Leveraging Large Language Models for Identifying Interpretable Linguistic Markers and Enhancing Alzheimer's Disease Diagnostics

Tingyu Mo [1], Jacqueline C. K. Lam* [1], Victor O. K. Li* [1], and Lawrence Y. L. Cheung [2]

[1]Department of Electrical and Electronic Engineering, The University of Hong Kong

[2]Department of Linguistics and Modern Languages, Chinese University of Hong Kong

Abstract

Alzheimer's Disease (AD) is a progressive irreversible neurodegenerative disorder. Early AD detection is crucial for timely intervention. This study proposes a novel LLM framework to identify interpretable linguistic markers from LLMs and incorporate them to supervised AD detection transformers, while evaluating corresponding model performance and interpretability. Our work carries three major novelties: First, we design in-context Few-Shot and Zero-Shot prompting strategies to facilitate LLMs in summarising the high-level linguistic markers discriminative of AD from Normal Control (NC), providing interpretation and assessment of their strength, reliability and relevance for AD classification. Second, we incorporate LLM-summarised linguistic markers into a smaller transformer model to enhance the performance of AD detection. Third, we investigate whether the LLM-summarised linguistic markers can enhance accuracy and interpretability of our downstream supervised transformer model when used with the original speech transcripts. Our findings have shown that using LLM-summarised linguistic markers solely may yield lower accuracy using supervised learning, while combining these markers with the features extracted from the original speech transcripts can enhance the AD detection model's diagnostic capabilities. Our hybrid speech-based AD detection framework capitalizes both the interpretability of linguistic markers and the rich information embedded within the original transcripts. When used as complementary features, the distinguishable AD linguistic markers have enhanced the AD detection performance, while providing clinically meaningful insights into speech-based AD detection. These insights have enabled healthcare professionals to gain deeper understandings about what linguistic variations have dictated individuals suffering from AD, enabling the clinical professionals to make more informed interventions

*Corresponding authors: V.O.K. Li (vli@eee.hku.hk), J.C.K. Lam (jcklam@eee.hku.hk)

and healthcare decision-makings.

Keywords: LLM-based Linguistic Markers Extraction and Identification; Zero-shot and Few-shot Prompting Strategies; Interpretability-accuracy Tradeoff; LLM-based Linguistic Interpretation Generation; Linguistic-based AD Diagnosis

1 Introduction

Alzheimer's Disease is a devastating neurodegenerative disorder that progressively impairs cognitive function, affecting millions of individuals worldwide. Timely and precise diagnosis of AD is of utmost importance, as it enables early therapeutic intervention and optimal healthcare management, potentially improving patient outcomes and their qualities of life. Recent advancements in natural language processing (NLP)[24,26] and deep learning offer new opportunities for developing automated tools to assist the detection and diagnosis of AD[1].[2,27,28] NLP models, such as BERT (Bidirectional Encoder Representations from Transformers),[4] have achieved remarkable performance in various NLP tasks, including language understanding, generation and classification. However, their applications in healthcare contexts demand careful balance between model performance and interpretability: First, it is necessary for clinicians to interpret and validate diagnostic model findings against clinical decisions, based on a conventional set of clinical indicators. Second, accurate early-stage AD detection is dependent on clear, interpretable clinical biomarkers that discriminate cognitive decline from normal aging. Third, healthcare providers must be able to communicate findings effectively with patients, particularly when it is suitable for early intervention. While more complex NLP models might detect subtle linguistic patterns at a higher accuracy, their "black box" nature may limit clinical adoption. Conversely, more interpretable models might sacrifice statistical performance, though they may offer clearer insights for clinical validation, thus enhancing the potential of integrating the model results with the process of clinical diagnostics.

Traditional NLP methods for AD detection can achieve reasonable accuracy in distinguishing AD from NC. Automated speech analyses[11] have shown promising results in predicting AD progression, achieving an accuracy of 78.5%. However, they are weak in capturing /interpreting the underlying linguistic patterns. The complex nature of linguistic change in AD requires more sophisticated models that not only detects linguistic variations but also articulates them in meaningful ways that are intelligible by the healthcare professionals. Additionally, the limited availability of annotated speech data from individuals suffering from AD pose a substantial challenge for conventional supervised learning models operated on

large datasets. Large Language models (LLMs) offer a promising direction for addressing these challenges through Few-Shot and Zero-Shot in-context learning, capable of learning based on limited data size/samples. LLMs are deep learning models trained on vast amounts of textual data, enabling them to capture rich linguistic information by retrieving domain-specific knowledge from big clinical datasets. More importantly, they are capable of identifying and interpreting high-level linguistic patterns which may be used as interpretable indicators of cognitive decline, complementing existing diagnostic biomarkers. Interpretability is essential for integrating model results with clinical practice, as healthcare providers need clear, explainable evidence to support clinical diagnostic decisions. For instance, interpretable linguistic biomarkers of AD identified from interpretable LLM models, such as repetition and coherence, can potentially serve as clinical linguistic biomarkers to assist clinicians in making more informed AD diagnostics.

In this paper, we propose a novel approach for AD detection based on interpretable linguistic biomarkers summarised from LLMs. Our method leverages Few-Shot and Zero-Shot learning techniques to enable LLMs to identify high-level meaningful linguistic patterns from limited transcription data. We design a set of prompting strategies to guide LLMs in summarising key linguistic features and generate explanations. Furthermore, we incorporate LLM-summarised linguistic biomarkers into LLMs for downstream supervised learning, with the final goal of assessing AD versus NC classification performance (for AD detection). While LLMs have achieved impressive results in various NLP performance tasks, their AD classification performance is still lagging behind traditional supervised learning approaches, given the limited data constraint. To overcome this challenge, we propose a novel AD detection architecture that capitalizes on the strengths of both LLMs and supervised learning techniques. Our main contributions and novelties are as follows:

- Use Large Language Models (LLMs) to extract interpretable linguistic markers for AD vs. NC classification, or AD detection, using in-context Few-Shot and Zero-Shot prompting strategies.
- Design different prompting strategies under different learning contexts to facilitate LLMs in identifying high-level linguistic markers discriminative of AD/NC, providing interpretation and assessment, in terms of their strength, reliability and relevance to AD vs. NC classification, or AD detection.
- Incorporate linguistic markers summarised by LLMs into a supervised-learning model to enhance the performance of downstream supervised learning for AD vs. NC classification, or AD detection, by fusing high-level linguistic markers/features summarised from LLMs with the original speech

transcript.

