What Do LLMs Know About Alzheimer's Disease?
Fine-Tuning, Probing, and Data Synthesis for AD Detection

Anonymous ACL submission

Abstract

Reliable early detection of Alzheimer's disease (AD) is challenging, particularly due to limited availability of labeled data. While large language models (LLMs) have shown strong transfer capabilities across domains, adapting them to the AD domain through supervised fine-tuning remains largely unexplored. In this work, we fine-tune an LLM for AD detection and investigate how task-relevant information is encoded within its internal representations. We employ probing techniques to analyze intermediate activations across transformer layers, and we observe that, after fine-tuning, the probing values of specific words and special markers change substantially, indicating that these elements assume a crucial role in the model's improved detection performance. Guided by this insight, we design a curated set of task-aware special markers and train a sequence-to-sequence model as a data-synthesis tool that leverages these markers to generate structurally consistent and diagnostically informative synthetic samples. We evaluate the synthesized data both intrinsically and by incorporating it into downstream training pipelines.

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 both timely intervention and improve patient outcomes as populations age globally (de la Fuente Garcia et al., 2020). Recently, LLMs have demonstrated success at a myriad of downstream tasks through fine-tuning on task-relevant datasets, thus 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). However, fine-tuning language models within the AD domain is currently very underexplored (Zhang et al., 2025; Hou et al., 2025; Li et al., 2025b; Dhinagar et al., 2025).

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 large language models (LLMs) presents a promising opportunity to leverage their strong linguistic understanding for this clinical application, but fine-tuning LLMs for AD detection remains underexplored, 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 study the use of fine-tuned LLMs for AD detection, and the extent to which they encode meaningful information pertaining to AD. We find that fine-tuning leads to pronounced shifts in probe values for certain lexical items and for spe-

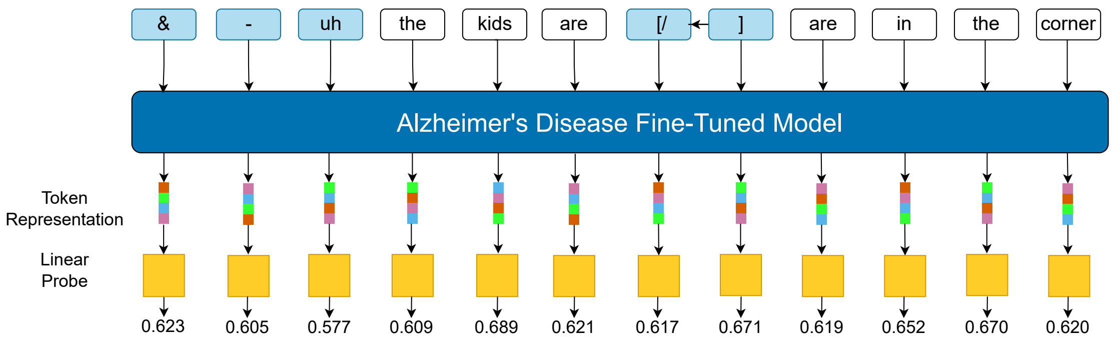

Figure 1: We acquire the representation of each token and project it by the linear probe. Tokens marked with blue are special markers that capture critical aspects of speech.

cial annotation markers (i.e., indicators of pauses, repetitions, and unintelligible speech), suggesting that these elements constitute key signals underlying the model's improved AD detection capability.

Crucially, our token-level representation analysis provides an avenue for data synthesis: by identifying and inserting AD-relevant linguistic markers into plain text, we can generate synthetic transcripts that exhibit AD–related linguistic features. Guided by this insight, we preprocess AD transcripts from publicly available datasets and fine-tune a T5 model in a sequence-to-sequence manner to learn to transform text-only transcripts to richer transcripts with AD-relevant speech and disfluency markers. Thus, overall our contributions are fourfold: (1) **we investigate the use of SFT in the AD domain**, achieving competitive performance on standard AD detection benchmarks; (2) **we train linear probes on LLM representations** to capture and quantify AD–related linguistic information; (3) **we conduct token-level probe-based analysis** to identify representations and tokens that are most informative for AD classification; and (4) **we propose a sequence-to-sequence data synthesis method** to generate text exhibiting linguistic features characteristic of Alzheimer's disease.

