CROSS-LINGUAL ALZHEIMER'S DISEASE DETECTION WITH MULTIMODAL LLMS VIA SPEECH CUE-AUGMENTED PROMPTING AND INSTRUCTION TUNING

Yin-Long Liu[1], *Yuanchao Li*[2], *Yu-Ang Chen*[1], *Liu He*[1], *Rui Feng*[1], *Jia-Xin Chen*[1], *Jiahong Yuan*[1✉]

[1]National Engineering Research Center of Speech and Language Information Processing,
University of Science and Technology of China, Hefei, P. R. China

[2]Centre for Speech Technology Research, University of Edinburgh, Edinburgh, UK

lyl2001@mail.ustc.edu.cn, jiahongyuan@ustc.edu.cn

ABSTRACT

Speech-based Alzheimer's Disease (AD) detection holds significant promise, yet conventional supervised models often struggle to generalize across languages and datasets. In this study, we introduce a novel approach leveraging Multimodal Large Language Models (MLLMs) for cross-lingual AD detection. We first conduct a systematic evaluation of three MLLMs (MiDashengLM, Qwen2-Audio, and Qwen2.5-Omni) in a zero-shot setting across three AD datasets in English (ADReSS, PROCESS) and Mandarin (iFLYTEK), using four types of prompt. Our results show that a contextual chain-of-thought prompt achieves the strongest zero-shot performance. Next, we propose Speech Cue–Augmented Prompting (SCAP), which prepends natural-language descriptors of four AD-related speech cues to the prompt. Experimental results demonstrate that MLLMs with SCAP not only improve zero-shot performance but also surpass supervised models. Furthermore, we explore instruction tuning of MLLMs via low-rank adaptation, and find that instruction tuning with SCAP improves both in-domain accuracy and out-of-domain transfer, outperforming all supervised baselines in both settings. These findings suggest that carefully designed prompting strategies, together with lightweight tuning, can transform MLLMs into robust and language-agnostic tools for AD detection.

***Index Terms*—** Alzheimer's disease detection, multimodal LLMs, prompt engineering, chain-of-thought, instruction tuning

1. INTRODUCTION

Alzheimer's Disease (AD), the leading cause of dementia, results in irreversible cognitive decline, including impairments in memory, language, and executive function [1]. Early detection of AD is crucial for timely intervention, yet traditional clinical methods are often costly and invasive [2]. As speech and language impairments are among the earliest clinical manifestations of AD [3], speech analysis offers a cost-effective, non-invasive, and scalable approach to detection of AD. Consequently, research has focused on identifying linguistic and acoustic biomarkers from naturalistic tasks like picture description [4–6], with progress accelerated by community benchmark challenges [7, 8].

Despite recent advances reporting promising accuracy in AD detection, developing robust, real-world–ready speech-based systems remains challenging. First, although several datasets have been released (e.g., ADReSS [7] and PROCESS [8]), labeled AD speech remains scarce, the overall amount of data is limited, and the corpora vary in recording conditions, speaker demographics, and elicitation protocols. Moreover, most existing models are trained and evaluated in-domain and therefore fail to generalize across languages and collection sites, limiting their broader applicability. These challenges underscore the need for robust models that not only perform strongly In-Domain (ID) but also generalize well Out-of-Domain (OOD).

Multimodal Large Language Models (MLLMs) extend the powerful zero-shot and few-shot generalization capabilities of text-only LLMs to reasoning across multiple modalities (e.g., audio, text, image, video). In particular, Audio LLMs couple an audio front end with a language backbone to support instruction following directly from speech, enabling end-to-end reasoning without requiring intermediate Automatic Speech Recognition (ASR) [9]. However, applications of MLLMs to clinical AD speech remain limited. Most prior LLM-based AD studies operate on transcripts and assess text-only models, rather than audio-native reasoning [10, 11]. To our knowledge, only one recent study has begun to explore cognitive impairment detection directly from speech with an Audio LLM [12]. The study only evaluates the zero-shot ability of a single off-the-shelf model and does not explore how to carefully design prompts, represent speech cues, or adapt models for robust cross-lingual AD detection. As a result, this area remains largely underexplored.