The remainder of this paper is organized as follows. Section 2 provides an overview of the related work covering AD classification/detection using NLP and deep learning techniques. Section 3 describes the methodology, including prompt engineering strategies, and outlines our proposed model framework for evaluation. Section 4 presents the experimental results and discusses the performance of our proposed model. Finally, Section 5 and Section 6 concludes the paper and outlines the potential for future studies.

2 Related Work

Recently, there has been a growing interest in using linguistic biomarkers in AD diagnostics. Linguistic markers refer to the distinctive linguistic features and patterns that are indicative of cognitive decline. Several studies have explored the potential of linguistic markers in AD detection. Fraser et al.[5] analyzed a wide range of linguistic features, including grammatical complexity, semantic content, and discourse coherence, in the language samples of AD patients and healthy controls. They found that a combination of linguistic features could effectively distinguish between AD and normal aging. Similarly, Orimaye et al.[6] used natural language processing techniques to extract syntactic, lexical, and semantic features from the transcripts of AD patients and achieved promising results in AD classification. Other studies have focused on specific linguistic aspects, such as speech fluency, word-finding difficulties, and semantic fluency. For example, Pakhomov et al.[19] investigated the use of verbal fluency tests as a screening tool for frontotemporal lobar degeneration, while demonstrating their effectiveness in detecting cognitive impairment. Kavé and Goral[7] explored the semantic and phonemic fluency deficits in AD and highlighted their potentials to serve as linguistic markers of AD.

In the medical domain, LLMs have been used for tasks such as medical language translation, clinical note summarization, and medical question answering. For example, Rasmy et al.[10] developed a medical language model called MED-BERT, which was trained on a large corpus of medical text and achieved state-of-the-art performance on several medical natural language processing tasks. Some studies have explored the potential of LLMs in mental health analysis. Yang et al.[13,14] proposed interpretable mental health analysis frameworks using LLMs to identify linguistic patterns and generate explanations from text data. Xu et al.[15] introduced ExpertPrompting, instructing LLMs to behave as domain experts, enhancing performance and reliability in mental health applications. Luo et al.[16] explored ChatGPT's capability in evaluating factual inconsistencies, highlighting the potential of LLMs in assessing the reliability of

generated insights. Jeon et al.[17] proposed a dual-prompting approach to improve the interpretability and reliability of mental health analysis using LLMs. These studies have demonstrated the growing interest in leveraging LLMs for accurate, interpretable, and trustworthy mental health analyses. Our research builds upon these advancements, exploring the use of LLMs for identifying interpretable linguistic markers in Alzheimer's Disease detection. We evaluate the feasibility of using LLMs to identify interpretable linguistic markers for detection of AD based on Few-Shot and Zero-Shot learning. We design various prompting strategies under different learning contexts to facilitate LLMs in identifying linguistic biomarkers discriminative of AD, assessing their strength and reliability, and providing interpretation concerning their relevance to AD detection. We use novel linguistic biomarkers summarised by LLMs to enhance the performance of downstream supervised learning, by proposing a novel framework that incorporates the LLM-identified markers into a multimodal AD detection model. Our work contributes significantly to the growing body of literature covering LLM-assisted AD diagnostics.

3 Methodology

We propose a feasible approach to evaluate linguistic markers identified by LLMs for Alzheimer's disease (AD) detection. We design two prompting strategies that leverage the knowledge and capabilities of LLMs to analyze the linguistic patterns that govern individual speeches.

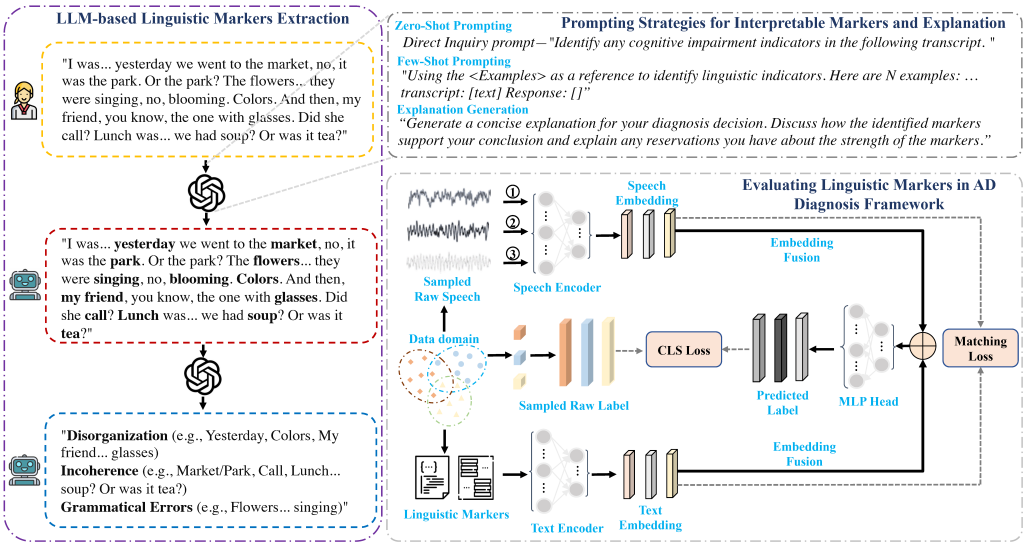


Figure 1: A framework for leveraging LLMs for identifying interpretable linguistic markers and enhancing Alzheimer's Disease diagnostics

3.1 LLM Configuration for Model Development

We explore the capabilities and applications of several leading LLMs, including ChatGPT3.5, GPT4, GPT4o, to extract linguistic patterns and markers relevant to Alzheimer's disease (AD) diagnosis. These models are representatives in LLM technology, each having unique characteristics and strengths in generating and analyzing human languages.

