What Do LLMs Know About Alzheimer's Disease? Multi-loss Fine-Tuning and Probing for AD Detection

Anonymous ACL submission

Abstract

Reliable early detection of Alzheimer's disease (AD) is challenging, particularly due to the limited availability of labeled data. While large language models (LLMs) have shown strong transfer capabilities across do mains, adapting them to the AD domain through supervised fine-tuning remains largely unexplored. In this work, we empirically evaluate various model architectures across three heterogeneous transcript corpora (Pitt, CCC, ADRC) to investigate their effectiveness for text-based AD detection and analyze how task-relevant information is encoded within their internal representations. To the best of our knowledge, our fine-tuned BERT and T5 models establish a new state-of-the-art on the Pitt and CCC datasets, while achieving strong performance on ADRC. In parallel, the decoder-only Llama-1B achieves highly competitive results comparable to BERT and T5 across all three corpora, highlighting its effectiveness for AD detection. We further conduct a comprehensive evaluation of the Llama-1B backbone, analyzing cross-corpus transferability, optimal input chunk-size granularity, and the impact of clinical transcript markers. Also, we use linear probing to empirically show that fine-tuning shifts the representations of individual tokens, both linguistic markers and content words, in ways that reflect AD-related signal.

1 Introduction

Alzheimer's disease (AD) leads to progressive cognitive decline and poses a major burden on patients and caregivers (Skaria, 2022), and early detection of AD is increasingly crucial to enable timely intervention and improve patient outcomes as populations age globally (de la Fuente Garcia et al., 2020). Recently, large language models (LLMs) have demonstrated success at a myriad of downstream tasks through fine-tuning on task-relevant datasets, allowing them to combine the wealth of knowledge learned during pre-training with more task-specific details (Hu et al., 2022; Dettmers et al., 2023). This creates promise for advanced, AI-enabled solutions to challenging problem domains, including those related to healthcare problems growing in prevalence, such as AD (Farzana and Parde, 2024; Han et al., 2025; Li et al., 2025a).

Current methods for AI-enabled AD detection rely on standardized linguistic and cognitive assessments and face challenges in accuracy and scalability (Martinc and Pollak, 2020; Balagopalan et al., 2020; Yuan et al., 2020; Li et al., 2021; Rohanian et al., 2021; Farzana and Parde, 2022, 2023). The rapid advancement of LLMs presents opportunity to leverage their strong linguistic understanding for this clinical application, but fine-tuning language models within the AD domain is underexplored (Zhang et al., 2025; Hou et al., 2025; Li et al., 2025b; Dhinagar et al., 2025), in large part due to the limited availability of labeled clinical data (Duan et al., 2023). Addressing these challenges can not only enhance diagnostic tools for AD but can also contribute to broader research on how LLMs can be effectively adapted to specialized, low-resource clinical tasks without compromising their general language understanding.

Research on other language tasks has demonstrated that supervised fine-tuning (SFT) allows models to learn directly from curated examples that reflect the desired outputs, improving reliability, consistency, and alignment with domain expectations or usage goals (Wei et al., 2022b; Harada et al., 2025). This process not only enhances task performance and reduces unpredictable behavior, but also enables efficient domain adaptation, safety and policy alignment, and personalization without requiring that the model is retrained from scratch (Chung et al., 2024; OpenAI et al., 2024; Ouyang et al., 2022; Peng et al., 2023).

In this work, we comprehensively evaluate three language model architectures for AD detection, as-

sessing their capacity to encode clinically meaningful linguistic markers. Within our multi-loss framework, encoder and encoder-decoder architectures such as BERT(Devlin et al., 2019) and T5(Raffel et al., 2023) yield superior classification performance—achieving, to the best of our knowledge, state-of-the-art results on the Pitt (Becker et al., 1994) and CCC (Pope and Davis, 2011) corpora, alongside strong performance on the SLaCAD (Farzana et al., 2024) dataset. Importantly, the decoder-only Llama3.2-1B-Instruct(Grattafiori et al., 2024)(refer to it as Llama-1B throughout the paper) achieves highly competitive results comparable to BERT and T5 across all three heterogeneous corpora, motivating further investigation into its broader diagnostic potential. While BERT and T5 excel at raw classification performance, larger decoder-only architectures offer strong contextual modeling capabilities that may benefit future clinical applications. Although exploring these capabilities falls outside the scope of this study, our extensive evaluation establishes an important baseline for future work on scalable, multi-corpus clinical AD detection workflows.

Our contributions are fourfold. **(1)** We propose a multi-loss SFT recipe for a Llama-1B backbone that achieves strong AD-detection performance across three heterogeneous transcript corpora (Pitt, CCC, SLaCAD), demonstrating that a small open-weight LLM can match or exceed prior specialized models on this task. **(2)** Through a per-corpus loss-term ablation, we show that all three corpora contribute complementary signal: the full joint objective is best on every per-corpus metric, and removing any single $\mathcal{L}_d$ degrades performance on the remaining corpora as well as on the dropped one. **(3)** We identify chunk-size selection as a critical preprocessing choice: aligning the chunk size to the length scale of the shortest representative segment (512 tokens in our setting) gives all training examples a comparable input granularity and improves F_1 uniformly across corpora. **(4)** We show that the CHAT/CLAN markers carried by the Pitt prompt are a transferable AD-domain feature. The markers explicitly tag the disfluencies, retraces, and word-finding pauses that other corpora leave unannotated, and retaining them improves generalization on *both* Pitt and unmarked CCC, indicating that a small but genuine AD-domain signal is extracted from Pitt and reused on inputs that lack the markers. **5** We use a layer-wise linear probing analysis on LLM representations and empirically observe that, after fine-tuning, the representations of individual tokens, including both linguistic markers and content words, shift in ways that may reflect AD-related signal.

2 Related Work

2.1 Alzheimer's Disease Detection

A relatively large body of work within the NLP community has studied AD detection from different angles (Farzana and Parde, 2023; Zhu et al., 2024; Gkoumas et al., 2023; Li et al., 2022), with prior work focusing mainly on prompt construction (Wang et al., 2023; Farzana and Parde, 2024), model choice (Di Palo and Parde, 2019), and multi-step system design (Ye et al., 2021; Wang et al., 2022). Some work has leveraged speech-based approaches that extract acoustic and prosodic features to distinguish AD patients from healthy controls (Ding et al., 2024; El Hallani et al., 2025), while others have used speech–language hybrid methods combining linguistic and acoustic cues to better capture cognitive decline (Shi et al., 2023). Other work has explored language-only LLM frameworks that analyze discourse structure and lexical patterns using pretrained models (Zhang et al., 2025; Hou et al., 2025; Li et al., 2025b) and vision–language models for multimodal detection (Dhinagar et al., 2025).