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, 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).

Overall, these studies emphasize system design and performance, seeking to gradually move the dial forward on overall AD detection performance according to 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

SFT is a widely used approach 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 through instruction tuning 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-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).

2.3 Linear probes

Finally, recent advances in natural language processing reveal that linguistic and stylistic concepts often manifest as linear features within the high-dimensional spaces of language model representations. Mikolov et al. (2013) first demonstrated that word embeddings learned by neural language models capture syntactic and semantic regularities through nearly constant vector offsets, enabling analogical reasoning (e.g., the popular example *king – man + woman* ≈ *queen*). Building on this empirical foundation, Park et al. (2024) formalized the Linear Representation Hypothesis, defining what it means for concepts to be linearly represented in both input (embedding) and output (unembedding) spaces. They connected these notions respectively to intervention and measurement, and introduced a causal inner product that aligns concept subspaces and respects semantic independence. Gurnee and Tegmark (2024), Kim et al. (2025), and Nanda et al. (2023) used linear probes to measure the concepts encoded in the representations of LLMs including space, time and political concepts, and show evidence of the linear representation hypothesis. In our work, we apply linear probes to estimate the direction associated with AD in the model's representations and to measure the extent to which AD-related information is captured in the token-level representations.

3 Methodology

In this work, we explore the performance of SFT with LLMs for AD detection. We also investigate how the concept of AD is encoded within the internal representations of LLMs and how it can be quantitatively measured. Building on evidence that concepts are linearly encoded in representation and can be detected (Mikolov et al., 2013; Park et al., 2024; Gurnee and Tegmark, 2024; Nanda et al., 2023; Kim et al., 2025), 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 Problem Description

Our goal is to identify a fine-tuning framework that promotes strong LLM performance on both a specialized healthcare domain and, more broadly, within the general knowledge domain. We cast *AD detection* as a classification problem; given a transcript T from an AD or healthy control patient, the LLM's task is to predict to which group the patient belongs. Within the present scope, we assume access to a dataset with transcriptions of audiorecorded interviews, with each transcript having a binary AD label (*AD* or *healthy control*).

3.2 Supervised Finetuning Loss

We use a standard cross-entropy loss to demonstrate the baseline performance of SFT. We then add a contrastive learning loss term (Gao et al., 2021; Su et al., 2022; Jain et al., 2023; Kumar et al., 2023; Gong et al., 2025; Klein and Nabi, 2025) to let the language model learn differences in representation space to enhance its discriminative ability. Separately, we investigate the use of a focal loss term due to its reported ability to handle unbalanced features in prior work (Lin et al., 2017; Huang et al., 2021). The last loss that we study, label smoothing, transforms the prediction label from a one-hot encoding into a softer label (e.g., [0,1] into [0.05, 0.95]), preventing the model from becoming overconfident about the training labels.

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

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

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

Standard SFT with Contrastive Loss. This setting augments CE loss with a contrastive learning objective to improve representation separability:

$$\mathcal{L}_{\text{total}} = (1-\alpha)\cdot\mathcal{L}_{\text{LM}} + \alpha\cdot\mathcal{L}_{\text{CL}} \qquad (2)$$

where λ controls the contribution of the contrastive term. We adopt the contrastive learning loss(Chopra et al., 2005):

$$\mathcal{L}_{\text{CL}}(x_1, x_2, y) = (1-y) \cdot d^2(x_1, x_2) \\ + y \cdot \max(0,\, m - d(x_1, x_2))^2 \quad (3)$$

The embeddings x_1 and x_2 are d-dimensional vectors from the last token's hidden state of the final layer, paired within a batch (even indices with odd indices). The label y is binary: 0 for same class, 1 for different class, derived from comparing the original labels. The distance d is the Euclidean distance between the two embeddings. The margin m is a positive hyperparameter (default 1.0) that sets the minimum required separation for different-class pairs.

Focal Loss. Focal loss down-weights easy examples and emphasizes hard samples during training:

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

where $\gamma \geq 0$ is the focusing parameter.