• **GPT3.5**: As a version of GPT (Generative Pre-trained Transformer) series developed by OpenAI, it serves as one of the most powerful language models available during the time of its release. It excels in generating human-like text, completing given text prompts at high coherence, and understanding subtle linguistic nuances. We leverage GPT-3.5's capability to identify the linguistic patterns and anomalies indicative of cognitive impairment or AD.

• **GPT4/4o**: GPT-4/4o is the more sophisticated version fine-tuned for better understanding and generating natural language. While specific details regarding its size and architecture are proprietary, it is known for its enhanced ability to process and generate text across a wide range of languages and dialects, providing more accurate and contextually relevant responses. While GPT-4 excels at text processing and generation, GPT-4o is designed for comprehensive multimodal interaction, accepting inputs in text, audio, image, and video formats while generating outputs in text, audio, and image modalities. We evaluate GPT4/4o's advanced language understanding to extract more refined and contextually-aware linguistic markers.

3.2 LLM-based Prompting Strategies for Linguistic Markers Identification

We propose a novel approach to identify linguistic markers for Alzheimer's Disease (AD) detection using LLMs. We design a prompting strategy that leverages the knowledge and capabilities of LLMs to analyze linguistic patterns in descriptions of the Cookie Theft picture from the Boston Diagnostic Aphasia Examination. The methodology is outlined in Algorithm 1.

As shown in Line 3-4, the prompting process begins by assigning an expert identity[15] to the LLM, portraying it as a medical professional tasked with analyzing linguistic patterns in the Cookie Theft picture descriptions. This identity establishment helps to guide the LLM's behavior and outputs towards the desired task. Next, we assign the specific task to the LLM, which is to identify linguistic indicators that distinguish between individuals with AD and those with normal control. We emphasize the importance of being cautious and avoiding overdiagnosis of AD based solely on linguistic patterns. We use two prompting strategies to enable LLM's potential abilities, including Zero-Shot prompting and Few-Shot

prompting.

• **Zero-Shot Prompting**: Zero-Shot learning[9] enables LLMs to apply learned knowledge from one domain to perform tasks in another without any task-specific training data. We aim to prompt the general LLMs to describe linguistic attributes found in transcripts, then relate these attributes to known symptoms or stages of AD without explicit example-based guidance. Zero-Shot prompting strategy includes two approaches: a. Inferential Question Prompt: This kind of prompt directly asks the LLM to infer cognitive decline based on the language patterns observed in the transcript. For example, "What language patterns in this transcript suggest cognitive decline?" This prompt encourages the LLM to identify and analyze linguistic features that may indicate cognitive impairment. b. Direct Inquiry Prompt: This prompt explicitly instructs the LLM to identify any indicators of cognitive impairment in the given transcript. For example, "Identify any cognitive impairment indicators in the following transcript." This prompt directs the LLM's attention towards detecting specific linguistic markers associated with cognitive decline. In our method, we simply adopt Direct Inquiry Prompt to instruct the LLM, as shown in Line 11 in Algorithm.1 as follows:

"Identify linguistic patterns, keywords, or phrases that could potentially indicate cognitive impairment or Alzheimer's disease. However, be cautious and consider whether these indicators are strong enough to warrant a diagnosis on their own. Present these indicators and your assessment of their strengths in a list format."

• **Few-Shot Prompting**: Few-Shot learning[8] involves training the LLM models with extremely limited labelled data. We aim to fine-tune the LLMs with AD-specific linguistic markers using limited data from transcripts. We design a general LLM model without specifying any specific domains using Few-Shot prompting. Few-Shot prompting is conducted using extremely limited speech transcript data collected from AD and NC cohorts, followed by a question that seeks to identify similar linguistic patterns in the much larger dataset employed for LLM model training. Few-Shot prompting is conducted based on the following Chain-of-Thoughts sequences:

"<Examples> Transcripts from people with Alzheimer's Disease describing the Cookie Theft picture: Script: [example1]...[exampleK]

<Examples> Transcripts from people who are normal control describing the Cookie Theft picture: Script: [example1]...[exampleK]".

In this prompting sequence, the LLM model is provided with a small number of example transcripts (K examples) indicating the presence or absence of a specific condition (e.g., AD or NC). By employing the Zero-Shot and Few-Shot prompting strategies, we aim to evaluate whether LLMs can leverage domain-

specific knowledge and extract linguistic markers that discriminate AD from NC. The summarised linguistic markers will be utilized in downstream supervised learning for examining AD detection performance.

3.3 LLM-based Interpretable Explanation Generation

To analyse further the LLM-identified linguistic markers and enhance model interpretability, we assign two novelites/tasks for the LLM-based AD detection model. As shown in Line 13 in Algorithm 1, Task (b) focuses on making a diagnostic decision based on the LLM-identified linguistic indicators and their AD classification strength. The prompt is given as follows:

"Based on the identified linguistic indicators and their strength, make a diagnosis decision (Alzheimer's Disease or normal control) for the individual who provided the description. If the indicators are not strong or conclusive enough, lean towards a diagnosis of Not Sure".

We instruct the LLM to classify the individuals who provided the Cookie Theft picture description as either suffering from Alzheimer's Disease (AD) or cognitively normal, which we termed as Normal Control (NC). However, we emphasize the importance of caution in this decision-making process. If the LLM-identified linguistic indicators are not strong or conclusive enough to confidently determine whether an individual has suffered from AD, we guide the LLM to lean towards a decision of "Not Sure." This approach aims to prevent over-diagnosis and ensures that the LLM has taken into account the reliability and significance of linguistic markers before making any definitive AD classification.

As shown in Line 14 in Algorithm.1, Task (c) involves generating a concise explanation for the diagnosis decision made in Task (b). Our prompt for this task is as follows:

Generate a concise explanation for your diagnosis decision. Discuss how the identified keywords, phrases, and patterns from the transcript support your conclusion, and explain any reservations you have about the strength of the indicators, while emphasizing the importance.