In this paper, we propose Speech Cue-Augmented Prompting (SCAP) and combine it with instruction tuning of MLLMs for robust cross-lingual AD detection. Code is available at: `https://github.com/YinlongLiu-source/AD_MLLMs_SCAP`. Our main contributions are as follows:

- We present the first systematic zero-shot evaluation of three MLLMs across three AD datasets in English and Mandarin using four prompt types, and find that Contextual Chain-of-Thought (CoT) is the most effective.
- We propose SCAP, which prepends natural-language descriptors of four AD-related speech cues to the prompt. Experimental results show that MLLMs with SCAP improve zero-shot performance and surpass supervised models.
- We explore instruction tuning of MLLMs via Low-Rank Adaptation (LoRA), and find that tuning with SCAP not only improves in-domain (ID) accuracy but also strengthens out-of-domain (OOD) transfer, outperforming all supervised baselines in both settings.

2. DATASETS

For our cross-lingual and cross-dataset evaluation, we utilize three distinct datasets spanning both English and Chinese: the English ADReSS dataset from the INTERSPEECH 2020 Challenge [7], the English PROCESS dataset from the ICASSP 2025 Challenge [8], and a proprietary Mandarin Chinese dataset from iFLYTEK [13]. A statistical summary of these datasets is presented in Table 1. For the **ADReSS** dataset, where subjects performed the "Cookie Theft"

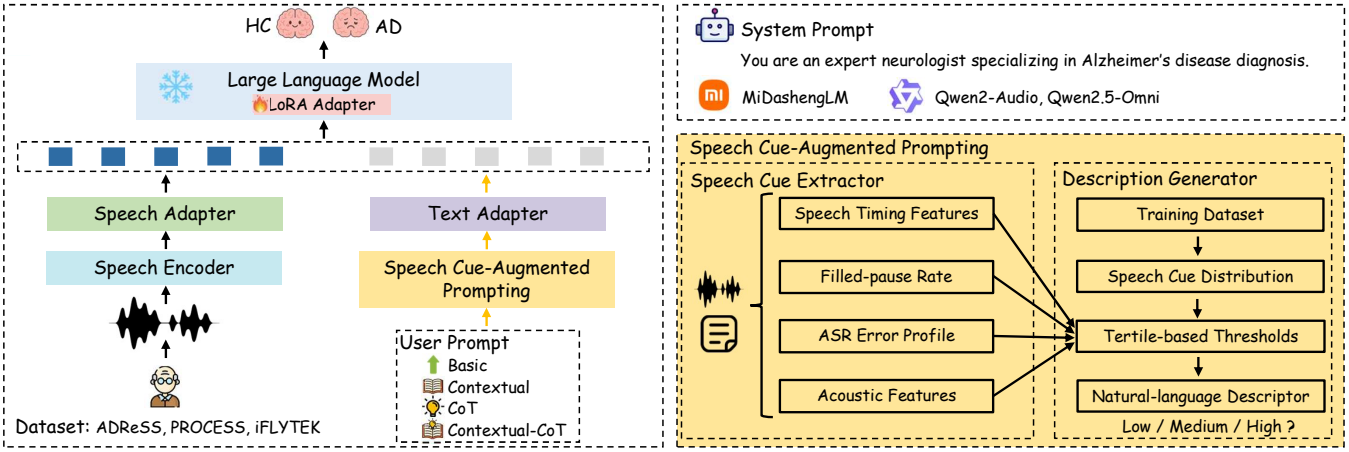

Fig. 1. Overview of the proposed methods, featuring a simplified MLLM that takes speech and text prompts as input.

picture description task [14], we follow the official train-test split. The **PROCESS** dataset involves subjects undertaking three different cognitive assessment tasks. However, we exclusively use the data from the "Cookie Theft" picture description task to ensure task consistency. This dataset originally comprises three classes: HC, Mild Cognitive Impairment (MCI), and Dementia. To unify the binary labels across our experiments, we merge the MCI and Dementia classes into a single AD category. As the organizers did not release the ground-truth labels for the official test set, we randomly partition the original training set into our own training and test sets using a 3:1 ratio. The **iFLYTEK** dataset is a proprietary collection where subjects were recruited from the Department of Neurology and the Department of Memory Clinic of Shanghai Tongji Hospital and were instructed to undertake the same picture description task. We also randomly split this dataset into training and test sets with a 3:1 ratio. For each subject across all three datasets, the data includes a single audio recording, a corresponding manual transcription, and the subject's diagnostic label, age, and gender.

