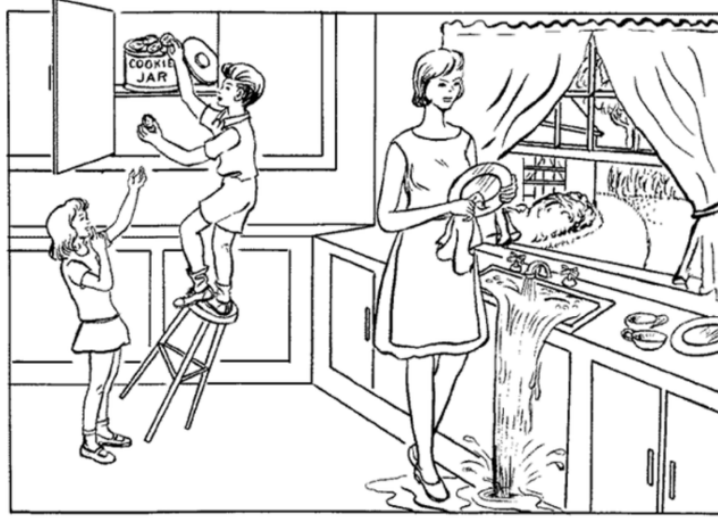


Figure B.1: The Cookie Theft picture from the Boston Diagnostic Aphasia Examination.

**VLM Captioning.** To incorporate visual context from the PDT for the VCG setting, we generate textual scene descriptions of the Cookie Theft Image (Fig. B.1) using multiple vision-language models (LLaVA-1.5-7B, GPT 5.2, and InternVL3.5-241B). Each model is prompted with a standardized instruction presented in Table B.3, producing natural-language captions that serve as a proxy for visual input. This design allows us to incorporate scene context without requiring multimodal architectures and to evaluate whether variations in caption generation influence downstream performance.

| Captioning Instruction                                                                                                                          |
|-------------------------------------------------------------------------------------------------------------------------------------------------|
| You are a vision-language assistant trained to extract structured scene information from images.<br>Describe the image.<br>{COOKIE THEFT IMAGE} |

Table B.3: Instruction used for generating cookie theft image context with VLM variants.

**Full Prompt Templates.** We construct input prompts by combining transcripts with generated scene captions under three prompting configurations that were derived in consultation with clinical and linguistic experts: VCG:Neutral, VCG:Linguistic, and VCG:Clinical. These prompts are designed to help evaluate how different forms of contextual and task-specific guidance influence model behavior as explained in Sec. 3.1. Full prompt templates are provided in Table ??.

**Prompt Length.** To observe whether performance differences are potentially driven by prompt length, we evaluate both short (50 tokens) and long (150 tokens) versions of each prompting configuration (Sec. 3.1). All three prompt types (VCG:Neutral, VCG:Linguistic, VCG:Clinical) are matched in length within each setting. For the neutral condition, additional tokens are filled with task-relevant but non-diagnostic information about the Cookie Theft image (Fig. B.1) to avoid introducing clinical or linguistic cues. Long prompts expand on the targeted features provided in the short versions of prompts with more detailed descriptions defined by in-house clinical and linguistic experts, allowing us to assess whether increased instructions improve model performance. Short and long prompt templates are provided in Table B.13.

## B.5 Additional Results

**Zero-shot performance.** Table B.4 reports zero-shot performance across a range of general-domain and biomedical LLMs. Overall, zero-shot accuracy remains close to random for most models (approximately 47–59%), indicating limited ability to perform dementia classification without task-specific adaptation. Among the evaluated models, PubMedBERT and Qwen2.5-7B-Instruct achieve the highest zero-shot accuracy, while PubMedBERT also attains the strongest F1 score. Biomedical encoder models with a lightweight classifier head generally outperform instruction-tuned LLMs, suggesting that domain-specific pretraining provides some advantage even without finetuning. However, performance remains substantially below finetuned results, highlighting the necessity of task adaptation for reliable dementia detection.

**Dataset Ablation.** Table B.5 reports model performance when incorporating transcripts from the Lu corpus which consists of 54 participants (26 control, 28 AD) and provides transcripts annotated in the CHAT format, consistent with ADReSS-IS2020. Training on the Lu corpus alone resulted in substantially lower performance, likely due to its small sample size. We therefore report results using a combined dataset (ADReSS-IS2020 + Lu). Performance decreases relative to training on ADReSS-IS2020 alone, suggesting the presence of domain mismatch between the two corpora. Other ADReSS variants were

Table B.4: Zero-shot performance of tested LLMs on dementia classification.

\*Encoder-only models are included with a lightweight linear classification head trained on frozen representations to provide a comparable baseline.

| Model                 | Acc. (%)    | F1 (%)    |
|-----------------------|-------------|-----------|
| Qwen2.5-7B-Inst.      | 58.33       | 61.54     |
| Bio-ClinicalBERT*     | 53.75±6.32  | 60.4±5.94 |
| Gemma-3-4B-IT         | 47.92       | 63.77     |
| MedGemma-4B-IT        | 47.92       | 63.77     |
| PuBMedBERT*           | 59.17±11.56 | 67.2±5.84 |
| BioGPT                | 47.92       | 62.69     |
| Mistral-7B-Inst.-v0.2 | 50          | 62.5      |
| LLaMA-3.2-1B-Inst.    | 47.92       | 63.77     |
| LLaMA-3.2-3B-Inst.    | 47.92       | 63.77     |
| LLaMA-3.1-8B-Inst.    | 47.92       | 63.77     |

not considered in this setting, as the Lu corpus is the only additional dataset identified that shares compatible CHAT-style annotations with ADReSS-IS2020, enabling consistent feature extraction. These results highlight the challenges of cross-dataset generalization even under shared annotation formats.

Table B.5: Performance of tested LLMs on additional dataset.

| Captioning Model          | Acc. (%)   | F1 (%)     |
|---------------------------|------------|------------|
| ADReSS-IS2020             | 86.25±3.78 | 85.79±4.6  |
| ADReSS-IS2020 + Lu Corpus | 80.50±3.49 | 77.80±4.89 |

**Additional VLM Backbones.** Appendix B.7 reports additional LLM backbones were evaluated to assess robustness across model families, scales, and domain specialization. These include general-purpose instruction-tuned models such as Qwen2.5-7B (64), Gemma-3-4B (65), MedGemma-4B (55), and Mistral-7B (66), alongside encoder-based biomedical models—BioClinicalBERT (67), PubMedBERT (68), and BioGPT (69). This comparison allows us to contrast general-domain language understanding with domain-specialized representations trained on clinical and biomedical text. For encoder-only models, a lightweight linear classification head is trained on frozen representations to provide a controlled and comparable baseline.