In general, these studies emphasize system design and performance, with the aim of gradually improving AD detection using standardized metrics and benchmark datasets. They have paid less attention to how AD-related information is encoded within model representations. Our work seeks to fill this gap.

2.2 Supervised Fine-tuning and Linear Probing

SFT is widely used for adapting pretrained LLMs to downstream tasks by training them on task-specific labeled examples. Early work demonstrated that fine-tuning pretrained Transformers can yield substantial performance gains across NLP benchmarks compared with training models from scratch (Devlin et al., 2019; Radford et al., 2019). Subsequent research refined SFT techniques to improve generalization, stability, and data efficiency, including by tuning instruction prompts to align models with human-written prompts and task formats (Chung et al., 2022; Wei et al., 2022a). SFT has also been used as a core component in multi-

Corpus	Elicitation	Transcription	Total	Ctrl : AD
Pitt	Cookie Theft picture description	CHAT/CLAN with disfluency markers	1,289	19% : 81%
CCC	Open-domain dyadic interview	Orthographic, no markup	105	46% : 54%
SLaCAD	Autobiographical narrative	Orthographic, no markup	91	90% : 10%

Table 1: Corpora differ in genre, annotation richness, and direction of class imbalance.

stage alignment pipelines, where models are first fine-tuned on curated supervised datasets before further improvement via reinforcement learning from human feedback (RLHF) or preference optimization (Ouyang et al., 2022; Bai et al., 2022). Recent work further explores domain-specific SFT for specialized applications, such as biomedical and clinical language tasks, demonstrating that carefully selected, high-quality supervision can enable strong downstream adaptation even in low-resource settings (Tran et al., 2024; Singhal et al., 2023).

To better understand how supervised fine-tuning shapes internal representations, we next turn to probing as a lightweight and interpretable analysis tool. Linguistic and stylistic concepts often manifest as linear features within the high-dimensional spaces of language model representations. Linguistic and stylistic concepts often manifest as linear features within the high-dimensional spaces of language model representations (Mikolov et al., 2013; Nanda et al., 2023; Park et al., 2024; Gurnee and Tegmark, 2024; Kim et al., 2025). In our work, we use linear probes to estimate the direction of AD in the model's representations and to measure the extent to which AD-related information is captured at the token level.

3 Methodology

In this work, we first explore the performance of SFT with LLMs for AD detection. Additionally, we investigate how the concept of AD is encoded in LLMs' internal representations. Building on evidence that concepts are linearly encoded in representation and can be detected, we train linear probes on LLM hidden states to predict AD-specific labels, yielding probes that isolate features in the representations relevant to AD. Using these probes, we examine token-level activations in AD transcripts and identify the tokens that are most informative for downstream AD classification when processed by the models.

3.1 Multi-Loss Learning across Heterogeneous Corpora

As shown in Table 1, our three corpora, DementiaBank's Pitt Corpus (Becker et al., 1994), Carolinas Conversation Collection (CCC) (Pope and Davis, 2011), and SLaCAD (Farzana et al., 2024), share the same diagnostic target (AD vs. Control) but differ markedly in transcription convention, elicitation protocol, and class balance. We phrase the entire fine-tuning objective as a sum of per-corpus losses, each conditioned on a corpus-specific prompt template and a corpus-specific class weighting scheme. The same formulation is instantiated across three model families: encoder-only (BERT), encoder–decoder (T5), and decoder-only LLMs (Llama-1B).

Through the remainder of this discussion, we use the following formalisms. We let $\mathcal{D} = \{\text{Pitt}, \text{CCC}, \text{SLaCAD}\}$. Correspondingly, we let $\mathcal{T}_d = \{(x_i, y_i)\}_{i=1}^{N_d}$ denote the training split of corpus $d \in \mathcal{D}$, with binary label $y_i \in \{\text{Control}, \text{AD}\}$ shared across $\mathcal{D}$ and $d(i)$ the corpus of example i.

Corpus Conditional Prompt Templates. DementiaBank transcripts follow the CHAT/CLAN convention and embed annotation markers (e.g., retracing `[/]`, revision `[//]`, replacement `[: . . . ]`, filled pauses `&-uh`, `&-um`) that are themselves linguistically informative cues for AD detection. A more detailed example is shown in Appendix A. CCC is composed of cleaned conversational interviews with no such markup, and SLaCAD contains structured language-test recordings (autobiographical interview, normal read-aloud, and autocorrect-cloze paragraph) that are also marker-free.

To preserve this corpus-specific information, we use a distinct user prompt $P_d(\cdot)$ for each corpus. The Pitt prompt embeds a glossary of CHAT/CLAN markers together with marker-aware diagnostic criteria, whereas the CCC and SLaCAD prompts list only generic linguistic criteria (word-finding difficulty, repetition, reduced coherence, and simpler or broken grammar). All three templates terminate in the same supervi-

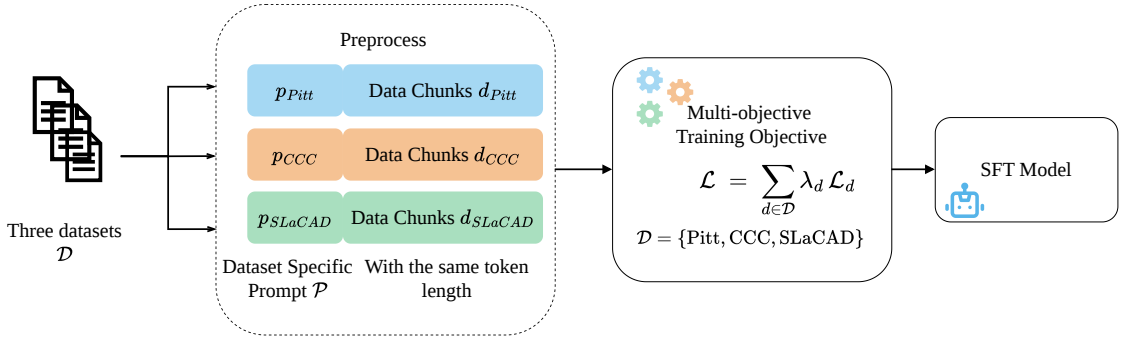

Figure 1: The pipeline of multi-objective fine-tuning for language models.