Label Smoothing. Label smoothing regularizes training by softening hard target labels:

$$y_i^{\text{LS}} = \begin{cases} 1-\epsilon, & \text{if } i = y, \\ \frac{\epsilon}{C-1}, & \text{otherwise}, \end{cases} \quad (5)$$

$$\mathcal{L}_{\text{LS}} = -\sum_{i=1}^{C} y_i^{\text{LS}} \log(p_i), \quad (6)$$

where $\epsilon \in [0, 1]$ is the smoothing factor.

3.3 Probing

After studying the performance of SFT for AD detection, we investigate internal LLM representations of linguistic information and their relevance to AD using probing methods. By training probes on representations related to AD, we acquire probes that represent the direction of AD at the linguistic level. We then measure how strongly the AD concept is expressed in the representation.

Probes are diagnostic models used to analyze the linguistic and semantic information encoded in the hidden representations of pretrained language models. Letting $h \in \mathbb{R}^d$ denote a hidden representation extracted from a specific layer for an input token or sentence, a probe is a lightweight function f_ϕ trained to predict a linguistic property y (e.g., part-of-speech tags, syntactic dependencies, or semantic roles) from h. The standard probing objective minimizes a supervised loss over a labeled dataset $\mathcal{D}_{\text{probe}}$:

$$\mathcal{L}_{\text{probe}}(\phi) = \mathbb{E}_{(h,y)\sim\mathcal{D}_{\text{probe}}}\left[\ell(f_\phi(h), y)\right], \quad (7)$$

where $\ell(\cdot)$ is an appropriate loss (e.g., squared error for regression or cross-entropy for classification). Following Gurnee and Tegmark (2024), we train the linear probes using Mean Squared Error (MSE) loss together with L2 regularization, also known as Ridge regression:

$$\ell = \ell_{\text{MSE}} + \lambda_{\text{ridge}} \cdot \ell_{\text{ridge}} \quad (8)$$

where:

$$\ell_{\text{MSE}} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \quad (9)$$

$$\ell_{\text{ridge}} = \|\mathbf{W}\|_2^2 = \sum_{j=1}^{d} w_j^2 \quad (10)$$

where N is the number of training samples, $y_i \in \{0, 1\}$ is the true label for sample i (0 = Healthy Control, 1 = AD), $\hat{y}_i$ is the probe prediction for sample i (raw regression output, before sigmoid), $\mathbf{W} = [w_1, w_2, \ldots, w_d]$ are the weights of the linear probe head, d is the hidden size of the representation, and λ_{ridge} is the ridge regularization coefficient.

4 Experiments

4.1 Experimental Settings

Hyperparameters. For SFT experiments, we employ a consistent set of general training hyperparameters across all loss functions: a learning rate of 2e-5, per-device batch size of 4 with gradient accumulation of 4 steps, yielding an effective batch size of 16, and 10 training epochs. For the standard cross-entropy loss, no additional loss-specific parameters are required. The label smoothing loss introduces a single hyperparameter, the label smoothing factor set to 0.1, which applies 10% smoothing to the target labels to improve generalization. The focal loss configuration includes two parameters: focal_alpha =0.25 to weight the rare class and focal_gamma=2.0 to focus learning on hard examples, making it particularly effective for handling class imbalance. Finally, the contrastive learning loss combines the standard language modeling loss with a contrastive component using a weighted sum, controlled by contrastive_alpha =0.1 which determines the mixing ratio between the two loss components, along with a contrastive margin of 1.0 for the triplet loss formulation.

Data. We use the **DementiaBank** (Becker et al., 1994) dataset for our experiments. DementiaBank is a widely used resource for research on cognitive decline, particularly Alzheimer's disease, and is part of the broader *TalkBank* project, which provides standardized corpora for the study of human communication. DementiaBank contains audio recordings and corresponding transcriptions of participants performing structured language tasks, with the most commonly employed task being the *Cookie Theft* picture description task from the Boston Diagnostic Aphasia Examination. The dataset includes speech samples from individuals diagnosed with AD as well as age-matched healthy controls, enabling comparative analyses of linguistic and cognitive patterns. Transcriptions in DementiaBank follow the CHAT (Codes for the Human Analysis of Transcripts) format, which is designed to encode both verbal and non-verbal speech features systematically. Each line begins with a speaker label (e.g., `*PAR:` for participant, `*INV:` for interviewer) and may include additional tiers such as `%mor:` for morphosyntactic annotation or `%com:` for researcher comments. Special marks in the CHAT format capture critical aspects of speech, including pauses, repetitions, nonverbal actions, and unintelligible segments. Table 2 summarizes commonly used annotations.