3. METHODOLOGY

This section outlines our proposed methods (as shown in Fig. 1), including MLLMs for AD Detection, Prompt Design, Speech Cue–Augmented Prompting (SCAP), and Instruction Tuning.

3.1. MLLMs for AD Detection

To investigate the capabilities of various MLLMs, we select three distinct and state-of-the-art MLLMs: MidashengLM-7B [15], Qwen2-Audio-7B-Instruct [16], and Qwen2.5-Omni-7B [17]. The former two MLLMs accept both audio and text as bimodal inputs, while Qwen2.5-Omni-7B is a quadru-modal MLLM capable of processing audio, text, image, and video inputs. In this paper, we focus on utilizing the speech comprehension and instruction-following capabilities of these MLLMs by providing speech and text prompts. Here, we present an overview of a simplified MLLM workflow, while noting that the specific implementation details may slightly vary given their structures. As illustrated in Fig. 1, for a given speech sample X and a text prompt P, the speech input X is first fed into a speech encoder (e.g., Whisper encoder), generating the speech representation $H_s = \text{SpeechEncoder}(X)$. Subsequently, a speech adapter module (e.g., fully-connected layer) maps H_s into H'_s via $H'_s = \text{SpeechAdapter}(H_s)$, aligning its dimensionality with that of the LLM's textual embeddings. Concurrently, the text prompt P is mapped into its textual embedding $H_t = \text{TextAdapter}(P)$ by a text adapter module (e.g., embedding layer). Both H'_s and H_t are then typically concatenated and fed into the LLM (e.g., Qwen). The LLM then generates the output tokens $Z = \{z_1, z_2, \ldots, z_k\}$ in an autoregressive manner, where each subsequent token is generated based on the preceding context: $z_i = \text{LLM}(H'_s, H_t, z_{<i})$. In this study, we specifically instruct the LLM to output either "AD" or "HC" as its final prediction.

Table 1. Statistics of the three datasets for AD detection.

Dataset	Language	Training set			Test set		
		Total	AD	HC	Total	AD	HC
ADReSS	English	108	54	54	48	24	24
PROCESS	English	117	56	61	40	19	21
iFLYTEK	Chinese	219	90	129	74	30	44

3.2. Prompt Design

To investigate the effectiveness of different prompt strategies, we design four prompt types, as detailed in Fig. 2, which incorporate varying levels of information and instructions.

The first type, **Basic**, directly instructs the model to perform AD detection based only on the provided speech, without any additional context. The second, **Contextual**, enriches the prompt with contextual information, including the subject's age, gender, language, and the specific cognitive task being performed (placeholders in curly braces are replaced with actual data at input time). This design aims to determine whether such context can aid the model in making more accurate predictions. The third type, **CoT**, explicitly instructs the model to follow an internal, step-by-step reasoning process before providing its final answer. We design a structured, three-step reasoning pathway to explore whether this can enhance the model's reasoning capabilities. Finally, **Contextual-CoT** integrates both the contextual information and the CoT reasoning.

To ensure the task remains a constrained binary classification, all four prompt types explicitly require the MLLMs to output only the word "AD" or "HC". Furthermore, to mitigate the MLLM's sensitivity to specific phrasing, an existing problem of LLMs [18], we employ ChatGPT-4o to generate ten paraphrased variants for each of the four prompt types to enhance robustness, while preserving their core semantic content. For each prompt type, we then select the top five performing variants based on their zero-shot accuracy on the training set. The final performance on the test set, for both zero-shot