We prompt the LLM to discuss how the identified keywords, phrases, and patterns from the transcript support its conclusion. The LLM-based explanation should highlight the specific linguistic features that contribute to the classification decision, providing insights into the reasoning process of the LLM. Additionally, we instruct the LLM to provide any reservations or uncertainties regarding the strength of the linguistic indicators. By doing so, we encourage the LLM to critically assess the reliability and capture the richness and complexity of the linguistic markers, while disclosing the limitations or ambiguities in the concerned LLM-based decision-making process. This step highlights the cautious approach we

have taken in model design to avoid over-diagnosis, while promoting transparency and interpretability of model results.

Algorithm 1: Prompting LLMs for Analysing Linguistic Patterns in Cookie Theft Picture Descriptions

Input:
Designed prompt for summarizing linguistic markers; transcripts from dataset.
Output:
Linguistic markers summarized from LLM and explanation for summarized results;

Assigning an expert identity:
You are a medical professional tasked with analyzing linguistic patterns in descriptions of the Cookie Theft picture from the Boston Diagnostic Aphasia Examination.
Assigning task:
Your goal is to identify linguistic indicators that distinguish between individuals with Alzheimer's Disease and those with normal control, while being cautious not to overdiagnose Alzheimer's Disease.
In-context Few-Shot Learning
<Examples> Transcripts from people with Alzheimer's Disease describing the Cookie Theft picture:
Script: [example1]...[exampleK]
<Examples> Transcripts from people with normal control describing the Cookie Theft picture:
Script: [example1]...[exampleK]
Assigning more specific tasks:
Using the <Examples> as a reference, please analyze the following transcript [text]. please complete the following tasks:
a. Identify linguistic patterns, keywords, or phrases that could potentially indicate cognitive impairment or Alzheimer's Disease. However, be cautious and consider whether these indicators are strong enough to warrant a diagnosis on their own. Present these indicators and your assessment of their strength in a list format.
b. Based on the identified linguistic indicators and their strength, make a diagnosis decision (Alzheimer's Disease or normal control) for the individual who provided the description. If the indicators are not strong or conclusive enough, lean towards a diagnosis of Not Sure.
c. Generate a concise explanation for your diagnosis decision. Discuss how the identified keywords, phrases, and patterns from the transcript support your conclusion, and explain any reservations you have about the strength of the indicators, while emphasizing the importance of a cautious approach to avoid overdiagnosis.
Extra output regulation:
Please provide your response in a structured format, separating each task with clear headings and limiting your total response to no more than 150 words
Collect response from LLM;

3.4 Evaluating Linguistic Markers in AD Diagnostic Framework

We propose an evaluation framework to assess the effectiveness and interpretability of the linguistic markers summarised by the LLMs for AD detection. We aim to investigate whether these linguistic markers can replace the original transcripts in training AD detection models, while evaluating the performance of different prompting strategies and interpretability of our LLMs. We train the AD diagnostic

models, using the LLM-identified linguistic markers and interpretations as the input features, then compare their performance to other models trained on the original speech transcript data. We employ various evaluation metrics, including, accuracy, precision, recall, and F1 score, to assess the model performance. If the models trained on the linguistic markers have achieved comparable or better performance than those trained on the original transcripts, it indicates that the summarised markers have captured the relevant information for AD detection, thereby potentially obviating the less interpretable speech transcripts.

Referring to Fig.1, our framework consists of two main components: (1) the speech embedding branch and (2) the text embedding branch. As demonstrated in,[12] representation learning can improve the naturalness of speech-driven gesture generation. On the one hand, the speech embedding branch in our framework processes the sampled raw speech using a speech encoder to generate speech embeddings. On the other hand, the text embedding branch utilizes the summarised linguistic markers and their corresponding interpretations. The linguistic markers serve as an alternative representation of the original speech, capturing the relevant information for AD detection. These markers are then fed into the text encoder, to be converted into the text embeddings for the subsequent processing steps, which consist of embedding fusion and classification using the Multi-layer Perceptrons (MLP) head.

To train the AD detection model, we employ two loss functions: the cross-entropy loss and the matching loss. The cross-entropy loss is used to measure the discrepancy between the predicted labels and the ground truth labels, guiding the model to give accurate AD diagnosis predictions. The matching loss, on the other hand, is designed to align the speech embeddings with text embeddings. By minimizing the matching loss, we encourage the model to learn a shared representation space where the speech and text embeddings are closely aligned. The optimization objects are listed in follows:

$$\underset{\Theta_e,\Theta_f}{\operatorname{argmin}}\ \mathcal{L}_{\text{cls}} = \frac{1}{\|D_T\|}\sum^{\|D_T\|}\mathcal{L}_{ce}(l, f; \Theta_e, \Theta_f) \tag{3.1}$$

$$\underset{\Theta_e,\Theta_f}{\operatorname{argmin}}\ \mathcal{L}_{\text{matching}} = \frac{1}{\|D_T\|}\sum^{\|D_T\|}\sum_{\mathbf{e}s,\mathbf{e}t}\mathcal{L}cosine(\mathbf{e}s, \mathbf{e}t; \Theta e, \Theta_f) \tag{3.2}$$

$$\underset{\Theta}{\operatorname{argmin}}\ \mathcal{L}_{\text{total}} = \mathcal{L}_{\text{cls}} + \beta\mathcal{L}_{\text{matching}} \tag{3.3}$$

where f is the fused embedding, Θ_e and Θ_f are the parameters of encoders and MLP-Head for classification respectively. $\mathcal{L}_{ce}$ represents the cross-entropy loss and $\mathcal{L}_{\text{matching}}$ represents the matching loss.

$|D_T|$ denotes the size of the training dataset. $\mathbf{e}s$ and $\mathbf{e}t$ are the speech embedding and text embedding, respectively. $\mathcal{L}_{cosine}$ is the cosine distance based loss. β is the weight for balancing these objects.