Other Modeling Details. We use llama3-1b-instruct(Meta AI, 2024) and qwen-2.5-1.5b-instruct as the backbone models(Qwen Team, 2024). We experiment with each loss function condition on both backbone models.There are 1044 AD data points and 247 control data points. We split the dataset into 80/20 for training/evaluation. All experiments use the same base training configuration, with only the loss-specific parameters varying to enable fair comparison across different loss function strategies. We conducted all experiments using four A100 PCIe GPUs, each equipped with 40 GB of memory. Fine-tuning a single model on one GPU takes approximately 9 minutes. Perplexity evaluation requires about 3 minutes per model on a single GPU, while the AD detection task takes roughly 12 minutes per model per round on a single GPU.

4.2 Supervised Fine-tuning Results

Figure 7 compares total loss trajectories during language-model fine-tuning across multiple experimental runs for all four SFT loss conditions. This figure illustrates the training loss over successive epochs, where all runs exhibit a consistent downward trend, indicating stable optimization and effective learning. Despite minor differences in initial loss values, the curves converge toward similarly low training losses by the final epochs. However, the detection performance of this method is poor. In Table 1, by evaluating their impact under identical fine-tuning conditions, we quantify how each objective influences the model's learning dynamics and downstream performance.

The experimental results reveal notable differences in model behavior depending on the chosen loss function. In terms of classification metrics, label smoothing and standard CE perform competitively. Label smoothing delivers a high accuracy at 0.853, the highest recall at 0.975, and achieves an F_1=0.919, placing it among the top-performing methods. Standard CE also achieves an accuracy of 0.868, while also demonstrating strong precision of 0.881 and recall of 0.96.

CE with a contrastive learning term underperforms across all metrics except recall, where it reaches 1.0 but fails to maintain balanced precision at 0.772, leading to the lowest F_1=0.875 and also a low accuracy at 0.772. This suggests that contrastive fine-tuning may distort token or representation distributions, harming prediction consistency. Finally, focal loss yields reasonable recall at 0.94 and F_1=0.902, but trails behind label smoothing and CE. It provides moderate improvements over CE in some metrics, but does not surpass label smoothing in any category.

Label smoothing achieves a low perplexity at 36.79, indicating that it produces the most stable and calibrated next-token probability distribution among the four approaches. In contrast, CE and CE with contrastive learning exhibit extremely high perplexity at 39.76 and 283,439, respectively, suggesting training instability or overfitting to the supervised objective without adequately modeling the token distribution. Focal loss shows moderate perplexity at 1806.18, outperforming CE but still substantially worse than label smoothing.

Integrating pairwise contrastive learning into CE degrades perplexity and overall classification quality, implying a misalignment between con-

Model + Loss Function	PPL	Accuracy	Precision	Recall	F_1
Vanilla Llama	12.37	0.441(0.200)	0.758(0.017)	0.405(0.027)	0.528(0.024)
Llama + focal loss	28.23	0.842(0.010)	0.866(0.010)	0.940(0.005)	0.902(0.006)
Llama + label smoothing	36.79	0.853(0.009)	0.855(0.007)	0.975(0.004)	0.911(0.006)
Llama + CE	39.76	0.868(0.012)	0.881(0.008)	0.960(0.008)	0.919(0.008)
Llama + CE w/CL	283439	0.772(0.000)	0.772(0.000)	1.000(0.000)	0.875(0.000)
Vanilla Qwen	9.16	0.769(0.002)	0.771(0.000)	0.996(0.002)	0.869(0.001)
Qwen + focal loss	12.54	0.791(0.008)	0.793(0.004)	0.988(0.006)	0.880(0.004)
Qwen + label smoothing	13.66	0.799(0.005)	0.806(0.008)	0.975(0.007)	0.882(0.002)
Qwen + CE	13.54	0.810(0.013)	0.808(0.010)	0.990(0.009)	0.890(0.007)
Qwen + CE w/CL	4211.1	0.772(0.000)	0.772(0.000)	1.00(0.000)	0.872(0.000)
T5	N/A	0.903(0.000)	0.940(0.000)	0.935(0.000)	0.937(0.000)
Bert	N/A	0.915(0.004)	0.945(0.004)	0.945(0.004)	0.945(0.004)