First, performance does not scale monotonically with model size: smaller instruction-tuned models (e.g., LLaMA-3.2-1B) achieve competitive or superior results relative to larger counterparts, suggesting that task-specific alignment and prompt conditioning play a more critical role than raw capacity in this setting. Second, domain-specific pretrain-

ing (e.g., BioClinicalBERT, PubMedBERT) does not consistently translate to improved performance, indicating that biomedical pretraining settings alone may not capture the discourse-level linguistic patterns characteristic of dementia. Third, results are more sensitive to prompt design than backbone choice, with clinically framed and structured inputs often yielding larger improvements than switching between architectures.

Together, these findings suggest that lightweight adaptation strategies and prompt conditioning can be primary drivers of performance, while backbone selection contributes comparatively less variability. However, our analysis treats each backbone as a black-box classifier and does not probe differences in internal representations, reasoning behavior, or attribution structure across model families. A more fine-grained investigation of how different architectures encode and utilize linguistic signals (e.g., through cross-model attribution alignment or representation probing) remains an important direction for future work.

**Additional VLM Captioning Models.** We evaluate the sensitivity of our best-performing configuration (VCG:Clinical with LLaMA-3.2-1B) to the choice of VLM used for caption generation as described in App. B.5. As shown in Table B.6, performance remains consistent across captions generated by LLaVA-1.5-7B, GPT-5.2, and InternVL3.5-241B, with only marginal differences in both accuracy and F1-score.

This stability suggests that downstream classification performance is largely resilient to the specific phrasing or stylistic variations introduced by different VLMs. These results support the robustness of the evaluation setting and indicate that observed performance gains are attributable to the prompting and modeling strategy rather than generated artifacts of a particular captioning model.

Table B.6: Performance of tested LLMs on captions generated by varied VLMs.

| Captioning Model | Acc. (%)   | F1 (%)     |
|------------------|------------|------------|
| LLaVA-1.5-7B     | 86.25±3.78 | 85.79±4.6  |
| GPT-5.2          | 85.42±2.08 | 85.61±1.84 |
| InternVL3.5-241B | 84.58±1.86 | 84.71±2.26 |

**Ablation on Learning Rates.** Figure B.2 shows the impact of learning rate across prompt formulations for two representative backbones (LLaMA-3.2-1B and LLaMA-3.1-

8B). For the smaller 1B model, performance improves consistently with higher learning rates, with both VCG and instruction-guided prompting reaching peak accuracy at  $5 \times 10^{-5}$ . In contrast, transcript-only prompting exhibits less stable behavior, with a drop at intermediate learning rates before partial recovery. For the larger 8B model, performance is comparatively less stable across learning rates, with smaller gains from increasing the learning rate. Across both model sizes, VCG:Transcript prompting demonstrates the most consistent improvement with increasing learning rate, suggesting that structured contextual grounding benefits more from stronger parameter updates during fine-tuning. In contrast, transcript-only inputs appear more sensitive to optimization settings and less robust to higher learning rates. These results indicate an interaction between prompt structure and optimization dynamics, where smaller models benefit more from aggressive learning rates and structured prompting, while larger models exhibit greater stability but reduced sensitivity to hyperparameter tuning.

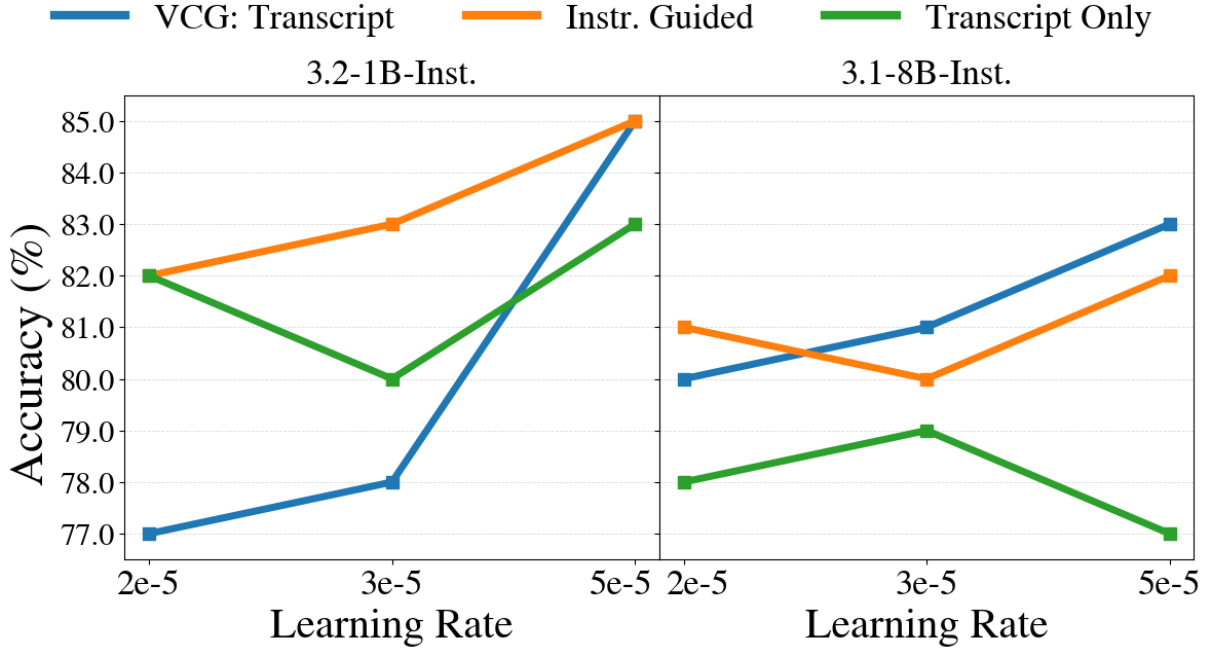


Figure B.2: Accuracy comparison across learning rates ( $2 \times 10^{-5} - 5 \times 10^{-5}$ ) for LLaMA-3.2-1B (left) and LLaMA-3.1-8B (right) under three prompting strategies: transcript-only, instruction-guided, and VCG:Transcript.

## B.6 Additional Explainability Results.

**Token Attribution by Prediction.** Figure B.3 shows the most influential tokens under prediction-conditioned attribution (i.e., computed with respect to the model’s predicted class). For AD predictions, high-attribution tokens include disfluency markers (“uh”, “well”) and salient scene entities (“sink”, “mother”), consistent with known linguistic markers of cognitive impairment such as hesitation and lexical retrieval difficulty. In contrast, NC predictions exhibit more functional and descriptive tokens (e.g., “getting”, “of”, “dishes”), possibly pointing to more fluent and structured narrative production. While these token-level patterns appear intuitively interpretable, qualitative inspection alone does not provide evidence of reliability. We therefore complement this analysis with quantitative measures of attribution stability and clinical alignment, detailed in Section 4.5.