4 Experiment

4.1 Evaluation Datasets

We utilize the ADReSSo Challenge 2021 dataset,[20] the Pitt corpus[33] and FDRD dataset to investigate the feasibility of using LLMs for identifying interpretable linguistic markers of AD and improving the performance of AD detection. The datasets deployed consist of spontaneous speech from diverse AD-related and NC cohorts, making it a valuable resource for advancing our understanding on what linguistic markers are characterizing symptomatic AD. More specifically, the speech recordings were obtained as a result of the Cookie Theft picture description task performed by different bodies, in accordance with the Boston Diagnostic Aphasia Examination. For this task, participants are shown a picture depicting a scene where a boy is stealing cookies from a jar, while his mother is washing her dishes. The scene also depicts other objects such as a sink overflowing with water and a boy trying hard to get hold of a cookie jar located at the top part of the shelf. Participants are asked to describe what they have observed in the picture, in order to collect spontaneous speech samples based on which their language abilities and cognitive functions can be evaluated.

Dataset	Number of Samples			Description Task Type
	Total	AD	NC	
ADReSSo	166	86	80	BDAE Cookie Theft
Pitt	549	306	243	BDAE Cookie Theft
FDRD	80	45	35	BDAE Cookie Theft

Table 1: An overview of the speech datasets adopted for this study. AD: Alzheimer's Disease; NC: Normal Control; BDAE: Boston Diagnostic Aphasia Examination

4.1.1 The ADReSSo Dataset

The ADReSSo Challenge 2021 dataset is widely used in the research community, serving as a benchmark for developing and evaluating methods for AD detection through speech analysis. These spontaneous speech samples have been divided into two main categories based on the cognitive labels: Alzheimer's Disease (AD) and Normal Control (NC). The AD category covers speech recordings and transcripts from individuals diagnosed as Alzheimer's Disease patients. The speech samples in this category exhibit linguistic patterns and features associated with AD. NC category includes the speech recordings and

the transcripts from individuals who are cognitively normal, meaning they do not display any signs of cognitive impairment or dementia. The speech samples in this category serve as the control samples and provide a baseline for comparing the linguistic characteristics of AD individuals. In our experiment, our dataset contains a total of 166 speech recordings and transcripts, with 86 samples in the AD category and 80 samples in the NC category.

4.1.2 The Pitt Corpus

The DementiaBank Pitt Corpus, part of the larger TalkBank project, offers a longitudinal perspective on cognitive decline through speech analysis. This dataset complements the ADReSSo dataset by providing a larger sample size and multiple recordings for some participants over time. Our experimental subset of the Pitt Corpus contains a total of 549 audio recordings from 291 unique participants. The audio recordings are divided into two main categories: 306 recordings from participants diagnosed with dementia and 243 recordings from cognitively normal participants. Similar to the ADReSSo dataset, participants have completed the Cookie Theft picture description task, providing consistency in the speech elicitation method across both datasets.

4.1.3 The FDRD Dataset

The FDRD dataset[1] from Cambridge Centre for Frontotemporal Dementia and Related Disorders (FDRD), a clinical database consisting of speech samples from participants suffering from various neurodegenerative conditions, as well as healthy controls. Complementary data include assessments from both baseline and follow-up visits, enabling longitudinal analysis. Participants underwent standardized picture description tasks using three validated stimuli: the Boston Diagnostic Aphasia Examination (BDAE) Cookie Theft picture, the Mini Linguistic State Examination (MLSE) beach scene, and the Western Aphasia Battery (WAB) picnic scene. The dataset covers multiple diagnostic groups, including semantic variant primary progressive aphasia (svPPA), non-fluent variant PPA (nfvPPA), logopenic variant PPA (lvPPA), mixed PPA, corticobasal syndrome (CBS), progressive supranuclear palsy (PSP), and Alzheimer's Disease (AD). In our study, we focus on two main groups: participants with AD (n=45, including those with multi-domain dementia, lvPPA, and typical amnestic AD) and neurologically healthy controls (n=35). Comprehensive participant information was collected, including age, sex, handedness, educational background, and cognitive performance as measured by the Addenbrooke's Cognitive Examination (ACE,

[1]The FDRD dataset used in this study was derived from Cambridge Centre for Frontotemporal Dementia and Related Disorders. IRB approval has been obtained by our collaborator, Prof. James Rowe, from the University of Cambridge.

scoring out of 100).

4.2 Experimental Set-up

We provide a detailed description of the experimental setup for evaluating our proposed methodolody. Our experiments were conducted using the Python 3.8 programming language on an Ubuntu 22.04 LTS operating system. We utilized the PyTorch 1.9 deep learning framework along with additional libraries such as Huggingface for establishing our model. The experiments were performed on a system equipped with an Intel(R) Xeon(R) Platinum 8370C CPU @ 2.80GHz and four NVIDIA GeForce RTX 4090 GPUs with 24GB of VRAM. During training, we employed the AdamW optimizer with an initial learning rate of 2e-5 and a weight decay of 0.05. The batch size was set to 4. The model was trained for a total of 100 epochs. Gradient accumulation was used with a step size of 2 to effectively increase the batch size and stabilize the training process. The coefficient weight β for balancing our objectives was set to 0.1. We applied a linear learning rate scheduler with a warmup parameter of 0.05 to gradually increase the learning rate during the initial training stages.

4.3 Evaluation Metrics

To assess the performance of our proposed method for detecting Alzheimer's Disease using the LLM-identified linguistic markers, we employ two primary evaluation metrics: Accuracy and F1 score. Accuracy is a widely used metric that measures the overall correctness of model prediction. It is calculated as the ratio of the number of correctly classified samples to the total number of samples in the dataset. In our case, Accuracy represents the percentage of correctly identified AD and NC samples. Accuracy is computed using the following formula:

$$\text{Accuracy} = \frac{\text{label}(x_t) = l \cap \text{predict}(x_t) = l}{num(\text{TestSet})} \tag{4.1}$$

where, x_t is the test data, num(TestSet) is the total counts of testset samples, $label(x_t)$ is the ground truth label of x_t and $predict(x_t)$ is the predicted label of x_t by the trained classifier for AD detection. While Accuracy provides an overall measure of the model's performance, it may not be sufficient when dealing with imbalanced datasets or when the costs of misclassification differ across classes. Therefore, we also utilize the F1 score, which is the harmonic mean of precision and recall.