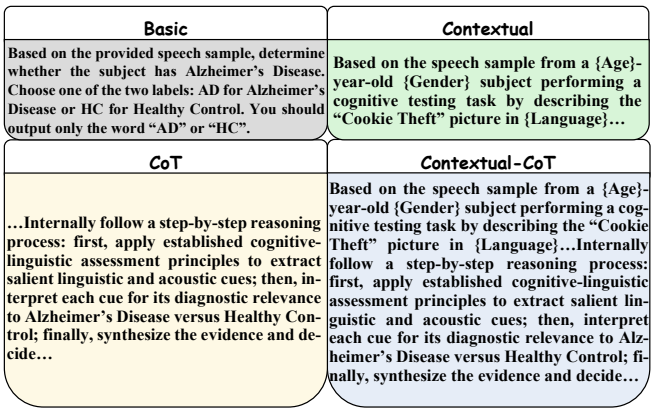

Fig. 2. The four designed prompt types. For brevity, "..." denotes content identical to the corresponding part of the Basic prompt.

evaluation and instruction tuning, is determined by a majority vote among these five selected variants.

3.3. Speech Cue–Augmented Prompting (SCAP)

To enrich prompts with fine-grained, AD-related information, we propose SCAP, an approach inspired by [19]. SCAP is a two-stage pipeline designed to automatically extract crucial speech cues and convert them into natural-language descriptors that can be prepended to any prompt. This process enhances the MLLM's contextual understanding of the speaker's cognitive state. The two main components of SCAP are the Speech Cue Extractor (SCE) and the Description Generator (DG).

The **SCE** is responsible for computing numerical values for four categories of speech cues that are known to be associated with AD. Specifically, we extract: (1) **Speech Timing Features**, which capture the temporal dynamics of speech and are supported by prior work [20]. Using a Voice Activity Detector, we first segment each audio sample into speech and pause segments. Subsequently, we compute a set of 11 metrics: the count of pause segments per second, the count of speech segments per second, the ratio of pause count to speech count, and statistical metrics (maximum, minimum, mean, and standard deviation) for the duration of both pause and speech segments. (2) **Filled-pause Rate**, a well-established biomarker for cognitive decline [21], is computed as the ratio of filled pauses to the total word count in the manual transcription. We identify English fillers ("uh", "um", "er", and "ah") and Mandarin fillers ("呃(e)", "嗯(en)", "啊(a)", "然后(ranhou)", "就是(jiushi)", "那个(neige)", and "这个(zhege)"). (3) **ASR Error Profile**. Building on our prior work showing that ASR errors can provide valuable cues for AD detection [22], we first fine-tune five sizes of the Whisper model [23] (tiny, base, small, medium, large) on data from DementiaBank. The profile then consists of four metrics: word error rate, substitution rate, deletion rate, and insertion rate, each averaged across the transcripts generated by these five fine-tuned ASR models. (4) **Acoustic Features**. Based on established literature [24], we use Praat to extract a set of classic acoustic features associated with dementia, including Pitch, Intensity, Jitter, and Shimmer.

The **DG** then converts these continuous numerical values into discrete natural-language descriptors (e.g., "Low", "Medium", "High"). This conversion is achieved using tertile-based thresholds derived from the training set's data distribution. Specifically, for each cue, we determine the 33rd (T_{33}) and 66th (T_{66}) percentile values from its distribution in the training set. A given sample's cue value v is then mapped to "Low" if $v < T_{33}$, "Medium" if $T_{33} \leq v \leq T_{66}$, and "High" if $v > T_{66}$. This process is applied to all extracted cues. The individual descriptors are then concatenated to form a single descriptive sentence that is prepended to the main task prompt. For example, a generated sentence might be: *"The speech exhibits a high filled-pause rate and a low pitch..."*

3.4. Instruction Tuning

While zero-shot prompting enables MLLMs to perform AD detection directly without any fine-tuning, we apply instruction tuning to further improve their performance. We utilize LoRA for efficient tuning of each MLLM. The instruction-response pairs for this process are constructed using the best-performing prompt type from our zero-shot experiments (CoT-Contextual, both with and without SCAP) and the corresponding ground-truth labels. To ensure generalization, we investigate a single-source fine-tuning strategy: each MLLM is tuned on a single dataset (e.g., the ADReSS dataset) and subsequently evaluated on all three datasets to rigorously test its cross-lingual and cross-dataset transfer capabilities.