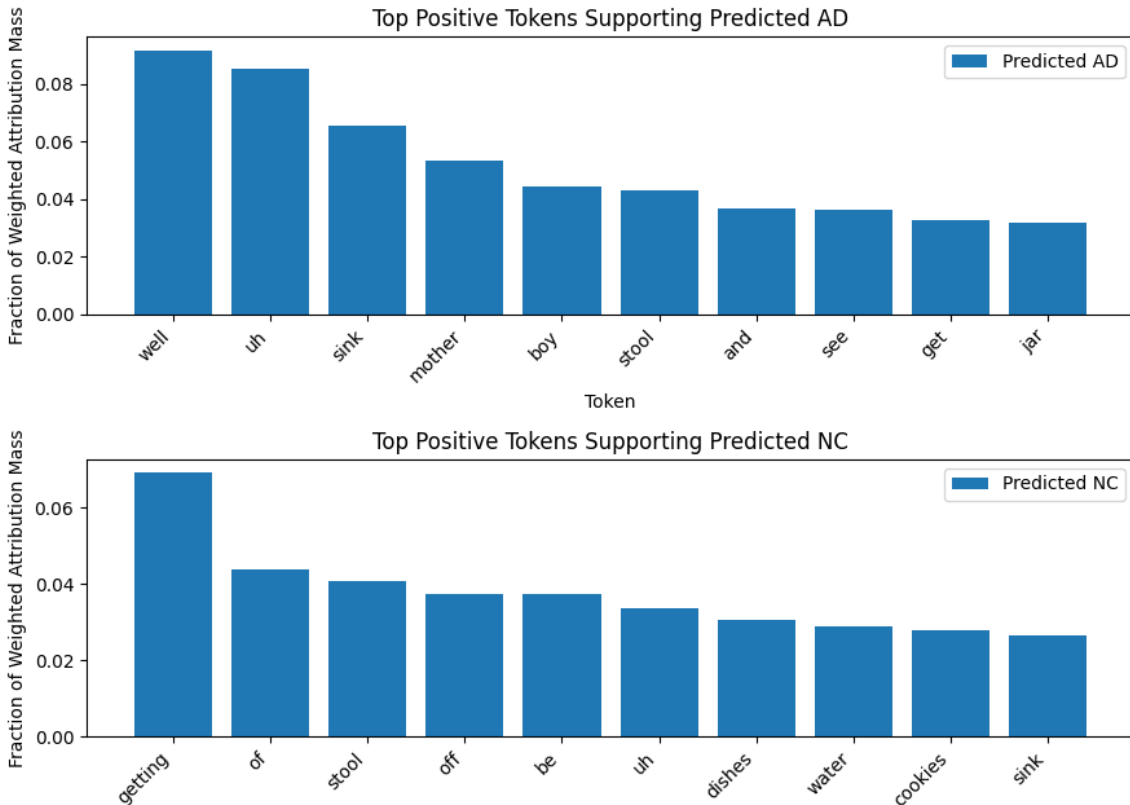


Figure B.3: Top 10 tokens with the highest attribution mass for samples predicted as AD (top) and NC (bottom), computed using DeepLIFT, our most stable attribution method. Attributions are conditioned on the model’s *predicted class* rather than fixed target labels, reflecting features that directly contribute to each decision.

**Attribution Stability Results.** Table B.7 summarizes attribution stability across methods and prompt conditions. Across all prompts, DL and DLS achieve consistently higher Jaccard overlap than IG, indicating more stable selection of top- $k$  influential tokens across training runs. However, this improvement in top-token consistency does not translate to global agreement: Spearman correlation remains moderate across methods, and DLS in particular exhibits substantially lower and more variable rank correlation. This suggests that while DLS may identify similar high-attribution tokens, the overall ordering of token importance is less stable. Notably, stability patterns are largely consistent across prompt types, indicating that attribution variability is driven more by the choice of attribution method than by prompt setting. Overall, these results highlight that different attribution methods capture different aspects of stability, and that agreement on salient tokens does not necessarily imply reliable global explanations.

Table B.7: **Attribution stability across prompts and methods.** Top: Jaccard overlap (Top- $k$ ) measuring consistency of salient tokens across seeds. Bottom: Spearman rank correlation measuring global consistency of attribution scores. IG=Integrated Gradients, DL=DeepLIFT, and DLS=DeepLIFTSHAP.

| Prompt                              | Attr. | Mean      | Bootstrap 95% CI |
|-------------------------------------|-------|-----------|------------------|
| <i>Jaccard (Top-<math>k</math>)</i> |       |           |                  |
| Clinical                            | IG    | 0.16±0.09 | [0.148, 0.165]   |
|                                     | DL    | 0.23±0.12 | [0.219, 0.240]   |
|                                     | DLS   | 0.23±0.12 | [0.219, 0.241]   |
| Linguistic                          | IG    | 0.16±0.10 | [0.156, 0.174]   |
|                                     | DL    | 0.24±0.12 | [0.233, 0.254]   |
|                                     | DLS   | 0.24±0.11 | [0.231, 0.251]   |
| Neutral                             | IG    | 0.16±0.10 | [0.152, 0.170]   |
|                                     | DL    | 0.23±0.11 | [0.216, 0.236]   |
|                                     | DLS   | 0.23±0.11 | [0.223, 0.243]   |
| <i>Spearman (<math>\rho</math>)</i> |       |           |                  |
| Clinical                            | IG    | 0.29±0.13 | [0.275, 0.299]   |
|                                     | DL    | 0.32±0.14 | [0.312, 0.336]   |
|                                     | DLS   | 0.25±0.26 | [0.225, 0.272]   |
| Linguistic                          | IG    | 0.31±0.16 | [0.300, 0.329]   |
|                                     | DL    | 0.31±0.13 | [0.301, 0.325]   |
|                                     | DLS   | 0.24±0.22 | [0.220, 0.261]   |
| Neutral                             | IG    | 0.33±0.19 | [0.313, 0.346]   |
|                                     | DL    | 0.30±0.12 | [0.290, 0.312]   |
|                                     | DLS   | 0.25±0.19 | [0.231, 0.266]   |

**Pairwise Bootstrap Comparison Results.** Table B.9 compares prompting strategies under short and long configurations. Under short prompting, clinical prompts consistently outperform neutral and linguistic variants, with significant gains in both balanced accuracy and macro-F1 (confidence intervals excluding zero), particularly over linguistic prompts. In contrast, linguistic prompting shows no consistent advantage over neutral. For long prompting, these differences diminish: all confidence intervals overlap zero and results are more balanced, indicating no reliable differences between prompt types. Overall, prompt design matters more under short inputs, where clinical framing is most effective.