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4.2}$$

The F1 score balances precision (the proportion of correct positive predictions among all positive predictions) and recall (the proportion of actual positive cases correctly identified), making it particularly valuable for medical diagnostic tasks, where both false positives and false negatives carry significant consequences. These evaluation metrics were reported for each fold in our stratified 5-fold cross-validation setup, while the final performance were determined by averaging the results across all folds.

4.4 Results & Analyses

4.4.1 Model Performance

Table 2 presents a comprehensive comparison of the performance based on different AD detection LLM configurations trained on the ADReSSo dataset. Table 2 evaluates the various combinations of LLM (GPT-3.5, GPT-4.0, and GPT4o) input modalities, and prompting strategies. In the modality column, ”T” represents the original transcript, ”B” denotes the linguistic markers summarised by LLMs, whereas ”A” represents the raw audio features. The Audio Model column shows whether Wav2vec is used for audio processing, while the Text Model column indicates whether Bert-Large is used for text processing. The Prompting Strategy column specifies whether the Zero-Shot or Few-Shot (with 10 examples) Learning has been employed. Performance is measured through both Accuracy and F1 score metrics, enabling a balanced assessment of model effectiveness. The results in Table 2 have demonstrated how different combinations of modalities and models affect AD detection performance, with multiple modalities (B+A+T) plus Few-Shot Learning generally achieving better results as compared to single-modality approaches.

Among the single-modality approaches, the audio-based Wav2vec[34] model achieves the highest performance (Accuracy = 0.92; F1 score = 0.91), followed by Bert-Large[4] using speech transcripts (Accuracy = 0.81; F1 score = 0.76). When LLMs are used alone for diagnosis based on the original transcripts under Zero-Shot Learning, the performance is notably lower, with GPT-3.5 achieving an accuracy of 0.58 (F1 score = 0.68), GPT-4.0 (Accuracy = 0.64; F1 score = 0.72), and GPT4o (Accuracy = 0.65; F1score = 0.69). These results have suggested that these LLMs may not be sufficient for accurate AD detection under the Zero-Shot Learning. The poor performance of the GPT models have raised concerns regarding the feasibility of using the linguistic markers summarised by LLMs to replace the original transcripts. However, the combination of linguistic markers in different modalities have yielded different results. When combining the linguistic markers with the transcripts (B+T), Bert-Large with GPT4o using the Few-Shot Learning (10 shots) has achieved an accuracy of 0.86 (F1 score = 0.83), showing significant

improvement over Zero-Shot Learning, which only reaches an accuracy around 0.62. The combination of audio features and linguistic markers (B+A) with GPT4o under the Few-Shot Learning demonstrates impressive performance, achieving an accuracy of 0.95 (F1 score = 0.94).

Table 2: Performance comparison of different methods on the ADReSSo dataset

LLMs	Audio Model	Text Model	Modality	Prompting Strategy	Accuracy	F1 Score
GPT3.5	-	-	T	Zero-Shot	0.58	0.68
GPT4.0	-	-	T	Zero-Shot	0.64	0.72
GPT4o	-	-	T	Zero-Shot	0.65	0.69
-	-	Bert-Large	T	-	0.81	0.76
-	Wav2vec	-	A	-	0.92	0.91
GPT3.5	-	Bert-Large	B	Few-Shot	0.69	0.75
GPT4.0	-	Bert-Large	B	Few-Shot	0.62	0.56
GPT4o	-	Bert-Large	B	Few-Shot	0.62	0.59
-	Wav2vec	Bert-Large	A+T	-	0.89	0.88
GPT4o	-	Bert-Large	B+T	Few-Shot	0.86	0.83
GPT4o	Wav2vec	-	B+A	Few-Shot	0.95	0.94
GPT3.5	Wav2vec	Bert-Large	B+A+T	Zero-Shot	0.83	0.83
GPT4.0	Wav2vec	Bert-Large	B+A+T	Zero-Shot	0.71	0.73
GPT3.5	Wav2vec	Bert-Large	B+A+T	Few-Shot	0.85	0.85
GPT4.0	Wav2vec	Bert-Large	B+A+T	Few-Shot	0.73	0.74
GPT4o	Wav2vec	Bert-Large	B+A+T	Few-Shot	0.97	0.97

Note: In the Modality column, "T" represents the original speech transcript, "B" denotes the linguistic markers summarised by the LLMs, and "A" denotes the raw audio features.

Most notably, the integration of all three modalities (B+A+T) with GPT4o under Few-Shot learning achieves the best overall performance (Accuracy = 0.97; F1 score = 0.97). This significant improvement over the single-modality approach has suggested that the combination of linguistic markers, acoustic features, and transcript provides the complementary information conducive for more accurate AD detection. These results have revealed that using the acoustic features alone can be effective for AD detection, whilst using the linguistic features alone are not as effective as the acoustic features due to the lack of contextual information. The combination of linguistic features with the original speech transcript or with the audio features, can achieve better results, than using linguistic features alone. The study highlights the potential effectiveness of using linguistic features/markers in combined modalities for AD detection, which are interpretable non-invasive markers highly desirable for clinical diagnosis, when compared to the more invasive biomarkers, such as MRI, or blood biomarkers etc.