4. EXPERIMENTS

4.1. Experimental Setup

We conducted experiments in two settings (zero-shot inference and instruction-tuning), performing an ablation study in each by comparing model performance with and without SCAP. For instruction-tuning, we employed LoRA with a rank of 8 and an alpha of 32, applying adapters to all linear layers of the LLM while keeping all other parameters frozen. The models were trained for 10 epochs with a batch size of 8, using the AdamW optimizer, a learning rate of 1×10^{-4}, and a cosine annealing scheduler. Performance was evaluated using classification accuracy. All experiments were conducted on NVIDIA RTX 4090 and A100 GPUs.

4.2. Experimental Results and Analysis

4.2.1. Supervised Baseline Results

We compared our method against three supervised baselines from recent literature: (1) **eGeMAPS+Naïve Bayes** [25], a classifier trained on the eGeMAPS feature set; (2) **Wav2Vec2+Linguistic+MLP** [26], an MLP classifier trained on the concatenation of dimension-reduced Wav2Vec2 embeddings and linguistic features from manual transcripts; and (3) **Whisper+MLP** [27], an MLP trained on Whisper-large encoder embeddings. The results, presented in Table 2, reveal several key findings. First, all baselines exhibit a significant performance drop in OOD scenarios (cross-lingual and cross-dataset), indicating poor generalization under domain shift. Second, performance on PROCESS is generally lower than on ADReSS, potentially because merging MCI samples (a transitional

Table 2. AD detection accuracy (%) of supervised baseline models. Each model is trained on a single source dataset and evaluated on all three to assess OOD generalization. Bold indicates ID results.

Baseline Model	Year	Training Set	Test Set		
			ADReSS	PROCESS	iFLYTEK
eGeMAPS+Naïve Bayes [25]	2023	ADReSS	**62.50**	55.00	52.70
		PROCESS	54.17	**60.00**	51.35
		iFLYTEK	52.08	52.50	**68.92**
Wav2Vec2+Linguistic+MLP [26]	2024	ADReSS	**70.83**	62.50	56.76
		PROCESS	58.33	**67.50**	55.41
		iFLYTEK	56.25	55.00	**74.32**
Whisper+MLP [27]	2025	ADReSS	**79.17**	65.00	59.46
		PROCESS	64.58	**70.00**	58.11
		iFLYTEK	60.42	62.50	**81.08**

Table 3. Zero-shot AD detection accuracy (%) of MLLMs, comparing performance with and without SCAP. The comparison spans three MLLMs and three test sets, with results for four prompt types detailed in separate columns. Arrows indicate performance change relative to the "w/o SCAP". Bold indicates the best result per dataset.

Model	SCAP	Test on ADReSS				Test on PROCESS				Test on iFLYTEK			
		Basic	Contextual	CoT	Contextual-CoT	Basic	Contextual	CoT	Contextual-CoT	Basic	Contextual	CoT	Contextual-CoT
MiDashengLM	w/o	52.08	52.08	54.17	56.25	50.00	55.00	57.50	55.00	54.05	58.11	58.11	60.81
	w/	54.17↑	56.25↑	56.25↑	58.33↑	52.50↑	57.50↑	55.00↓	57.50↑	55.41↑	58.11→	56.76↓	60.81→
Qwen2-Audio	w/o	52.08	54.17	56.25	56.25	52.50	55.00	52.50	57.50	55.41	56.76	59.46	62.16
	w/	52.08→	56.25↑	58.33↑	60.42↑	52.50→	57.50↑	55.00↑	60.00↑	56.76↑	58.11↑	59.46→	63.51↑
Qwen2.5-Omni	w/o	54.17	58.33	58.33	62.50	55.00	55.00	57.50	60.00	55.41	58.11	62.16	67.57
	w/	58.33↑	62.50↑	62.50↑	**66.67**↑	55.00→	57.50↑	60.00↑	**62.50**↑	58.11↑	62.16↑	66.22↑	**71.62**↑

state with ambiguous boundaries) into the AD class introduces label noise. Finally, the "Whisper+MLP" model achieves the best performance among the baselines in both ID and OOD settings.