Table B.8 compares short and long prompts within each strategy. Clinical prompting benefits from shorter inputs, with significant improvements in both metrics. Linguistic prompting shows no benefit from longer prompts and slightly degrades, while neutral prompting shows inconsistent gains with high variance. Overall, shorter prompts are sufficient—and for clinical prompting, preferable—while longer prompts introduce variability without consistent improvement.

Table B.8: **Short vs. long prompt comparisons within each prompt setting.** Top: Balanced Accuracy. Bottom: Macro-F1.

| Prompt                   | Seed $\Delta$ Mean   | Bootstrap 95% CI        | W/T/L        |
|--------------------------|----------------------|-------------------------|--------------|
| <i>Balanced Accuracy</i> |                      |                         |              |
| Clinical                 | $0.0333 \pm 0.0378$  | <b>[0.0042, 0.0625]</b> | 3/2/0        |
| Linguistic               | $-0.0083 \pm 0.0280$ | [-0.0292, 0.0125]       | 2/0/3        |
| Neutral                  | $0.0458 \pm 0.0851$  | [-0.0125, 0.1208]       | <b>4/0/1</b> |
| <i>Macro-F1</i>          |                      |                         |              |
| Clinical                 | $0.0332 \pm 0.0378$  | <b>[0.0041, 0.0624]</b> | 3/2/0        |
| Linguistic               | $-0.0086 \pm 0.0279$ | [-0.0294, 0.0123]       | 2/0/3        |
| Neutral                  | $0.0501 \pm 0.0942$  | [-0.0125, 0.1338]       | <b>4/0/1</b> |

Building on the POS-based attribution analysis in Sec. 4.5.3, we next construct clinically motivated proxy features by aggregating attribution mass across linguistically interpretable categories (Table B.10). Specifically, we group pronouns and disfluencies to approximate word-finding difficulty, while retaining individual measures for disfluency and pronoun use, and defining a residual category to capture all remaining attribution mass. We then quantify how much of the model’s attribution is allocated to these clinically

Table B.9: **Pairwise prompt comparisons across short and long prompting settings.** Top: Balanced Accuracy. Bottom: Macro-F1.

| Length                   | Compar. | Seed $\Delta$ Mean   | Bootstrap 95% CI        | W/T/L        |
|--------------------------|---------|----------------------|-------------------------|--------------|
| <i>Balanced Accuracy</i> |         |                      |                         |              |
| Short                    | C – N   | $0.0208 \pm 0.0255$  | <b>[0.0000, 0.0375]</b> | <b>4/1/0</b> |
|                          | L – N   | $-0.0292 \pm 0.0480$ | [-0.0667, 0.0083]       | 2/0/3        |
|                          | C – L   | $0.0500 \pm 0.0432$  | <b>[0.0167, 0.0833]</b> | <b>4/1/0</b> |
| Long                     | C – N   | $0.0333 \pm 0.0788$  | [-0.0167, 0.1042]       | 4/0/1        |
|                          | L – N   | $0.0250 \pm 0.0851$  | [-0.0667, 0.0083]       | 2/2/1        |
|                          | C – L   | $0.0083 \pm 0.0238$  | [-0.0083, 0.0250]       | 3/1/2        |
| <i>Macro-F1</i>          |         |                      |                         |              |
| Short                    | C – N   | $0.0208 \pm 0.0253$  | <b>[0.0001, 0.0373]</b> | <b>4/1/0</b> |
|                          | L – N   | $-0.0294 \pm 0.0483$ | [-0.0667, 0.0083]       | 2/0/3        |
|                          | C – L   | $0.0502 \pm 0.0436$  | <b>[0.0166, 0.0840]</b> | <b>4/1/0</b> |
| Long                     | C – N   | $0.0377 \pm 0.0882$  | [-0.0168, 0.1173]       | 2/1/2        |
|                          | L – N   | $0.0292 \pm 0.0947$  | [-0.0667, 0.0083]       | 2/2/1        |
|                          | C – L   | $0.0084 \pm 0.0242$  | [-0.0085, 0.0295]       | 2/1/2        |

Table B.10: **Mapping from POS-aligned attribution categories to expert-informed proxy features.** Proxy features are constructed by grouping token-level attribution mass according to linguistic and clinical insights.

| Proxy Feature           | POS Categories Used                     |
|-------------------------|-----------------------------------------|
| Word-Finding Difficulty | Pronoun + Disfluency                    |
| Disfluency              | Disfluency                              |
| Pronoun Use             | Pronoun                                 |
| Residual                | Other + Punctuation                     |
| Clinical Proxy Total    | Pronoun + Disfluency + Discourse Marker |

Table B.11: **Quantitative alignment of attribution mass with clinically motivated proxy features.** Values are reported as mean  $\pm$  standard deviation across seeds, with approximate 95% confidence intervals.

| Metric                     | Mean $\pm$ Std                       | 95% CI                  |
|----------------------------|--------------------------------------|-------------------------|
| Clinical Proxy Total       | $0.174 \pm 0.016$                    | [0.158, 0.190]          |
| Residual                   | $0.426 \pm 0.022$                    | [0.404, 0.448]          |
| <b>Clinical – Residual</b> | <b><math>-0.252 \pm 0.027</math></b> | <b>[-0.279, -0.224]</b> |
| <b>Alignment Ratio</b>     | <b><math>0.290 \pm 0.029</math></b>  | <b>[0.261, 0.319]</b>   |

relevant proxies compared to residual features (Table B.11).