4.5 Expert-verified Linguistic Markers

To systematically understand the linguistic difference between AD and NC individuals, a linguistic expert has been consulted in order to categorize the various linguistic properties identified by LLMs that distinguish AD from NC cohorts. The expert-categorized linguistic markers are summarized in Table

4. The AD linguistic patterns identified by LLMs are further categorized by the linguistic expert into four major types in Table 3. These linguistic features generally align with what have previously been identified in the literature.[5,22,23] First, AD cohorts have displayed problems in lexical recall, resulting in hesitation, repetition, more frequent occurrence of fillers, vague descriptions or even erroneous use of words. In contrast, NC cohorts have displayed little problems with word retrieval and are able to produce fluent speeches. Second, AD cohorts are weak in organizing their thoughts and maintain thematic coherence of their discourse. Their narrations are characterized by unexpected topic switchings as well as incoherent picture descriptions. NC cohorts, in contrast, can organize their thoughts coherently and logically, providing consistent and relevant descriptions. Third, AD cohorts have difficulties in correcting their speech utterances. NC cohorts are able to self-correct their speeches whenever necessary. Lastly, AD cohorts have difficulties in generating grammatical structures while NC cohorts have no problems in generating the right grammatical sentences.

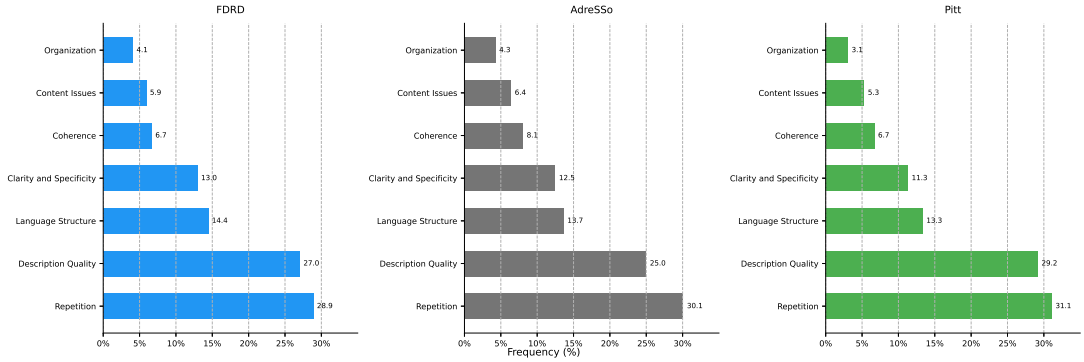


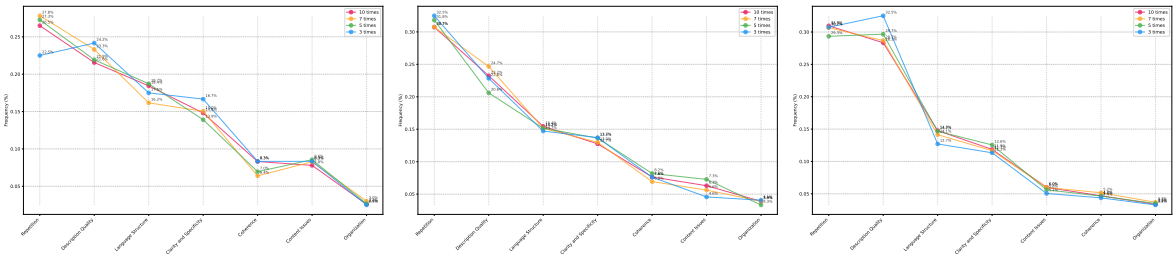
Figure 2: Linguistic marker consistency across three different datasets

4.6 LLM-summarised Linguistic Markers

4.6.1 Consistency of LLM-summarised Linguistic Markers Across Different Datasets

To evaluate the stability of LLMs in summarising linguistic patterns from speech samples, we have conducted a comprehensive analysis across multiple datasets. First, we have employed LLMs to identify distinctive language markers that differentiate AD and NC cohorts on three datasets (the Cambridge Dataset, the AdreSSo Dataset, and the Pitt Corpus).Through conducting Similarity Analysis on these LLM-summarised indicators, we have asked the LLMs to classify them into seven categories, based on the shared characteristics for each dataset. Table 4 presents these categories and their corresponding attributes. To determine the frequency of each category, we have computed the semantic similarity between individual indicators and category attributes, which have enabled us to classify the linguistic

indicators for each dataset. The frequency distributions across the three AD speech datasets have been shown in Figure 2. These distributions reveal consistent patterns in LLM-summarised linguistic markers, suggesting the robust generalizability of our proposed framework. The most frequently identified linguistic markers have shown remarkable consistency across all three datasets. Repetition has emerged as the predominant marker, with frequencies of 28.9%, 30.1%, and 31.1% in the Cambridge Dataset, the AdreSSo Dataset and the Pitt Corpus, respectively. The high frequency of repetition markers aligns with the previous clinical research,[30] which identified repetitiveness as a key predictor of AD. Similarly, Description Quality category corresponds to their finding. The category is the second most frequent across the datasets (FDRD: 27.0%, AdreSSo: 25.0%, Pitt: 29.2%), demonstrating the LLM's consistency in assessing narrative competence. Language Structure markers have shown minimal variance (0.56) across the three datasets (FDRD: 14.4%, AdreSSo: 13.7%, Pitt: 13.3%), while Clarity and Specificity demonstrated similar stability (FDRD: 13.0%, AdreSSo: 12.5%, Pitt: 11.3%). The stability of Language Structure and Clarity and Specificity markers across datasets is echoed in clinical research[30] that identifies grammatical structure and referential specificity as the reliable indicators of cognitive decline. This consistency has also suggested that LLMs can reliably detect syntactic and semantic alterations in speech patterns associated with cognitive decline. The stable hierarchy of markers has suggested that LLMs can reliably prioritize different aspects of linguistic analysis, potentially offering a standardized approach to speech assessment. The consistency across different datasets has demonstrated robust generalization capabilities, thus overcoming the common challenge of inconsistency that LLMs might exhibit when deployed to assist clinician-guided assessment.