Table 1: Model performance with different SFT loss functions. *PPL*=perplexity; *CE*=cross-entropy; *CL*=contrastive learning. Standard deviations are shown in (parentheses). We report perplexity only for causal language models.

trastive objectives and language-modeling objectives in this setup. These findings show that label smoothing consistently balances likelihood-based and classification-based performance, while CE retains strong discriminative ability but suffers from worse perplexity. Contrastive learning introduces instability that negatively impacts language-modeling quality. Overall, label smoothing emerges as the most robust and effective loss function for SFT, striking the best balance between perplexity and downstream predictive performance.

We report the performance of the Llama and Qwen models. Although Qwen generally achieves lower perplexity than Llama, its accuracy is lower. Therefore, we do not select Qwen as our final model. We also include smaller models such as T5 and BERT as baselines. While these discriminative (BERT) and sequence-to-sequence (T5) models achieve higher accuracy than Llama in our experiments, we evaluate Llama to explore the extent to which a causal language model can perform in the context of Alzheimer's disease detection. Our results show that, despite its lower task-specific performance, Llama remains competitive and offers desirable properties for broader generative and discourse-level modeling.

5 Probing Analyses

To understand how fine-tuning changes the model's internal representations of AD-related linguistic patterns, we conducted token-level probe analysis comparing vanilla and fine-tuned models. Our methodology involved training linear probes on layer-specific representations from the fine-tuned model, which was trained with a label smoothing loss. We then selected the best probe by comparing classification metrics and applied the probe to extract token-level probe values from evaluation transcripts. We computed the difference between fine-tuned and vanilla model representation probe predictions for each token, and we analyzed the distribution of probe values across all tokens to characterize how fine-tuning affected the overall representation space.

Our results reveal that fine-tuning increases mean probe values when measured with probes trained on fine-tuned representations. Distribution analysis demonstrated that while mean values increased, the separation between AD and Control samples was maintained, with AD samples consistently showing higher probe values than Control samples across both vanilla and fine-tuned models. Token-level difference analysis showed selective changes, with certain tokens (particularly marker tokens and content words) exhibiting larger differences, indicating that fine-tuning learned to emphasize or de-emphasize specific linguistic patterns.

5.1 Probes Training

We conducted a layerwise probing experiment in which a simple classifier was trained on the activation extracted from each layer. We use a simple linear model referred to as a probe that maps the model's internal representations to a single scalar value. Importantly, this probe is not trained to diagnose Alzheimer's disease, but instead defines a

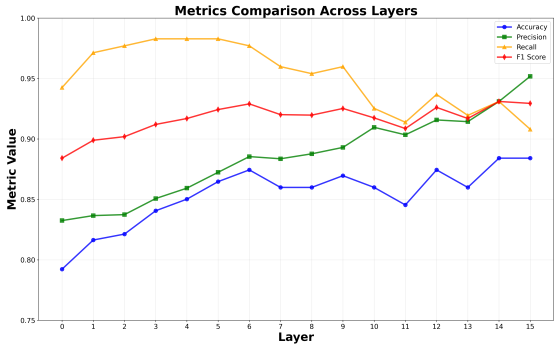

Figure 2: Progression of probe performance metrics across successive model layers.

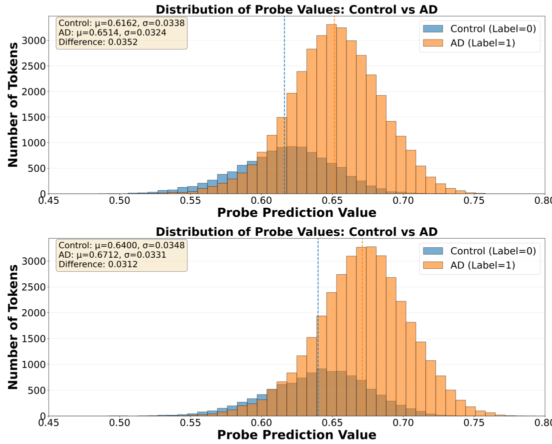

Figure 3: Token probing value distribution.

direction in the model's representation space.