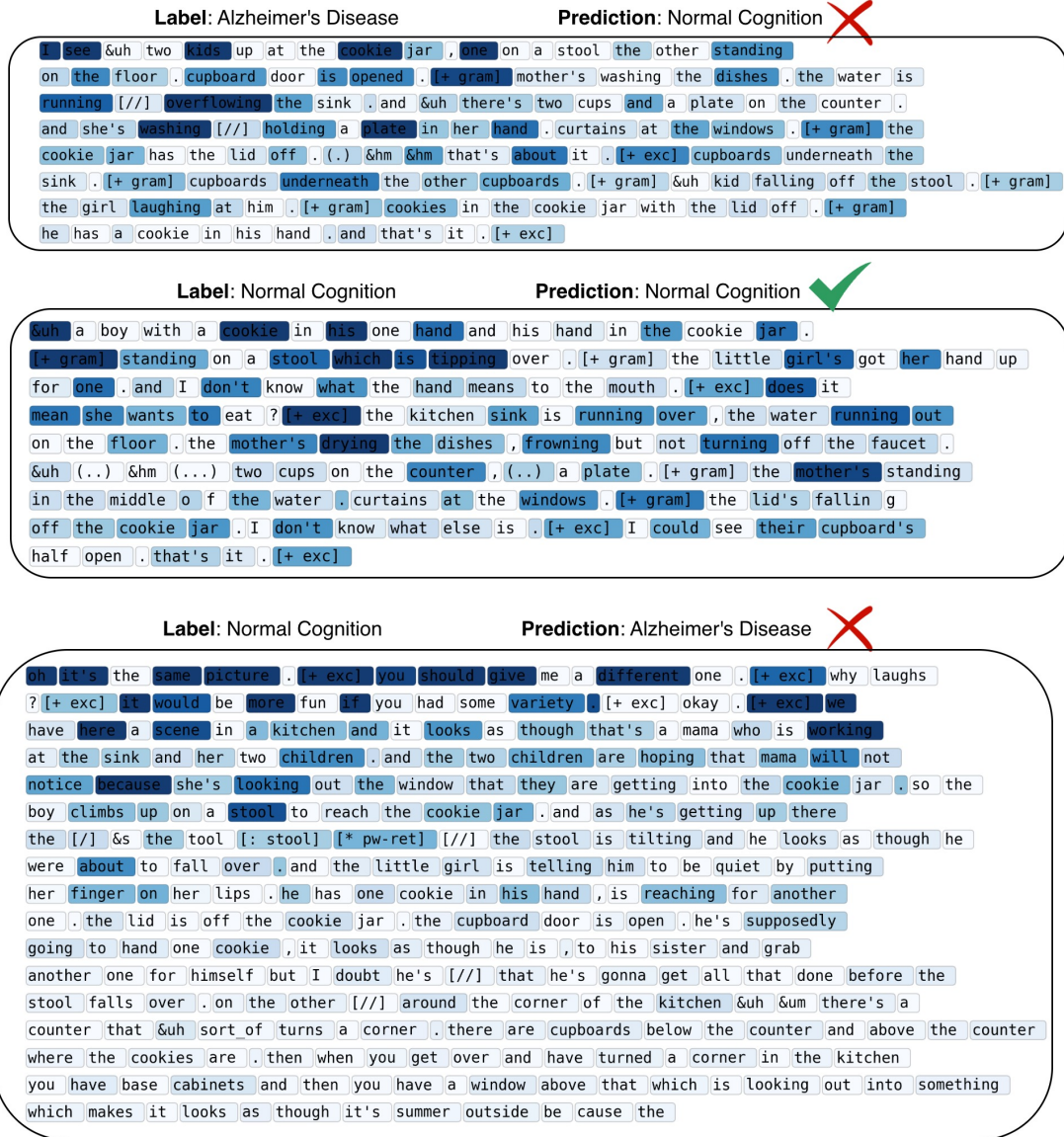


Figure B.4: Additional token heatmaps for patient transcripts using best model (VCG:Neutral).

Overall, these findings suggest that although the POS-based grouping enables a structured interpretation of model behavior, the learned decision signals are not strongly aligned with clinically motivated linguistic features, highlighting a gap between model reasoning and expert-informed expectations.

**Additional Heatmap Visualization.** Figure B.4 provides additional qualitative examples of token-level attribution heatmaps across both correct and incorrect predictions. These visualizations could potentially offer an interpretable window into the model's de-

cision process by highlighting which parts of the transcript contribute most strongly to each prediction. However, as demonstrated by our quantitative analyses of attribution stability and clinical feature alignment, token-level attributions exhibit substantial variability across seeds and only weak correspondence with clinically motivated linguistic features. Future work should focus on improving the stability and clinical alignment of attribution methods, as well as developing evaluation protocols that can quantify when such explanations can be trusted for the future development of clinically reliable tools.

## B.7 Comparison with Prior Work

Table B.12 compares our approach with prior work on the ADReSS dataset across both traditional and LLM-based methods. Among text-only approaches, our VCG:Clinical configuration achieves competitive performance (86.7% accuracy, 86.6% F1), outperforming several prior text-based models, including transformer and earlier LLM-based approaches, while remaining close to the strongest reported results. Notably, methods that incorporate both audio and text modalities can achieve higher performance (e.g., Zhu et al., 2021a), highlighting the advantage of multimodal information.

Within LLM-based approaches, our results are comparable to recent prompting-based methods, despite using a lightweight LoRA finetuning setup without additional modalities or complex inference strategies. While some recent works report higher accuracy (e.g., Xiao et al., 2025), these often rely on alternative evaluation settings or metrics that are not directly comparable. Overall, these results demonstrate that parameter-efficient finetuning of LLMs on transcript-only data can achieve strong performance, providing a competitive and lightweight alternative to more complex or multimodal systems.

\* indicates LLM-based methods. <sup>1</sup> Results from Park et al. were not reproducible using the authors’ released code.

| Method               | Venue             | Model                  | Audio | Text | Acc.        | F1          |
|----------------------|-------------------|------------------------|-------|------|-------------|-------------|
| (9)                  | Front. Comp. Sci. | LDA                    | ✗     | ✓    | 76.8        | 74.5        |
| (41)                 | IEEE Access       | SVM                    | ✗     | ✓    | 81.3        | –           |
| (42)                 | Interspeech       | Wav2Vec2               | ✓     | ✗    | 73.7        | 72.8        |
| (42)                 | Interspeech       | HuBERT                 | ✓     | ✗    | 74.2        | 73.6        |
| (24)                 | Interspeech       | Bi-LSTM (GloVe)        | ✗     | ✓    | 81.3        | 81.2        |
| (43)                 | Front. Comp. Sci. | BERT + SpeechBERT      | ✓     | ✓    | 82.9        | 82.9        |
| (12)*                | EMNLP             | LLaMA-3-8B-Inst.       | ✗     | ✓    | 83.3        | –           |
| (44)*                | Interspeech       | Transformer text model | ✗     | ✓    | 81.3        | –           |
| (14)*                | Interspeech       | Paired-perplexity LLM  | ✗     | ✓    | <b>91.7</b> | –           |
| (13)* <sup>1</sup>   | Interspeech       | LLaMA-3.2-1B + CoT     | ✓     | ✓    | 83.3        | 83.2        |
| (13)* <sup>1</sup>   | Interspeech       | LLaMA-3.2-1B + CoT     | ✗     | ✓    | 87.5        | <b>87.5</b> |
| <b>VCG:C (Ours)*</b> | –                 | LLaMA-3.2-1B (LoRA)    | ✗     | ✓    | 86.7        | 86.6        |
| <b>LFG:L (Ours)*</b> | –                 | LLaMA-3.2-1B (LoRA)    | ✗     | ✓    | 85.0        | 85.7        |