sion slot. This means that the gold answer string $y \in \{\texttt{Control}, \texttt{AD}\}$ is identical across corpora. The corpus-conditional input is

$$\tilde{x}_i = P_{d(i)}(x_i), \tag{1}$$

Per-Corpus Class-Weighted Aggregation. The three corpora have very different class priors: Pitt is AD-majority, CCC is mildly Control-leaning, and SLaCAD is strongly Control-majority with only a handful of AD examples. A uniform mean over the merged pool would therefore steer gradient updates toward whichever class is over-represented in the largest corpus. We associate each corpus with its own class-weight vector $w_d = (w_{d,\text{Ctl}}, w_{d,\text{AD}}) \in \mathbb{R}^2$, fixed from the inverse class frequency of $\mathcal{T}_d$ in isolation, and aggregate the per-example loss into a per-corpus loss by

$$\mathcal{L}_d = \frac{1}{Z_d} \sum_{i \in \mathcal{T}_d} w_{d,y_i} \, \ell_\theta\big(P_d(x_i), y_i\big), \tag{2}$$

where $Z_d = \sum_{i \in \mathcal{T}_d} w_{d,y_i}$ normalizes the weighted sum to a weighted mean, making $\mathcal{L}_d$ comparable across corpora of different sizes and classes.

Joint Multi-Objective Loss. The full fine-tuning objective is the sum of the three per-corpus losses,

$$\mathcal{L} = \sum_{d \in \mathcal{D}} \lambda_d \, \mathcal{L}_d, \qquad \lambda_d \geq 0, \tag{3}$$

with corpus mixing coefficients λ_d. We realize Eq. (3) at the data preprocessing level by upsampling the smaller corpora to match the size of a reference corpus, which is Pitt in our situation: we repeat CCC and SLaCAD examples with replacement until each contributes as many training rows as Pitt(1,041 in our splits), so each $\mathcal{L}_d$ receives equal exposure per epoch.

Standard SFT (Cross-Entropy Loss). Standard SFT employs cross-entropy (CE) loss to maximize the likelihood of the ground-truth labels:

$$\mathcal{L}_{\text{CE}} = -\sum_{i=1}^{C} y_i \log(p_i), \tag{4}$$

where y_i denotes the one-hot encoded ground-truth label and p_i is the predicted probability for class i.

Focal Loss. To explicitly down-weight easy, confidently classified examples and concentrate the gradient on hard or minority-class ones, we replace $\mathcal{L}_{\text{CE}}$ with the focal cross-entropy of ():

$$\mathcal{L}_{\text{focal}} = -\sum_{i=1}^{C} (1 - p_i)^\gamma \, y_i \log(p_i), \tag{5}$$

where the modulating factor $(1 - p_i)^\gamma$ vanishes as the model assigns high probability to the correct class, and $\gamma \geq 0$ controls the rate of this attenuation; we fix $\gamma = 2$ throughout.

Label Smoothing. Label smoothing () softens the one-hot target by mixing it with a uniform distribution over the C classes,

$$\begin{aligned} \tilde{y}_i &= (1 - \alpha)\, y_i + \frac{\alpha}{C}, \\ \mathcal{L}_{\text{LS}} &= -\sum_{i=1}^{C} \tilde{y}_i \log(p_i), \end{aligned} \tag{6}$$

where $\alpha \in [0, 1)$ controls the smoothing strength; we set $\alpha = 0.1$. Minimising $\mathcal{L}_{\text{LS}}$ instead of $\mathcal{L}_{\text{CE}}$ penalises overconfident next-token predictions on the ground truth label (*AD* or *Control*). This is useful in AD-detection SFT because the training corpora differ in task, annotation, and class balance; Smoothing discourages the model from overfitting to corpus-specific patterns and produces more reliable label predictions at generation time.

3.2 Probing

After studying SFT performance for AD detection, we ask whether AD-relevant linguistic information is linearly recoverable from the model's internal states. Following Gurnee and Tegmark (2024), we fit a per-layer ridge probe $\hat{y} = \mathbf{W}^\top h$ that maps a hidden representation $h \in \mathbb{R}^d$ to the AD label $y \in \{0, 1\}$ (0 = Control, 1 = AD), with objective

$$\ell(\mathbf{W}) = \frac{1}{N} \sum_{i=1}^{N} \left(y_i - \mathbf{W}^\top h_i\right)^2 + \lambda_{\text{ridge}} \|\mathbf{W}\|_2^2, \tag{7}$$

where N is the number of training samples and λ_{ridge} controls the L2 penalty. The fitted $\mathbf{W}$ defines a linear direction for AD in representation space, and $\hat{y}_i$ measures how strongly that concept is expressed in h_i.

4 Experiments

4.1 Experimental Settings

Hyperparameters. Across all experiments, we train for 10 epochs. The Llama-3.2-1B-Instruct model is fine-tuned with a learning rate of 2e-5, preserving long transcripts without truncation. The T5-large baseline is fine-tuned at a learning rate of 3e-4, while the BERT-base baseline uses a learning rate of 2e-5. To counter the class imbalance, which differs across corpora, we apply a per-task weighted cross-entropy loss with dataset-specific (Control, AD) weights that up-weight each corpus's minority class in inverse proportion to its frequency, using a stronger exponent on SLaCAD to absorb its extreme, inverted imbalance: (2.099, 0.477) for Pitt, (1.277, 0.783) for CCC, and (0.096, 10.429) for SLaCAD. Label smoothing and focal loss additionally introduce a label-smoothing factor of 0.1, focal_alpha=0.25 with focal_gamma=2.0, and a contrastive mixing weight of 0.1 with a margin of 1.0, respectively.

Other Modeling Details. We experiment with each loss function condition on both backbone models across three corpora. The DementiaBank corpus consists of 1,044 AD and 247 Control samples; the CCC corpus consists of 57 AD and 48 Control samples; and the SLaCAD corpus consists of 9 AD and 82 Control samples, exhibiting the opposite class imbalance to DementiaBank. Each corpus is split 80/20 into training and evaluation sets in a stratified manner. We conducted all experiments using four A100 PCIe GPUs, each equipped with 40 GB of memory. Fine-tuning a single model on 4 GPUs takes approximately 6 minutes.