Each training example consists of a full transcript as input and an associated label as output. The probe uses the activation extracted from layers at the last token position of each transcript as the input feature. For every layer l, we collected the activation produced in response to the same fixed dataset and trained an identical probe classifier to predict the target labels. Each probe was evaluated using accuracy, F_1, precision, and recall to capture complementary aspects of predictive performance.

In Figure 2, we show the progression of probe performance metrics across successive model layers. All metrics except recall exhibit overall improvement with increasing depth, indicating that later layers encode more AD-relevant linguistic information. We observe that across evaluation metrics, probes trained on representations from layer 14 consistently achieve the strongest performance, and their training loss converges more reliably than probes trained on other layers. Based on this empirical evidence, we select the representations from layer 14 for subsequent analysis. These results suggest that representations at this layer encode AD-related linguistic variation in a form that is particularly linearly accessible to a simple probe, providing an AD linguistic direction. This direction can be used as a measurement tool: projecting internal representations onto this direction reveals how strongly a given token's representation aligns with the features captured by the probe.

5.2 Probing Representation

After training, we use the probe not as a classifier, but instead as a measurement axis for AD linguistic direction. We apply the same linear mapping to the internal representations of every generated token in unseen transcripts. This produces a scalar value for each token, indicating how strongly that token's representation aligns with the learned direction. Crucially, these token-level values are not interpreted as "this token is AD-related." Instead, they indicate the degree to which the model's internal state at that moment aligns with the probe direction. Aggregating these values across tokens yields a distribution that characterizes how a group of transcripts is represented internally by the model.

We perform our probing analysis on two models: (1) a vanilla (pretrained) language model, and (2) a fine-tuned model trained for AD detection. We find in Figure 3 that fine-tuning does not merely improve predictions on AD detection but reshapes the internal representation space of the model. Specifically, fine-tuning aligns AD-related linguistic information in the representation along a coherent direction that is consistently expressed across tokens. Importantly, this does not imply that the model has learned an explicit or interpretable "AD concept." Rather, it suggests that fine-tuning consolidates multiple subtle linguistic patterns of fluency, lexical choice, or syntactic structure into a unified representational pattern.

We show an example result from our token representation analysis in Figure 4, quantifying how fine-tuning changes the AD-related signal strength for individual tokens by comparing probe predictions before and after fine-tuning. We set the probing value change threshold as 0.02, and the Color intensity corresponds to the magnitude of change. We find that some words' and markers' representations changed substantially after fine-tuning. This analysis provides fine-grained insights into which linguistic elements the model learns to emphasize or de-emphasize during fine-tuning. The probe

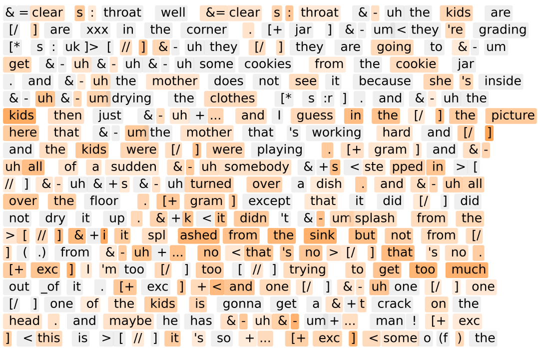

Figure 4: This figure shows the token probing value difference before and after fine-tuning

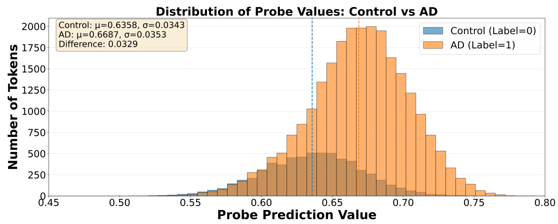

Figure 6: Token probing value distribution.

identifies a latent representational direction along which AD-related linguistic variation is expressed in the base language model, and token-level projections along this direction highlight regions of text where such variation is most strongly manifested. Based on this empirical observation, we manually make an AD linguistic marker set and use it to train a T5 model as a data synthesis tool. The linguistic marker set is in Appendix Table 3.