4.2.2. Zero-shot Prompting Results

Table 3 presents the zero-shot performance of the MLLMs. From this table, we observe several key trends. First, for the experiments **without** SCAP, we observe that: ***1)*** The Contextual and CoT prompts consistently outperform the Basic prompt across nearly all MLLMs and datasets. ***2)*** The CoT-Contextual prompt almost always yields the best performance, confirming that providing both context and a structured reasoning path is highly effective. ***3)*** Among the MLLMs, Qwen2.5-Omni generally achieves superior results. Second, **with** the introduction of our proposed SCAP method, we find that: ***1)*** Performance is consistently and significantly improved (as indicated by the red arrows) across nearly all MLLMs, datasets, and prompt types. ***2)*** Most notably, the zero-shot performance of Qwen2.5-Omni with SCAP using the CoT-Contextual prompt surpasses the supervised "eGeMAPS+Naïve Bayes" baseline on all three datasets (ADReSS: 66.67% vs. 62.5%, PROCESS: 62.5% vs. 60%, iFLYTEK: 71.62% vs. 68.92%), demonstrating both the strong generalization potential of our approach and the powerful out-of-the-box capabilities of modern MLLMs. These results indicate that incorporating expert-inspired speech cues in the prompt allows the MLLM to better identify AD-related anomalies.

Table 4. AD detection accuracy (%) of MLLMs after instruction tuning with the Contextual-CoT prompt. It compares results on three test sets after tuning on different source datasets, with vs. without SCAP. Arrows indicate performance change relative to the "w/o SCAP". Bold indicates ID results.

Model	SCAP	Tuning Set	Test Set		
			ADReSS	PROCESS	iFLYTEK
MiDashengLM	w/o	ADReSS	**62.50**	60.00	60.81
		PROCESS	58.33	**62.50**	62.16
		iFLYTEK	56.25	57.50	**66.22**
	w/	ADReSS	**64.58**↑	60.00→	62.16↑
		PROCESS	60.42↑	**62.50**→	62.16→
		iFLYTEK	60.42↑	60.00↑	**68.92**↑
Qwen2-Audio	w/o	ADReSS	**79.17**	60.00	64.86
		PROCESS	62.50	**65.00**	63.51
		iFLYTEK	60.42	60.00	**81.08**
	w/	ADReSS	**81.25**↑	62.50↑	64.86→
		PROCESS	64.58↑	**67.50**↑	67.57↑
		iFLYTEK	62.50↑	60.00→	**83.78**↑
Qwen2.5-Omni	w/o	ADReSS	**79.17**	62.50	71.62
		PROCESS	64.58	**67.50**	70.27
		iFLYTEK	64.58	62.50	**82.43**
	w/	ADReSS	**83.33**↑	67.50↑	72.97↑
		PROCESS	68.75↑	**72.50**↑	70.27→
		iFLYTEK	64.58→	65.00↑	**85.14**↑

4.2.3. Instruction Tuning Results

Table 4 shows the performance of MLLMs after instruction tuning with the Contextual-CoT prompt. First, for the results **without** SCAP, we can observe two primary benefits of instruction tuning: ***1)*** As expected, tuning on a single dataset substantially boosts ID performance over the zero-shot results. For instance, tuning Qwen2.5-Omni on ADReSS boosted its accuracy from a zero-shot performance of 62.5% to 79.17%. ***2)*** More importantly, instruction tuning also enhances OOD generalization. For example, Qwen2.5-Omni tuned on ADReSS achieved 62.5% on PROCESS and 71.62% on iFLYTEK, improving upon its zero-shot accuracy of 60% and 67.57%, respectively. Second, **with** the incorporation of SCAP during the instruction tuning process, we find that: ***1)*** The benefits are amplified across both ID and OOD settings. For instance, for Qwen2.5-Omni tuned on ADReSS, incorporating SCAP improves performance on all three test sets (ADReSS: 83.33% vs. 79.17%, PROCESS: 67.5% vs. 62.5%, iFLYTEK: 72.97% vs. 71.62%). ***2)*** Ultimately, the instruction-tuned Qwen2.5-Omni with SCAP significantly outperforms the strongest supervised baseline (Whisper+MLP) in both ID and OOD settings. For example, when both models are trained on ADReSS, our approach surpasses the baseline across all three test sets: ADReSS (83.33% vs. 79.17%), PROCESS (67.5% vs. 65%), and iFLYTEK (72.97% vs. 59.46%). These results demonstrate that MLLMs instruction-tuned with SCAP on a single ID dataset can achieve not only strong domain-specific performance but also robust OOD generalization.