4.2 Supervised Fine-tuning Results

Table 2 compares different loss functions for Llama-1B across the Pitt, CCC, and SLaCAD datasets, using Accuracy (Acc) and F1-score as evaluation metrics. Among all investigated loss functions, the **Weighted loss** consistently achieved the best performance across all datasets. Specifically, it obtained the highest mean Accuracy (0.87) and mean F1-score (0.76), outperforming the Standard, Focal, and Label Smoothing losses.

For the Pitt dataset, the Weighted loss improved the F1-score from 0.73 (Standard loss) to 0.83, indicating a substantial gain in classification balance and robustness. Similarly, on the CCC dataset, the Weighted loss achieved the best Accuracy (0.83) and F1-score (0.82), demonstrating its effectiveness in handling class imbalance. The largest relative improvement was observed on the SLaCAD dataset, where the F1-score increased from 0.46 under Standard loss to 0.63 using Weighted loss.

In contrast, the Focal loss consistently underperformed compared to the other approaches, yielding the lowest mean Accuracy (0.73) and mean F1-score (0.58). Label smoothing provided moderate improvements over Focal loss but did not surpass the Standard or Weighted configurations.

Compared with larger baseline models, T5-large achieved the best overall performance, with a mean Accuracy of 0.95 and a mean F1-score of 0.89. BERT-base also outperformed Llama-1B in most settings, particularly on the Pitt and CCC datasets. However, the Weighted-loss Llama-1B significantly narrowed the performance gap, suggesting that appropriate loss design can substantially improve the performance of smaller language models.

Summary of SFT Results Overall, the experimental results demonstrate that applying a Weighted loss function is the most effective strategy for improving Llama-1B performance across all evaluated datasets. The consistent gains in both Accuracy and F1-score suggest that class imbalance is a critical factor in these tasks, and weighting the loss successfully mitigates this issue. Although larger models such as T5-large still achieve superior overall results, the optimized Llama-1B configuration offers a competitive and computationally efficient alternative.

Model	Loss Type	Pitt		CCC		SLaCAD		Mean	
		Acc	F1	Acc	F1	Acc	F1	Acc	F1
Llama-1B	Vanilla	$0.78_{\pm 0.00}$	$0.44_{\pm 0.00}$	$0.61_{\pm 0.00}$	$0.50_{\pm 0.00}$	$0.26_{\pm 0.00}$	$0.24_{\pm 0.00}$	0.55	0.39
	Standard	$0.84_{\pm 0.01}$	$0.73_{\pm 0.02}$	$0.74_{\pm 0.02}$	$0.70_{\pm 0.02}$	$0.87_{\pm 0.05}$	$0.46_{\pm 0.01}$	0.82	0.63
	Weighted	$\mathbf{0.88}_{\pm 0.02}$	$\mathbf{0.83}_{\pm 0.02}$	$\mathbf{0.83}_{\pm 0.02}$	$\mathbf{0.82}_{\pm 0.02}$	$\mathbf{0.89}_{\pm 0.02}$	$\mathbf{0.63}_{\pm 0.02}$	**0.87**	**0.76**
	Focal	$0.77_{\pm 0.02}$	$0.68_{\pm 0.02}$	$0.63_{\pm 0.02}$	$0.62_{\pm 0.02}$	$0.78_{\pm 0.09}$	$0.44_{\pm 0.03}$	0.73	0.58
	Label smoothing	$0.79_{\pm 0.01}$	$0.75_{\pm 0.02}$	$0.69_{\pm 0.04}$	$0.69_{\pm 0.04}$	$0.69_{\pm 0.02}$	$0.48_{\pm 0.01}$	0.73	0.64
T5-large	Standard	0.92	0.89	0.96	0.95	0.97	0.83	0.95	0.89
BERT-base	Standard	0.91	0.87	0.93	0.92	0.95	0.49	0.93	0.76

Table 2: Accuracy and macro-F1 on Pitt, CCC, and SLaCAD. T5/BERT are single deterministic runs. "Standard" denotes uniform per-task, per-class loss weights, all set to 1.

Impact of Class Imbalance on SLaCAD. The SLaCAD dataset is highly imbalanced, containing only 7 Alzheimer's disease (AD) samples while the remaining subjects belong to the control class. This imbalance explains the discrepancy between Accuracy and F1-score, as models can achieve high Accuracy by favoring the majority class while failing to correctly identify AD samples. For example, although Standard loss achieved relatively high Accuracy on SLaCAD, the corresponding F1-scores remained low. In contrast, the Weighted loss substantially improved the F1-score, indicating better recognition of the minority AD class and demonstrating the importance of class-balanced optimization for highly skewed datasets.

5 Ablation Studies

We ablate three design choices that together specify the multi-loss objective in Eq. (3): (i) which per-corpus loss terms $\mathcal{L}_d$ are active, (ii) how each corpus is preprocessed before being fed to the model, and (iii) whether the corpus-conditional prompt P_d for Pitt should preserve its CHAT/CLAN markers. Unless otherwise noted, every ablation uses the same backbone, the same per-corpus class weights w_d, the same upsample-to-reference mixing policy ($\lambda_d = 1$), and the same held-out test splits; only the variable under study changes. We report per-corpus macro-F_1 on Pitt, CCC, and SLaCAD and the mean macro-F_1 across corpora.

5.1 Per-Corpus Loss-Term Ablation

This study tests whether the three per-corpus losses are complementary or whether a single dominant corpus accounts for the model's behavior. We enable every non-empty subset $\mathcal{S} \subseteq \{\text{Pitt}, \text{CCC}, \text{SLaCAD}\}$ in turn and train with $\mathcal{L}^{(\mathcal{S})} = \sum_{d \in \mathcal{S}} \mathcal{L}_d$, holding all other components fixed. Concretely, in each variant, the corpora in $\mathcal{S}$ contribute their full per-corpus weighted cross-entropy from Eq. (2), while the remaining corpora are removed from the training mix entirely. Table 3 reports the seven resulting models.