5.3 Data Sythesis with T5 Model

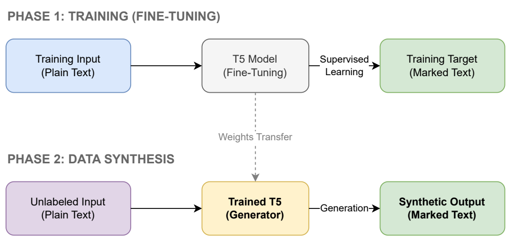

Figure 5: T5 marker pipeline

To operationalize insights obtained from the probing analysis, we finally employed a T5 model as a data synthesis tool. We first removed all special markers in the training subset, resulting in plain transcripts without any speech metadata. We trained the model on the training subset to map the plain data to richer, marked data, and evaluated the model by using T5 to generate synthetic data on the cleaned evaluation subset. We then used the fine-tuned llama3.2-1b-Instruct model to classify this synthetic data. We say the synthetic data belongs to the same dataset as the original dataset if the model performs similarly on both the original dataset and the synthetic dataset, and when using the probe to investigate the distribution of token representation, both datasets show similar distributions.

We then train a T5 model to rewrite text using these markers, so we can generate new sentences in a controlled way that reflect the patterns related to AD-related features in the model's representations.

Our experiments show that the fine-tuned Llama model can achieve 0.8417 accuracy, 0.8621 precision, 0.9569 recall, and 0.9070 F_1, which is comparable to the results in Table 1. Additionally, the token distribution is illustrated in Figure 6, displaying a similar mean value and standard deviation. These experimental observations suggest that the synthetic data distribution is similar to the original data distribution. This framework provides a way to empirically synthesize data with AD linguistic features from plain text.

6 Conclusion

In this study, we demonstrate that adapting LLMs to the AD domain through SFT is both feasible and effective. Through probing analysis of the model's representations, we find that, after fine-tuning, the probing values of certain words and special markers change substantially, indicating that these elements play a central role in the improved AD detection performance. Building on this insight, we curate a set of task-aware special markers and develop a sequence-to-sequence data-synthesis model that leverages these markers to generate structurally consistent and diagnostically meaningful synthetic samples.

7 Limitations

Despite the encouraging results, several limitations remain. First, the AD dataset used for fine-tuning is small and may not fully represent the linguistic diversity of real clinical populations. Future work should incorporate larger, multilingual, and more clinically diverse corpora to ensure robustness and fairness.

Second, while contrastive learning was included to enhance representational separability, results indicate that naive integration of contrastive objectives can destabilize language-modeling performance. More principled approaches, such as curriculum-based contrastive sampling, patient-level contrastive signals, or layer-wise representation alignment, may yield better synergy between discriminative and generative objectives.

Third, evaluation primarily focuses on binary classification metrics, which may not fully capture clinically meaningful performance. Downstream evaluations involving explainability, error analysis, and alignment with clinical assessment protocols are needed to ensure that LLM-based detectors behave safely and transparently.

Finally, this work does not explore parameter-efficient tuning strategies that could further alleviate catastrophic forgetting. Future directions include testing methods such as LoRA-based adapters, rehearsal-free continual learning, or regularization techniques tailored for long-form linguistic cognition tasks. Overall, expanding the dataset, refining loss formulations, and developing more clinically grounded evaluation frameworks will be critical for making AD detection with LLMs both reliable and practically deployable.

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 Boller, Oscar L. Lopez, Judith Saxton, and Kathy 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.

S. Chopra, R. Hadsell, and Y. LeCun. 2005. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 539–546 vol. 1.

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.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Junqing Gong, Binhan Yang, and Wei Shen. 2025. A triple-view framework for fine-grained emotion classification with clustering-guided contrastive learning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4970–4984, Vienna, Austria. Association for Computational Linguistics.

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 fornbsp;alzheimer's disease analysis withnbsp;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*.