5. CONCLUSIONS

In this paper, we present a comprehensive study on leveraging MLLMs for robust, cross-lingual AD detection. We first demonstrate that zero-shot MLLM performance can be significantly improved through sophisticated prompt engineering, identifying a Contextual-CoT prompt as the most effective. We then propose SCAP, a method that enriches prompts with natural-language descriptors of AD-related speech cues, and show that this approach enables zero-shot MLLMs to surpass supervised models. Furthermore, we demonstrate that instruction-tuning MLLMs with SCAP via LoRA yields superior performance, outperforming all supervised baselines and dramatically improving ID accuracy and, more critically, strengthening OOD generalization. Our findings indicate that MLLMs, when guided by carefully designed prompts and lightweight tuning, can be transformed into powerful, robust, and generalizable tools for AD detection. Future work involves exploring a wider range of AD-related speech biomarkers and designing MLLMs tailored for AD detection.

6. REFERENCES

[1] Peter J Nestor, Philip Scheltens, and John R Hodges, "Advances in the early detection of Alzheimer's Disease," *Nature medicine*, vol. 10, no. Suppl 7, pp. S34–S41, 2004.

[2] Ellen Elisa De Roeck, Peter Paul De Deyn, Eva Dierckx, and Sebastiaan Engelborghs, "Brief cognitive screening instruments for early detection of Alzheimer's Disease: a systematic review," *Alzheimer's research & therapy*, vol. 11, no. 1, pp. 21, 2019.

[3] Laura Calzà, Gloria Gagliardi, Rema Rossini Favretti, and Fabio Tamburini, "Linguistic features and automatic classifiers for identifying mild cognitive impairment and dementia," *Computer Speech & Language*, vol. 65, pp. 101113, 2021.

[4] Yilin Pan, Bahman Mirheidari, Daniel Blackburn, and Heidi Christensen, "A two-step attention-based feature combination cross-attention system for speech-based dementia detection," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 33, pp. 896–907, 2025.

[5] Yuanchao Li, Zixing Zhang, Jing Han, Peter Bell, and Catherine Lai, "Semi-supervised cognitive state classification from speech with multi-view pseudo-labeling," in *Proc. ICASSP*, 2025, pp. 1–5.

[6] Yin-Long Liu, Yuanchao Li, Rui Feng, Liu He, Jia-Xin Chen, Yi-Ming Wang, Yu-Ang Chen, Yan-Han Peng, Jia-Hong Yuan, and Zhen-Hua Ling, "Leveraging Cascaded Binary Classification and Multimodal Fusion for Dementia Detection through Spontaneous Speech," in *Proc. Interspeech*, 2025, pp. 544–548.

[7] Saturnino Luz, Fasih Haider, Sofia de la Fuente Garcia, Davida Fromm, and Brian MacWhinney, "Alzheimer's dementia recognition through spontaneous speech: The ADReSS challenge," in *Proc. Interspeech*, 2020, pp. 2172–2176.

[8] Fuxiang Tao, Bahman Mirheidari, Madhurananda Pahar, Sophie Young, Yao Xiao, Hend Elghazaly, Fritz Peters, Caitlin Illingworth, Dorota Braun, Ronan O'Malley, et al., "Early dementia detection using multiple spontaneous speech prompts: The PROCESS Challenge," in *Proc. ICASSP*, 2025, pp. 1–2.

[9] Siddhant Arora, Kai-Wei Chang, Chung-Ming Chien, Yifan Peng, Haibin Wu, Yossi Adi, Emmanuel Dupoux, Hung-Yi Lee, Karen Livescu, and Shinji Watanabe, "On the landscape of spoken language models: A comprehensive survey," *arXiv preprint arXiv:2504.08528*, 2025.

[10] Catarina Botelho, John Mendonça, Anna Pompili, Tanja Schultz, Alberto Abad, and Isabel Trancoso, "Macro-descriptors for Alzheimer's Disease detection using large language models," in *Proc. Interspeech*, 2024, pp. 1975–1979.