Table B.12: Comparison with prior work on the ADReSS dataset.

| Prompt Type | Short Prompt (50 tokens)                                                                                                                                                                                                                                                                                                                                                                            | Long Prompt (150 tokens)                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            |
|-------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Neutral     | <p>The following is a scene context and transcript from a Cookie Theft picture-description task. The Cookie Theft task is commonly used in dementia detection, where participants must describe everything happening in a picture. The task allows comparison between a participant’s performance and standardized norm.</p> <p>[SCENE CONTEXT]: {SCENE CONTEXT}<br/>[TRANSCRIPT]: {TRANSCRIPT}</p> | <p>The following is a scene context and transcript from a Cookie Theft picture-description task. The Cookie Theft task is commonly used in dementia detection, where participants are asked to describe everything happening in the picture. The task allows comparison between a participant’s performance and standardized norm. The task originates from the Boston Diagnostic Aphasia Examination developed by Harold Goodglass and Edith Kaplan in the 1970s–1980s. The task elicits spontaneous discourse rather than isolated words or sentences. Populations studied with Cookie Theft include: frontotemporal dementia, primary progressive aphasia, Parkinson’s disease, schizophrenia, and bipolar disorder. It has also been used as a cross-linguistic discourse elicitation tool for English, Mandarin Chinese, Spanish, German, and more.</p> <p>[SCENE CONTEXT]: {SCENE CONTEXT}<br/>[TRANSCRIPT]: {TRANSCRIPT}</p> |
| Linguistic  | <p>The following is a scene context and transcript from a Cookie Theft picture-description task. Focus on the following linguistic features commonly observed in dementia detection: disfluencies, syntactic complexity, lexical frequency, type-token ratio, discourse features, and lexical concreteness.</p> <p>[SCENE CONTEXT]: {SCENE CONTEXT}<br/>[TRANSCRIPT]: {TRANSCRIPT}</p>              | <p>The following is a scene context and transcript from a Cookie Theft picture-description task. Focus on the following linguistic features commonly observed in dementia detection:</p> <p>Disfluencies – fillers such as “uh,” or “like,” may indicate difficulty maintaining fluent speech.<br/>Pronoun use – frequent pronouns (he, she, it, this, that) instead of specific nouns.<br/>Syntactic complexity – use of embedded clauses and detailed descriptions rather than simple sentences.<br/>Lexical frequency – reliance on common versus less frequently used vocabulary.<br/>Discourse features – coherence and narrative structure across sentences.<br/>Type-token ratio – vocabulary diversity relative to total words.</p> <p>[SCENE CONTEXT]: {SCENE CONTEXT}<br/>[TRANSCRIPT]: {TRANSCRIPT}</p>                                                                                                                  |
| Clinical    | <p>The following is a scene context and transcript from a Cookie Theft picture-description task. Focus on the following clinical features commonly observed in dementia detection: word finding difficulties, repetition of words, length of sentences, topic repetition, idea density, and topic completeness.</p> <p>[SCENE CONTEXT]: {SCENE CONTEXT}<br/>[TRANSCRIPT]: {TRANSCRIPT}</p>          | <p>The following is a scene context and transcript from a Cookie Theft picture-description task. Focus on the following clinical features commonly observed in dementia detection:</p> <p>Aphasia-related symptoms – word-finding difficulty and pauses (“uh,” “um,” “like”).<br/>Repetition of words – excessive repetition of the same words or phrases.<br/>Sentence length – shorter or simplified sentences indicating reduced syntactic complexity.<br/>Topic repetition – repeating the same objects or ideas or forgetting prior mention.<br/>Idea density – number of distinct informational units relative to total words.<br/>Topic completeness – whether key elements of the scene are fully described.</p> <p>[SCENE CONTEXT]: {SCENE CONTEXT}<br/>[TRANSCRIPT]: {TRANSCRIPT}</p>                                                                                                                                     |

Table B.13: Comparison of short and long prompt formulations. Prompt length was controlled to examine how additional contextual information influences model behavior.

Table B.14: Performance across finetuning and conditioning strategies for LLaMA-3.2-1B-Instruct. Results are reported as mean  $\pm$  std. dev. over five random seeds.