Yi Huang, Buse Giledereli, Abdullatif Köksal, Arzucan Özgür, and Elif Ozkirimli. 2021. Balancing methods for multi-label text classification with long-tailed class distribution. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8153–8161, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Nihal Jain, Dejiao Zhang, Wasi Uddin Ahmad, Zijian Wang, Feng Nan, Xiaopeng Li, Ming Tan, Ramesh Nallapati, Baishakhi Ray, Parminder Bhatia, Xiaofei Ma, and Bing Xiang. 2023. ContraCLM: Contrastive learning for causal language model. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6436–6459, Toronto, Canada. Association for Computational Linguistics.

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*.

Tassilo Klein and Moin Nabi. 2025. Contrastive perplexity for controlled generation: An application in detoxifying large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2493–2508, Vienna, Austria. Association for Computational Linguistics.

Vaibhav Kumar, Hana Koorehdavoudi, Masud Moshtaghi, Amita Misra, Ankit Chadha, and Emilio Ferrara. 2023. Controlled text generation with hidden representation transformations. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9440–9455, Toronto, Canada. 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.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

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

Meta AI. 2024. Llama-3.2-1b-instruct. Large language model.

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.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

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

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.

Yixuan Su, Tian Lan, Yan Wang, Dani Yogatama, Lingpeng Kong, and Nigel Collier. 2022. A contrastive framework for neural text generation. In *Advances in Neural Information Processing Systems*.

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.

A SFT Loss

B Special Markers

A short excerpt from a DementiaBank 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.

C Linguistic Marker Set

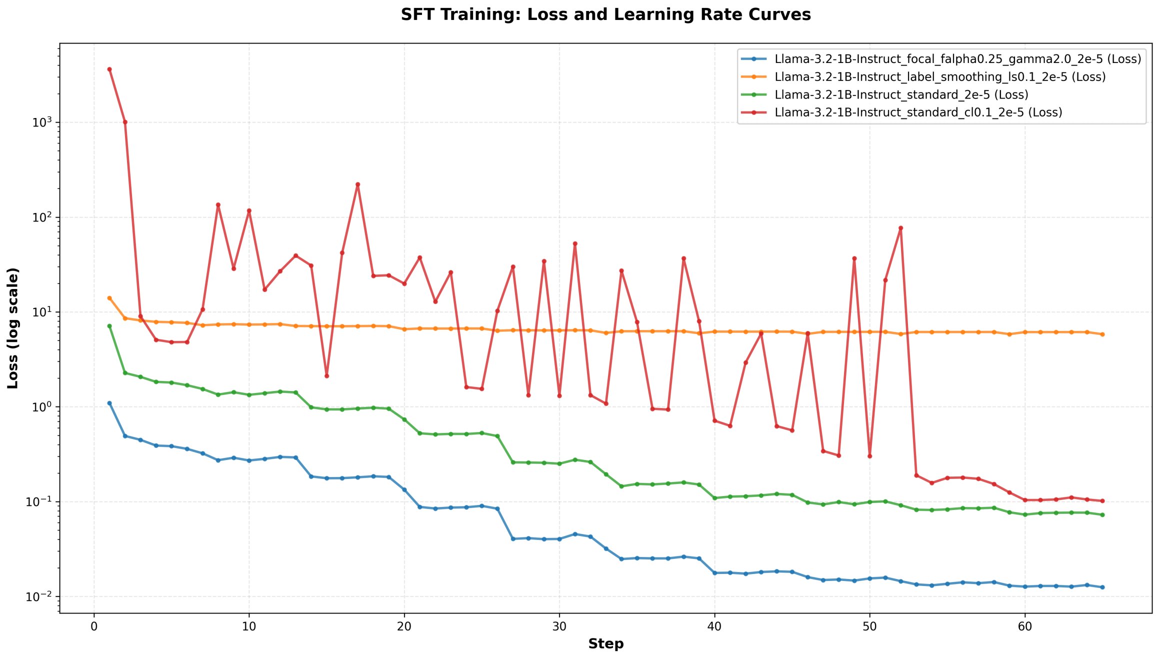

Figure 7: Llama SFT loss.

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 2: Common CHAT special marks in DementiaBank 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 3: Linguistic marker set.