Findings. All corpora contribute complementary signal. Activating all three losses yields the best results across all per-corpus metrics. The single-corpus models confirm that Pitt is the only corpus whose training signal alone produces a reasonable Pitt F_1 (0.804), but it still leaves both CCC and SLaCAD near majority-class performance. Crucially, removing Pitt from the loss (CCC+SLaCAD) collapses every metric, including SLaCAD's own F_1; this shows that the abundant, AD-rich Pitt signal acts as an anchor through which the smaller corpora's weighted terms become learnable. The result supports keeping all three $\mathcal{L}_d$ in the joint objective rather than deferring CCC or SLaCAD to a downstream fine-tune.

5.2 Chunking

Table 4 shows that reducing the chunk size from 1024 to 512 consistently improves performance across all datasets for the Llama-1B model. On the Pitt dataset, accuracy increases from 0.83 to 0.88, while the F1 score improves substantially from 0.69 to 0.83. Similar gains are observed on the CCC dataset, where accuracy rises slightly from 0.82 to 0.83, and F1 improves from 0.79 to 0.82. The largest improvement is seen on the SLaCAD dataset: accuracy increases from 0.86 to 0.89, and

Active Loss Terms	Pitt		CCC		SLaCAD	
	Acc	F1	Acc	F1	Acc	F1
Pitt + CCC + SLaCAD	$\mathbf{0.88 \pm 0.02}$	$\mathbf{0.83 \pm 0.02}$	0.83 ± 0.02	0.82 ± 0.02	0.89 ± 0.02	$\mathbf{0.63 \pm 0.02}$
Pitt + CCC	0.85 ± 0.01	0.79 ± 0.02	$\mathbf{0.85 \pm 0.03}$	$\mathbf{0.84 \pm 0.03}$	0.76 ± 0.03	0.49 ± 0.05
Pitt + SLaCAD	0.86 ± 0.01	0.81 ± 0.02	0.66 ± 0.02	0.57 ± 0.03	$\mathbf{0.90 \pm 0.02}$	0.47 ± 0.00
CCC + SLaCAD	0.68 ± 0.01	0.48 ± 0.02	0.64 ± 0.01	0.42 ± 0.03	0.17 ± 0.07	0.16 ± 0.06
Pitt only	0.86 ± 0.00	0.80 ± 0.01	0.57 ± 0.01	0.53 ± 0.02	0.57 ± 0.11	0.42 ± 0.04
CCC only	0.71 ± 0.02	0.48 ± 0.04	0.56 ± 0.05	0.49 ± 0.07	0.44 ± 0.05	0.33 ± 0.02
SLaCAD only	0.54 ± 0.01	0.49 ± 0.00	0.46 ± 0.07	0.45 ± 0.08	0.75 ± 0.05	0.54 ± 0.03

Table 3: Loss-term ablation for Llama-1B. Pitt + CCC + SLaCAD is the full version with all three loss terms.

Chunk Size	Pitt		CCC		SLaCAD	
	Acc	F1	Acc	F1	Acc	F1
1024	0.83 ± 0.01	0.69 ± 0.03	0.82 ± 0.01	0.79 ± 0.02	0.86 ± 0.03	0.46 ± 0.01
512	$\mathbf{0.88}_{\pm 0.02}$	$\mathbf{0.83}_{\pm 0.02}$	$\mathbf{0.83}_{\pm 0.02}$	$\mathbf{0.82}_{\pm 0.02}$	$\mathbf{0.89}_{\pm 0.02}$	$\mathbf{0.63}_{\pm 0.02}$

Table 4: Chunk-size ablation for the Llama-1B model.

Training Data	Pitt		CCC	
	Acc	F1	Acc	F1
Pitt only	$.85_{.01}$	$.74_{.03}$	$.60_{.02}$	$.45_{.01}$
CCC only	$.73_{.02}$	$.50_{.02}$	$.61_{.03}$	$.44_{.02}$
Pitt+CCC, w/ markers	$\mathbf{.85}_{.01}$	$\mathbf{.79}_{.02}$	$\mathbf{.85}_{.03}$	$\mathbf{.84}_{.03}$
Pitt+CCC, w/o markers	$\mathbf{.85}_{.01}$	$.78_{.01}$	$.84_{.01}$	$.82_{.01}$

Table 5: Marker ablation.

F1 scores from 0.46 to 0.63. Overall, the results indicate that a smaller chunk size of 512 yields more robust, balanced classification performance, particularly in terms of F1 score.

Findings. Chunk size should match the shortest corpus. The intuition behind using a chunk size of 512 is that it creates text segments with more uniform and comparable token lengths across data points, refers to Figure 3. We argue that a similar data segment length can help the model understand this domain better.

5.3 Corpus-Pair Composition and the Role of CHAT/CLAN Markers

In this experiment, we investigate whether the Pitt-specific CHAT/CLAN markers in P_{Pitt} help or hurt cross-corpus transfer. We test both questions at once on the Pitt+CCC pair, which is the only pair in which both corpora are sufficiently large for a stratified test split. We compare four configurations: each corpus trained alone with its own prompt; a balanced Pitt+CCC mix using the canonical P_{Pitt} that exposes CHAT/CLAN markers, and the same balanced mix with Pitt markers stripped and Pitt fed through the no-marker prompt variant (Section 3.1).

Findings. Pitt's CHAT/CLAN markers are a transferable AD-domain feature. First, the balanced Pitt+CCC mix beats both single-corpus models by a wide margin: mean F_1 improves by $+0.22$ over Pitt-only and $+0.35$ over CCC-only, with each joint model gaining most on the corpus it had *not* previously seen (CCC F_1 rises from 0.44 to 0.84 once Pitt is added). This matches the loss-term ablation. Per-corpus losses contribute a non-redundant signal, and the only way to fit both well is to learn a representation invariant to surface differences between the two corpora.

Second, keeping the CHAT/CLAN markers yields a small but consistent gain over the no-marker variant ($+0.01$ Pitt F_1, $+0.02$ CCC F_1, accuracies essentially unchanged). Markers such as `&-um`, `[/]`, and `+...` are unique to Pitt's transcription convention and explicitly tag the disfluencies, retraces, and word-finding pauses that characterize AD speech but are not annotated in any other

corpus we train on; the markers therefore expose a small AD-domain feature the model picks up from Pitt and reuses on unmarked CCC, where the same phenomena are signaled only lexically.