|                                           | Model                 | Acc. (%)                         | F1 (%)                           | Prec. (%)                        | Sens. (%)                         | Spec (%)                         | AUC                             |
|-------------------------------------------|-----------------------|----------------------------------|----------------------------------|----------------------------------|-----------------------------------|----------------------------------|---------------------------------|
| Transcript Only                           | LLaMA-3.2-1B-Inst.    | 77.92 $\pm$ 3.78                 | 75.86 $\pm$ 7.32                 | <b>82.56<math>\pm</math>5.36</b> | 72.5 $\pm$ 16.82                  | 83.33 $\pm$ 9.77                 | <b>0.89<math>\pm</math>0.02</b> |
|                                           | LLaMA-3.2-3B-Inst.    | <b>79.17<math>\pm</math>2.55</b> | <b>81.36<math>\pm</math>2.14</b> | 73.68 $\pm$ 2.41                 | <b>90.83<math>\pm</math>1.86</b>  | 67.5 $\pm$ 3.49                  | 0.85 $\pm$ 0.02                 |
|                                           | LLaMA-3.1-8B-Inst.    | 77.08 $\pm$ 4.17                 | 74.65 $\pm$ 7.29                 | 82.34 $\pm$ 4.01                 | 70 $\pm$ 15.42                    | <b>84.17<math>\pm</math>8.01</b> | 0.87 $\pm$ 0.02                 |
| Task Guided                               | LLaMA-3.2-1B-Inst.    | 76.67 $\pm$ 6.49                 | 73.57 $\pm$ 10.04                | 82.28 $\pm$ 3.04                 | <b>68.33<math>\pm</math>17.82</b> | 85 $\pm$ 5.59                    | 0.87 $\pm$ 0.02                 |
|                                           | LLaMA-3.2-3B-Inst.    | 72.92 $\pm$ 5.31                 | 73.66 $\pm$ 8.19                 | 71.27 $\pm$ 4.84                 | 78.33 $\pm$ 15.98                 | <b>67.5<math>\pm</math>11.56</b> | 0.81 $\pm$ 0.06                 |
|                                           | LLaMA-3.1-8B-Inst.    | <b>78.75<math>\pm</math>6.49</b> | <b>76.23<math>\pm</math>9.05</b> | <b>85.32<math>\pm</math>7.31</b> | 70 $\pm$ 13.31                    | 87.5 $\pm$ 7.8                   | <b>0.88<math>\pm</math>0.02</b> |
| <i>Vision-Aligned:</i><br>Transcript Only | Qwen2.5-7B-Inst.      | 81.25 $\pm$ 5.71                 | 79.8 $\pm$ 8.22                  | 84.52 $\pm$ 2.49                 | <b>76.67<math>\pm</math>13.69</b> | 85.83 $\pm$ 3.73                 | 0.91 $\pm$ 0.02                 |
|                                           | Bio-ClinicalBERT      | 71.67 $\pm$ 3.16                 | 64.67 $\pm$ 5.68                 | <b>85.19<math>\pm</math>0.73</b> | 52.5 $\pm$ 8.12                   | <b>90.83<math>\pm</math>1.86</b> | 0.87 $\pm$ 0.01                 |
|                                           | Gemma-3-4B-IT         | <b>85.83<math>\pm</math>2.72</b> | <b>86.19<math>\pm</math>2.53</b> | 84.28 $\pm$ 3.8                  | 88.33 $\pm$ 3.49                  | 83.33 $\pm$ 5.1                  | 0.9 $\pm$ 0.01                  |
|                                           | MedGemma-4B-IT        | 80.83 $\pm$ 3.42                 | 80.07 $\pm$ 4.8                  | 83.56 $\pm$ 6.78                 | 78.33 $\pm$ 11.93                 | 83.33 $\pm$ 10.62                | 0.89 $\pm$ 0.02                 |
|                                           | PuBMedBERT            | 75.42 $\pm$ 3.73                 | 71.35 $\pm$ 5.27                 | 85.14 $\pm$ 4.24                 | 61.67 $\pm$ 6.85                  | 89.17 $\pm$ 3.73                 | 0.87 $\pm$ 0.02                 |
|                                           | BioGPT                | 74.17 $\pm$ 11.47                | 69.67 $\pm$ 15.54                | 81.9 $\pm$ 14.51                 | 62.5 $\pm$ 20.41                  | 85.83 $\pm$ 12                   | 0.81 $\pm$ 0.15                 |
|                                           | Mistral-7B-Inst.-v0.2 | 80.83 $\pm$ 1.74                 | 80.95 $\pm$ 1.31                 | 81.32 $\pm$ 6.57                 | 81.67 $\pm$ 8.64                  | 80 $\pm$ 11.18                   | 0.88 $\pm$ 0.01                 |
|                                           | LLaMA-3.2-1B-Inst.    | 84.58 $\pm$ 4.32                 | 83.78 $\pm$ 6.06                 | 87.04 $\pm$ 3.34                 | 81.67 $\pm$ 11.64                 | 87.5 $\pm$ 5.1                   | 0.89 $\pm$ 0.02                 |
|                                           | LLaMA-3.2-3B-Inst.    | 80.42 $\pm$ 3.16                 | 81.1 $\pm$ 3.02                  | 78.64 $\pm$ 4.88                 | 84.17 $\pm$ 6.18                  | 76.67 $\pm$ 6.97                 | 0.86 $\pm$ 0.01                 |
|                                           | LLaMA-3.1-8B-Inst.    | 84.58 $\pm$ 1.14                 | 84.46 $\pm$ 1.74                 | 85.03 $\pm$ 2.29                 | 84.17 $\pm$ 5.43                  | 85 $\pm$ 3.73                    | 0.88 $\pm$ 0.01                 |
| <i>Vision-Aligned:</i><br>Neutral         | Qwen2.5-7B-Inst.      | 81.67 $\pm$ 3.09                 | 81.28 $\pm$ 4.44                 | 82.50 $\pm$ 3.05                 | 80.83 $\pm$ 10.03                 | 82.50 $\pm$ 5.43                 | 0.90 $\pm$ 0.00                 |
|                                           | Bio-ClinicalBERT      | 75.83 $\pm$ 1.14                 | 71.52 $\pm$ 2.12                 | <b>86.99<math>\pm</math>1.59</b> | 60.83 $\pm$ 3.73                  | <b>90.83<math>\pm</math>1.86</b> | 0.87 $\pm$ 0.04                 |
|                                           | Gemma-3-4B-IT         | 78.75 $\pm$ 2.72                 | 78.62 $\pm$ 3.07                 | 79.01 $\pm$ 2.55                 | 78.33 $\pm$ 4.56                  | 79.17 $\pm$ 2.95                 | 0.87 $\pm$ 0.02                 |
|                                           | MedGemma-4B-IT        | 80.42 $\pm$ 1.86                 | 79.42 $\pm$ 3.51                 | 83.49 $\pm$ 4.29                 | 76.67 $\pm$ 10.03                 | 84.17 $\pm$ 6.85                 | 0.88 $\pm$ 0.01                 |
|                                           | PuBMedBERT            | 75.00 $\pm$ 6.42                 | 70.50 $\pm$ 9.75                 | 84.16 $\pm$ 4.39                 | 61.67 $\pm$ 13.63                 | 88.33 $\pm$ 4.56                 | 0.87 $\pm$ 0.01                 |