[11] Chanwoo Park, Anna Seo Gyeong Choi, Sunghye Cho, and Chanwoo Kim, "Reasoning-Based Approach with Chain-of-Thought for Alzheimer's Detection Using Speech and Large Language Models," in *Proc. Interspeech*, 2025, pp. 2185–2189.

[12] Mostafa Shahin, Beena Ahmed, and Julien Epps, "Zero-Shot Cognitive Impairment Detection from Speech Using AudioLLM," *arXiv preprint arXiv:2506.17351*, 2025.

[13] Zhaoci Liu, Zhiqiang Guo, Zhenhua Ling, Shijin Wang, Lingjing Jin, and Yunxia Li, "Dementia detection by analyzing spontaneous mandarin speech," in *Proc. APSIPA*, 2019, pp. 289–296.

[14] Harold Goodglass and Edith Kaplan, *Boston diagnostic aphasia examination booklet*, Lea & Febiger, 1983.

[15] Heinrich Dinkel, Gang Li, Jizhong Liu, Jian Luan, Yadong Niu, Xingwei Sun, Tianzi Wang, Qiyang Xiao, Junbo Zhang, and Jiahao Zhou, "MiDashengLM: Efficient Audio Understanding with General Audio Captions," *arXiv preprint arXiv:2508.03983*, 2025.

[16] Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, et al., "Qwen2-Audio Technical Report," *arXiv preprint arXiv:2407.10759*, 2024.

[17] Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, et al., "Qwen2.5-Omni Technical Report," *arXiv preprint arXiv:2503.20215*, 2025.

[18] Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr, "Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I learned to start worrying about prompt formatting," in *proc. ICLR*, 2024.

[19] Yuanchao Li, Yuan Gong, Chao-Han Huck Yang, Peter Bell, and Catherine Lai, "Revise, reason, and recognize: LLM-based emotion recognition via emotion-specific prompts and asr error correction," in *Proc. ICASSP*, 2025, pp. 1–5.

[20] Kangdi Mei, Xinyun Ding, Yinlong Liu, Zhiqiang Guo, Feiyang Xu, Xin Li, Tuya Naren, Jiahong Yuan, and Zhenhua Ling, "The USTC System for ADReSS-M Challenge," in *proc. ICASSP*, 2023, pp. 1–2.

[21] Jiahong Yuan, Yuchen Bian, Xingyu Cai, Jiaji Huang, Zheng Ye, and Kenneth Church, "Disfluencies and fine-tuning pre-trained language models for detection of Alzheimer's Disease.," in *Proc. Interspeech*, 2020, pp. 2162–2166.

[22] Yin-Long Liu, Rui Feng, Ye-Xin Lu, Jia-Xin Chen, Yang Ai, Jia-Hong Yuan, and Zhen-Hua Ling, "Can Automated Speech Recognition Errors Provide Valuable Clues for Alzheimer's Disease Detection?," in *Proc. ICASSP*, 2025, pp. 1–5.

[23] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proc. ICML*, 2023, pp. 28492–28518.

[24] Juan José G Meilán, Francisco Martínez-Sánchez, Juan Carro, Dolores E López, Lymarie Millian-Morell, and José M Arana, "Speech in Alzheimer's Disease: can temporal and acoustic parameters discriminate dementia?," *Dementia and geriatric cognitive disorders*, vol. 37, no. 5-6, pp. 327–334, 2014.

[25] Saturnino Luz, Fasih Haider, Davida Fromm, Ioulietta Lazarou, Ioannis Kompatsiaris, and Brian MacWhinney, "Multilingual Alzheimer's dementia recognition through spontaneous speech: a signal processing grand challenge," in *Proc. ICASSP*, 2023, pp. 1–2.

[26] Saturnino Luz, Sofia De La Fuente Garcia, Fasih Haider, Davida Fromm, Brian MacWhinney, Alyssa Lanzi, Ya-Ning Chang, Chia-Ju Chou, and Yi-Chien Liu, "Connected Speech-Based Cognitive Assessment in Chinese and English," in *Proc. Interspeech*, 2024, pp. 947–951.

[27] Yifan Gao, Long Guo, and Hong Liu, "Leveraging multimodal methods and spontaneous speech for Alzheimer's Disease identification," in *Proc. ICASSP*, 2025, pp. 1–2.