5.4 Summary of Ablations

The three ablations converge on a single recommended configuration. **(1) All corpora contribute complementary signal.** Activating every per-corpus loss $\mathcal{L}_d$ in the joint objective yields the best score on every per-corpus metric; dropping any single $\mathcal{L}_d$, and most dramatically $\mathcal{L}_{\text{Pitt}}$, degrades performance on the remaining corpora as well as on the dropped one. **(2) Chunk size should match the shortest corpus.** Setting the chunking budget to the length scale of the smallest representative segment gives all training examples a comparable input granularity and improves F_1 uniformly across all three corpora. **(3) Pitt's CHAT/CLAN markers are a transferable AD-domain feature.** The diverse marker vocabulary in P_{Pitt} explicitly tags disfluencies, retraces, and word-finding pauses that other corpora leave unannotated; retaining it improves generalization on *both* Pitt and unmarked CCC, indicating that the model extracts a small but genuine AD-domain signal from these markers and reuses it on inputs that lack them.

5.5 Probing Representation

We trained probes by using hidden states from the SOTA Llama-1B to test how well dementia-vs-control information is recoverable from each layer's representation. Each probe is a ridge-regularized linear head on top of the frozen last-token hidden state at the target layer; the underlying language model is kept frozen. Because SLaCAD's extreme class imbalance (≈11:1) makes per-layer probe metrics unstable, we restrict probing to the Pitt and CCC training splits. The results are shown in 4. We chose layer-15 as the probe according to the 4 metrics

Figure 2 illustrates our token-level representation analysis on two example transcripts, showing how fine-tuning reshapes the AD-related signal carried by each token. For every token, we project its layer-15 hidden state onto the linear probe direction learned from the base and fine-tuned models, and visualize the signed difference; color intensity encodes the magnitude of change, with cells below a threshold of 0.02 left unshaded. Two complementary patterns emerge. On the picture-description sample, fine-tuning amplifies the signal

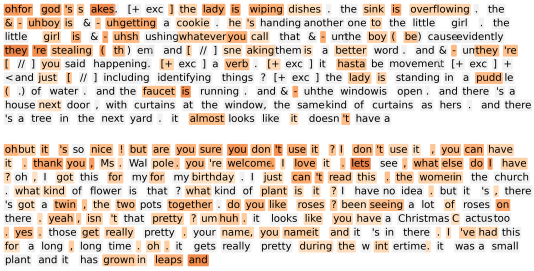

Figure 2: This figure illustrates the differences in token probing values before and after fine-tuning of the transcripts. The upper panel corresponds to the Pitt dataset, while the lower panel corresponds to the CCC dataset.

on CHAT-format disfluency and revision markers (`&-uh`, `&-un`, `[//]`, `[+ exc]`, partial-word fragments such as sneaking and puddle), together with scene-grounded content words (*lady*, *sink*, *wiping*, *overflowing*, *stealing*, *faucet*) and affective exclamations (*oh for god's sakes*). On the conversational sample, the largest shifts concentrate on interactional fillers and backchannels (*oh*, *yeah*, *um huh*, *yes*), repetitive or parallel constructions indicative of perseveration and circumlocution , and high-frequency pronouns and formulaic question scaffolds.

Together, these examples show that the probe isolates a latent representational direction along which AD-related linguistic variation is expressed in the base language model, and that token-level projections along this direction pinpoint the lexical, disfluency, and discourse-level elements the model learns to emphasize or de-emphasize during fine-tuning, yielding interpretable, fine-grained evidence of *which* linguistic phenomena drive the classifier.

6 Conclusion

This work demonstrates that a small, open-weight LLM, when trained with a multi-loss supervised fine-tuning framework, can achieve strong, consistent Alzheimer's disease detection performance across heterogeneous clinical speech corpora. Our results show that jointly leveraging complementary datasets, carefully selecting chunk granularity, and preserving clinically meaningful transcript annotations together improve robustness and generalization. Beyond establishing competitive SOTA performance, this study provides an important foundation for future exploration of generative LLMs as scalable and transferable tools for multi-corpus clinical AD assessment workflows.

7 Limitations

A limitation of this study is the strong class imbalance in SLaCAD, which may affect the stability and generalizability of the reported performance despite the robustness of the proposed multi-loss framework. In addition, although our approach demonstrates strong results across Pitt, CCC, and SLaCAD, the pipeline should be further evaluated on additional external and more diverse datasets to better assess its robustness, transferability, and clinical applicability across different populations, recording conditions, and transcript styles.

Another limitation is that this work primarily focuses on classification performance rather than interpretability or reasoning. While LLMs possess advanced contextual and reasoning capabilities, understanding how the model connects specific linguistic patterns in transcripts to its final AD detection decisions remains largely unexplored. Establishing more transparent reasoning pathways and clinically interpretable evidence for model predictions is an important direction for future research.

References

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, and 12 others. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *Preprint*, arXiv:2204.05862.

Aparna Balagopalan, Benjamin Eyre, Frank Rudzicz, and Jekaterina Novikova. 2020. To bert or not to bert: Comparing speech and language-based approaches for alzheimer's disease detection. In *Interspeech 2020*, pages 2167–2171.

James T. Becker, François Boiler, Oscar L. Lopez, Judith Saxton, and Karen L. McGonigle. 1994. The natural history of alzheimer's disease: Description of study cohort and accuracy of diagnosis. *Archives of Neurology*, 51(6):585–594.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tai, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, and 16 others. 2022. Scaling instruction-finetuned language models. *J. Mach. Learn. Res.*, 25(1).

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, and 16 others. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Sofia de la Fuente Garcia, Craig W. Ritchie, and Saturnino Luz. 2020. Artificial intelligence, speech, and language processing approaches to monitoring alzheimer's disease: A systematic review. *Journal of Alzheimer's Disease*, 78(4):1547–1574. PMID: 33185605.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. In *Advances in Neural Information Processing Systems*, volume 36, pages 10088–10115. Curran Associates, Inc.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Nikhil J. Dhinagar, Pavithra Senthilkumar, Sophia I. Thomopoulos, and Paul M. Thompson. 2025. Maveric-ad: Mixture-of-experts agentic vision-language ensemble for robust mri classification of alzheimer's disease. *bioRxiv*.

Flavio Di Palo and Natalie Parde. 2019. Enriching neural models with targeted features for dementia detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 302–308, Florence, Italy. Association for Computational Linguistics.