|                                           | BioGPT                | 75.83 $\pm$ 11.56                | 71.96 $\pm$ 14.08                | 84.36 $\pm$ 14.63                | 63.33 $\pm$ 15.42                 | 88.33 $\pm$ 10.37                | 0.81 $\pm$ 0.15                 |
|                                           | Mistral-7B-Inst.-v0.2 | 79.58 $\pm$ 1.74                 | 80.77 $\pm$ 0.58                 | 77.08 $\pm$ 6.05                 | <b>85.83<math>\pm</math>7.57</b>  | 73.33 $\pm$ 10.87                | 0.87 $\pm$ 0.01                 |
|                                           | LLaMA-3.2-1B-Inst.    | <b>84.58<math>\pm</math>2.38</b> | <b>84.79<math>\pm</math>2.10</b> | 84.02 $\pm$ 4.72                 | 85.83 $\pm$ 3.73                  | 83.33 $\pm$ 5.89                 | 0.88 $\pm$ 0.01                 |
|                                           | LLaMA-3.2-3B-Inst.    | 76.67 $\pm$ 6.49                 | 75.05 $\pm$ 10.54                | 79.62 $\pm$ 8.02                 | 75 $\pm$ 21.45                    | 78.33 $\pm$ 15.42                | 0.86 $\pm$ 0.03                 |
|                                           | LLaMA-3.1-8B-Inst.    | 79.58 $\pm$ 5.39                 | 77.89 $\pm$ 8.21                 | 83.01 $\pm$ 3.22                 | 75 $\pm$ 15.87                    | 84.17 $\pm$ 6.18                 | 0.88 $\pm$ 0.02                 |
| <i>Vision-Aligned:</i><br>Clinical        | Qwen2.5-7B-Inst.      | 80.83 $\pm$ 0.93                 | 79.83 $\pm$ 2.56                 | 84.60 $\pm$ 6.57                 | 76.67 $\pm$ 9.13                  | 85.00 $\pm$ 8.12                 | 0.89 $\pm$ 0.02                 |
|                                           | Bio-ClinicalBERT      | 75.83 $\pm$ 1.86                 | 71.24 $\pm$ 2.90                 | <b>87.77<math>\pm</math>0.70</b> | 60.00 $\pm$ 3.73                  | <b>91.67<math>\pm</math>0.00</b> | 0.86 $\pm$ 0.04                 |
|                                           | Gemma-3-4B-IT         | 83.33 $\pm$ 2.08                 | 83.84 $\pm$ 2.20                 | 81.48 $\pm$ 3.65                 | 86.67 $\pm$ 5.43                  | 80.00 $\pm$ 5.43                 | 0.89 $\pm$ 0.02                 |
|                                           | MedGemma-4B-IT        | 75.42 $\pm$ 12.36                | 76.32 $\pm$ 10.20                | 75.68 $\pm$ 12.74                | 78.33 $\pm$ 11.93                 | 72.50 $\pm$ 20.11                | 0.83 $\pm$ 0.16                 |
|                                           | PuBMedBERT            | 75.83 $\pm$ 6.00                 | 71.00 $\pm$ 10.41                | 86.39 $\pm$ 4.38                 | 61.67 $\pm$ 14.85                 | 90.00 $\pm$ 4.75                 | 0.87 $\pm$ 0.02                 |
|                                           | BioGPT                | 70.83 $\pm$ 9.55                 | 64.62 $\pm$ 12.04                | 82.49 $\pm$ 15.94                | 54.17 $\pm$ 13.82                 | 87.50 $\pm$ 11.79                | 0.81 $\pm$ 0.17                 |
|                                           | Mistral-7B-Inst.-v0.2 | 83.75 $\pm$ 4.01                 | 83.76 $\pm$ 4.19                 | 83.80 $\pm$ 5.03                 | 84.17 $\pm$ 7.45                  | 83.33 $\pm$ 7.22                 | 0.89 $\pm$ 0.02                 |
|                                           | LLaMA-3.2-1B-Inst.    | <b>86.67<math>\pm</math>2.80</b> | <b>86.58<math>\pm</math>3.21</b> | 86.86 $\pm$ 3.35                 | <b>86.67<math>\pm</math>6.85</b>  | 86.67 $\pm$ 4.56                 | <b>0.90<math>\pm</math>0.01</b> |
|                                           | LLaMA-3.2-3B-Inst.    | 82.5 $\pm$ 2.38                  | 81.70 $\pm$ 2.69                 | 85.56 $\pm$ 3.19                 | 78.33 $\pm$ 4.56                  | 86.67 $\pm$ 3.49                 | 0.87 $\pm$ 0.01                 |
|                                           | LLaMA-3.1-8B-Inst.    | 82.92 $\pm$ 4.00                 | 81.76 $\pm$ 5.75                 | 86.51 $\pm$ 2.70                 | 78.33 $\pm$ 11.18                 | 87.5 $\pm$ 4.17                  | 0.87 $\pm$ 0.03                 |
| <i>Vision-Aligned:</i><br>Linguistic      | LLaMA-3.2-1B-Inst.    | <b>81.67<math>\pm</math>5.39</b> | <b>81.47<math>\pm</math>6.22</b> | 81.66 $\pm$ 4.10                 | <b>81.67<math>\pm</math>9.59</b>  | 81.67 $\pm$ 4.75                 | <b>0.89<math>\pm</math>0.01</b> |
|                                           | LLaMA-3.2-3B-Inst.    | 77.5 $\pm$ 4.52                  | 75.55 $\pm$ 6.14                 | 83.14 $\pm$ 7.79                 | 70.83 $\pm$ 12.84                 | 84.17 $\pm$ 12.29                | 0.86 $\pm$ 0.01                 |
|                                           | LLaMA-3.1-8B-Inst.    | 79.17 $\pm$ 2.56                 | 77.38 $\pm$ 4.64                 | <b>84.20<math>\pm</math>3.73</b> | 72.5 $\pm$ 10.46                  | <b>85.83<math>\pm</math>6.32</b> | 0.88 $\pm$ 0.01                 |
| <i>Linguistic Feat.:</i><br>Numeric       | LLaMA-3.2-1B-Inst.    | 82.08 $\pm$ 7.88                 | 80.35 $\pm$ 13.18                | 83.88 $\pm$ 4.83                 | 80.83 $\pm$ 21.97                 | 83.33 $\pm$ 7.8                  | 0.89 $\pm$ 0.01                 |
|                                           | LLaMA-3.2-3B-Inst.    | 80.83 $\pm$ 2.72                 | 80.04 $\pm$ 5.16                 | 82.85 $\pm$ 5.25                 | 79.17 $\pm$ 13.82                 | 82.5 $\pm$ 8.54                  | 0.87 $\pm$ 0.01                 |
|                                           | LLaMA-3.1-8B-Inst.    | <b>82.92<math>\pm</math>4.52</b> | <b>81.82<math>\pm</math>5.93</b> | <b>86.58<math>\pm</math>4.62</b> | 78.33 $\pm$ 10.79                 | <b>87.5<math>\pm</math>5.89</b>  | 0.90 $\pm$ 0.02                 |
| <i>Linguistic Feat.:</i><br>Language      | LLaMA-3.2-1B-Inst.    | <b>85.00<math>\pm</math>0.93</b> | <b>85.69<math>\pm</math>1.32</b> | 81.85 $\pm$ 0.83                 | <b>90.00<math>\pm</math>3.73</b>  | 80 $\pm$ 1.86                    | 0.88 $\pm$ 0.01                 |
|                                           | LLaMA-3.2-3B-Inst.    | 80.42 $\pm$ 3.16                 | 81.93 $\pm$ 2.01                 | 76.57 $\pm$ 4.56                 | 88.33 $\pm$ 1.86                  | 72.5 $\pm$ 8.12                  | 0.87 $\pm$ 0.01                 |
|                                           | LLaMA-3.1-8B-Inst.    | 84.17 $\pm$ 4.32                 | 83.34 $\pm$ 4.47                 | <b>88.23<math>\pm</math>5.99</b> | 79.17 $\pm$ 5.1                   | <b>89.17<math>\pm</math>6.32</b> | 0.89 $\pm$ 0.016                |