Kewen Ding, Madhu Chetty, Azadeh Noori Hoshyar, Tanusri Bhattacharya, and Britt Klein. 2024. Speech based detection of alzheimer's disease: a survey of ai techniques, datasets and challenges. *Artificial Intelligence Review*, 57(12):325.

Junwen Duan, Fangyuan Wei, Jin Liu, Hongdong Li, Tianming Liu, and Jianxin Wang. 2023. CDA: A contrastive data augmentation method for Alzheimer's disease detection. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1819–1826, Toronto, Canada. Association for Computational Linguistics.

Anass El Hallani, Adil Chakhtouna, and Abdellah Adib. 2025. Advanced speech biomarker integration for robust alzheimer's disease diagnosis. *Annals of Telecommunications*, 80(5):427–444.

Shahla Farzana and Natalie Parde. 2022. Are interaction patterns helpful for task-agnostic dementia detection? an empirical exploration. In *Proceedings of the 23rd*

Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 172–182, Edinburgh, UK. Association for Computational Linguistics.

Shahla Farzana and Natalie Parde. 2023. Towards domain-agnostic and domain-adaptive dementia detection from spoken language. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11965–11978, Toronto, Canada. Association for Computational Linguistics.

Shahla Farzana and Natalie Parde. 2024. Domain adaptation via prompt learning for Alzheimer's detection. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15963–15976, Miami, Florida, USA. Association for Computational Linguistics.

Shahla Farzana, Edoardo Stoppa, Alex Leow, Tamar Gollan, Raeanne Moore, David Salmon, Douglas Galasko, Erin Sundermann, and Natalie Parde. 2024. SLaCAD: A spoken language corpus for early Alzheimer's disease detection. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14877–14897, Torino, Italia. ELRA and ICCL.

Dimitris Gkoumas, Matthew Purver, and Maria Liakata. 2023. Reformulating NLP tasks to capture longitudinal manifestation of language disorders in people with dementia. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15904–15917, Singapore. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Wes Gurnee and Max Tegmark. 2024. Language models represent space and time. In *The Twelfth International Conference on Learning Representations*.

Yang Han, Jacqueline C.k. Lam, Victor O.k. Li, and Lawrence Y. L. Cheung. 2025. An LLM-based temporal-spatial data generation and fusion approach for early detection of late onset Alzheimer's disease (LOAD) stagings especially in Chinese and English-speaking populations. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 14977–14990, Suzhou, China. Association for Computational Linguistics.

Yuto Harada, Yusuke Yamauchi, Yusuke Oda, Yohei Oseki, Yusuke Miyao, and Yu Takagi. 2025. Massive supervised fine-tuning experiments reveal how data, layer, and training factors shape LLM alignment quality. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 22360–22381, Suzhou, China. Association for Computational Linguistics.

Wenlong Hou, Guangqian Yang, Ye Du, Yeung Lau, Lihao Liu, Junjun He, Ling Long, and Shujun Wang. 2025. Adagent: Llm agent for alzheimer's disease analysis with collaborative coordinator. In *AI for Clinical Applications: First International Workshops, Agentic AI 2025, CREATE 2025, and Clinical MLLMs 2025, Held in Conjunction with MICCAI 2025, Daejeon, South Korea, September 23 and 27, 2025, Proceedings*, page 23–32, Berlin, Heidelberg. Springer-Verlag.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Junsol Kim, James Evans, and Aaron Schein. 2025. Linear representations of political perspective emerge in large language models. In *The Thirteenth International Conference on Learning Representations*.

Changye Li, David Knopman, Weizhe Xu, Trevor Cohen, and Serguei Pakhomov. 2022. GPT-D: Inducing dementia-related linguistic anomalies by deliberate degradation of artificial neural language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1866–1877, Dublin, Ireland. Association for Computational Linguistics.

Chuyuan Li, Giuseppe Carenini, and Thalia Field. 2025a. On large foundation models and Alzheimer's disease detection. In *Proceedings of the Second Workshop on Patient-Oriented Language Processing (CL4Health)*, pages 158–168, Albuquerque, New Mexico. Association for Computational Linguistics.

Jinchao Li, Jianwei Yu, Zi Ye, Simon Wong, Manwai Mak, Brian Mak, Xunying Liu, and Helen Meng. 2021. A comparative study of acoustic and linguistic features classification for alzheimer's disease detection. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6423–6427.

Rumeng Li, Xun Wang, Dan Berlowitz, Jesse Mez, Honghuang Lin, and Hong Yu. 2025b. Care-ad: a multi-agent large language model framework for alzheimer's disease prediction using longitudinal clinical notes. *npj Digital Medicine*, 8(1):541.

Matej Martinc and Senja Pollak. 2020. Tackling the adress challenge: A multimodal approach to the automated recognition of alzheimer's dementia. In *Interspeech*.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human*

Language Technologies, pages 746–751, Atlanta, Georgia. Association for Computational Linguistics.

Neel Nanda, Andrew Lee, and Martin Wattenberg. 2023. Emergent linear representations in world models of self-supervised sequence models. In *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 16–30, Singapore. Association for Computational Linguistics.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Kiho Park, Yo Joong Choe, and Victor Veitch. 2024. The linear representation hypothesis and the geometry of large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 39643–39666. PMLR.

Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. Instruction tuning with gpt-4. *Preprint*, arXiv:2304.03277.

Charlene Pope and Boyd H. Davis. 2011. Finding a balance: The carolinas conversation collection. *Corpus Linguistics and Linguistic Theory*, 7(1):143–161.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. Exploring the limits of transfer learning with a unified text-to-text transformer. *Preprint*, arXiv:1910.10683.

Morteza Rohanian, Julian Hough, and Matthew Purver. 2021. Alzheimer's dementia recognition using acoustic, lexical, disfluency and speech pause features robust to noisy inputs. In *Interspeech 2021*, pages 3820–3824.

Mengke Shi, Gary Cheung, and Seyed Reza Shahamiri. 2023. Speech and language processing with deep learning for dementia diagnosis: A systematic review. *Psychiatry Research*, 329:115538.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Abubakr Babiker, Nathanael Schärli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, and 13 others. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Anita Pothen Skaria. 2022. The economic and societal burden of alzheimer disease: managed care considerations. *The American journal of managed care*, 28(10 Suppl):S188–S196.

Hieu Tran, Zhichao Yang, Zonghai Yao, and Hong Yu. 2024. Bioinstruct: instruction tuning of large language models for biomedical natural language processing. *Journal of the American Medical Informatics Association*, 31(9):1821–1832.

Yi Wang, Jiajun Deng, Tianzi Wang, Bo Zheng, Shoukang Hu, Xunying Liu, and Helen Meng. 2023. Exploiting prompt learning with pre-trained language models for alzheimer's disease detection. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Yi Wang, Tianzi Wang, Zi Ye, Lingwei Meng, Shoukang Hu, Xixin Wu, Xunying Liu, and Helen Meng. 2022. Exploring linguistic feature and model combination for speech recognition based automatic AD detection. In *Interspeech 2022*, pages 3328–3332.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022a. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022b. Finetuned language models are zero-shot learners. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Zi Ye, Shoukang Hu, Jinchao Li, Xurong Xie, Mengzhe Geng, Jianwei Yu, Junhao Xu, Boyang Xue, Shansong Liu, Xunying Liu, and Helen Meng. 2021. Development of the cuhk elderly speech recognition system for neurocognitive disorder detection using the dementiabank corpus. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6433–6437.

Jiahong Yuan, Yuchen Bian, Xingyu Cai, Jiaji Huang, Zheng Ye, and Kenneth Church. 2020. Disfluencies and fine-tuning pre-trained language models for detection of alzheimer's disease. In *Interspeech 2020*, pages 2162–2166.

Meiwei Zhang, Yuwei Pan, Qiushi Cui, Yang Lü, and Weihua Yu. 2025. Multimodal llm for enhanced

alzheimer's disease diagnosis: Interpretable feature extraction from mini-mental state examination data. *Experimental Gerontology*, 208:112812.

Youxiang Zhu, Nana Lin, Kiran Sandilya Balivada, Daniel Haehn, and Xiaohui Liang. 2024. Adversarial text generation using large language models for dementia detection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21918–21933, Miami, Florida, USA. Association for Computational Linguistics.

A Special Markers

A short excerpt from a Pitt transcript illustrates these conventions:

```
*PAR: The boy is = is taking cookies.
%mor: pro|The n|boy v|be aux|is v|take
n|cookie
*INV: What is he doing?
*PAR: ((points to the picture)) I don't
know.
*PAR: He uh uh ((laughs)) is taking
cookies?
```

In this example, `*PAR:` and `*INV:` denote the participant and interviewer, = indicates repetition, uh uh represents hesitation, and `((laughs))` captures nonverbal behavior. Such detailed annotations allow researchers to analyze speech patterns, lexical retrieval, and cognitive markers, which are particularly relevant in studies of dementia.

B Linguistic Marker Set

C The Token Length Distribution of Each Dataset

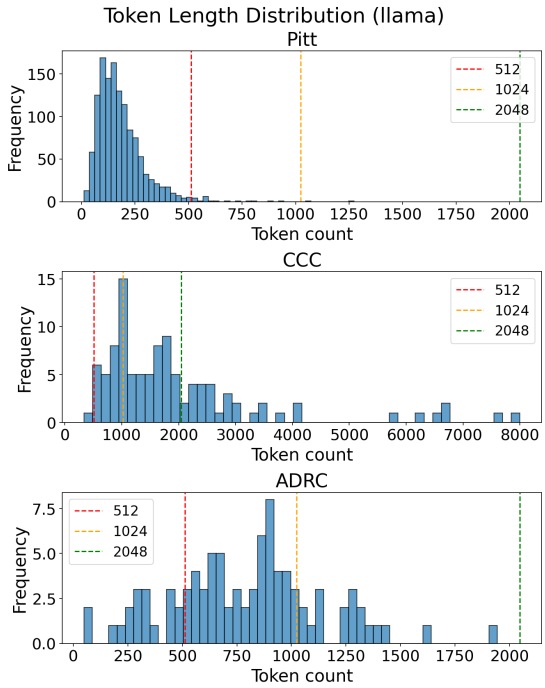

Figure 3: This figure shows the token length distribution of each dataset with the Llama3.2-1B-Instruct tokenizer.

D Probing Training

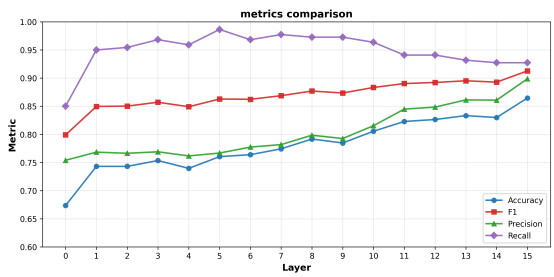

Figure 4: This figure shows probing training results of the SFT Llama3.2-1B-Instruct model.

Symbol / Mark	Meaning	Example
`*PAR:` / `*INV:`	Speaker label	`*PAR:` = participant, `*INV:` = interviewer
`[...]`	Unintelligible speech	I went to the [...] yesterday.
`xxx`	Unintelligible word	He xxx the cookies.
`?`	Uncertain transcription	I saw a dog?
`((...))`	Non-verbal actions	((laughs)), ((coughs))
`=`	Repetition / continuation	I = I went there.
`%mor:`	Morphosyntactic tier	pro|The n|boy v|be
`%com:`	Comment / notes	Optional researcher annotations

Table 6: Common CHAT special marks in Pitt transcripts.

Pattern	Regex	Example Markers	Description
1	`&-\w+`	`&-uh`, `&-um`	Filled pauses
2	`&=\w+[:\w]*`	`&=clears_throat`, `&=sighs`	Non-verbal sounds
3	`\[\+\s*[^\]]*\]`	`[+ gram]`, `[+ exc]`	Grammatical/exclamation
4	`\[/?/?\]`	`[/]`, `[//]`	Retracing
5	`\[:\s*[^\]]*\]`	`[: word]`, `[: ...]`	Replacement
6	`<[^>]*>`	`<...>`, `<word>`	Uncertain/omitted
7	`\[[^\]]*\]`	`[word]`, `[xxx]`	Any brackets
8	`\+\<+\>|`	`+<`, `+>`	Other markers
9	`\(\.+\)`	`(.)`, `(..)`, `(...)`	Pause markers
10	`\bxxx\b`	`xxx`	Unintelligible

Table 7: Linguistic marker set.