(a) FDRD (b) AdreSSo (c) Pitt

Figure 3: Consistency of linguistic marker across multiple LLM responses

4.6.2 Consistency of LLM-based Responses Across Multiple Iterations

To further ascertain the reliability of LLM-summarised linguistic markers, we examined whether the language patterns identified by LLMs remain consistent across multiple iterations. We repeated the analysis

for different number of iterations (3, 5, 7, and 10 times) across three datasets (FDRD, AdreSSo, and Pitt) to determine the optimal number of iterations needed for reliable marker identification. Figure 3 illustrates the distributional patterns of AD linguistic markers by iterations. For each number of iterations, we have analyzed both the variety and frequency of summarised markers, comparing how the patterns stabilize or vary with increased iterations. The visualization reveals that marker distributions has become increasingly stabilized as the number of iterations increases, though in diminishing returns after reaching 5 iterations. This conclusion is supported by quantitative stability metrics showing an average reduction of variability by 47% when the number of iterations increases from 3 to 5, indicating significant stability. Further increase from 7 and 10 iterations has yielded only a marginal improvement (less than 5% reduction in variability). The marker patterns identified across 5 iterations have demonstrated consistent distributions across all three datasets, indicating robust cross-dataset reliability. The threshold of 5 iterations offers a practical guideline for similar applications in AD detection in future. This finding has significant implications for optimizing computational resource while enhancing analytical reliability for consistent clinical decision-making.

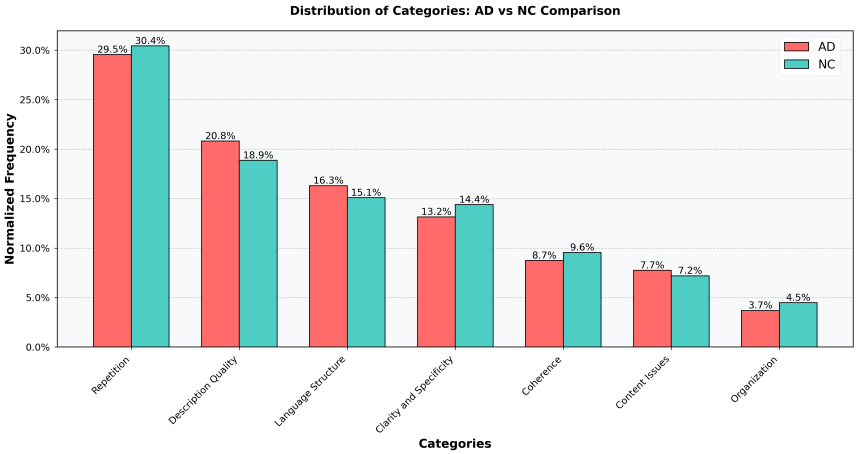


Figure 4: Comparative analysis of linguistic patterns across Alzheimer's Disease (AD) and Normal Control (NC) cohorts

4.6.3 Frequency of Linguistic Marker Categories across AD vs. NC Cohorts

To compare the distinctive language patterns across the AD and NC cohorts, we have analzyed the language characteristics exhibited by AD and NC across seven key categories. The results in Figure 4 have revealed the linguistic differences between the two groups. Speeches of AD exhibited notably higher frequencies in Description Quality (20.8% vs 18.9%) and Language Structure (16.3% vs 15.1%), as compared to those of NC. Conversely, speeches of NC have demonstrated a higher ability in Clarity and

Specificity (14.4% vs. 13.2%), Coherence (9.6% vs. 8.7%), and Organization (4.5% vs. 3.7%). Repetition has comparable occurrence frequency in both the AD and NC cohorts (AD: 29.5%, NC: 30.4%). Figure 4 has displayed minimal between-group variation (AD: 7.7%, NC: 7.2%). These findings suggest a distinctive linguistic pattern, whereas the AD cohorts have shown greater irregularity in descriptive and structural command of language, while the NC cohorts have maintained better command of organizational and coherence-related language acquisition skills. AD cohorts have experienced specific challenge in linguistic production and coherence.

Table 3: Linguistic features of Alzheimer's Disease (AD) and Normal Control (NC) cohorts.

Alzheimer's Disease (AD)	**Normal Control (NC)**
Lexical recall problems: • Hesitation, repetition, and frequent use of fillers • Vague descriptions and erroneous word usage • Difficulty in finding relevant words	**No issues with word retrieval:** • Fluent speech production • Accurate and relevant word usage
Weak thought organization and thematic coherence: • Unexpected topic switches • Irrelevant past memories • Mismatches between description and source image	**Coherent thought organization and thematic consistency:** • Logical and consistent narration • Relevant and accurate descriptions • Maintains focus on the given task
Difficulty in self-correction: • Struggle to correct errors or confusion • May persist with incorrect statements	**Ability to self-correct:** • Identifies and corrects mistakes • Maintains clarity and accuracy in speech
Weaker command of grammatical structures: • Inappropriate linguistic structures • Difficulty in formulating grammatically correct sentences	**Normal command of grammatical structures:** • Generates sentences with proper grammar • Maintains linguistic coherence and structure

5 Discussions and the Future Work

In future studies, we aim to refine our search for the distinctive linguistic markers for non-invasive speech-based AD diagnosis. This involves addressing any potential errors or biases arising from LLM-based linguistic marker identification. Linguistic markers are crucial for AD diagnostics for two reasons. First, while the current performance of LLMs in summarising the non-invasive markers may not be highly accurate, it is important to recognize the trade-off between interpretability and accuracy (Rowe, 2024[32]). Linguistic markers summarised by LLMs offer valuable insights into the language patterns and characteristics associated with AD, providing a level of interpretability that is crucial for medical diagnosis and understanding. Although the use of the interpretable linguistic markers may not directly improve accu-

Table 4: Categories of linguistic markers and their attributes

Category	Attributes	Category	Attributes
Repetition	– Repetition/repetitious – Repeated/repeating – Redundant/redundancy – Repetitive/repetitiveness – Echolalia	**Language Structure**	– Sentence/structural – Fragmented/incomplete – Mid-corrections – Vocabulary difficulties – Pronoun issues – Non-standard phrases – Confused/incorrect terms – Generic language – imprecise language – Overuse of conjunctions – Sequential structure – Unusual/filler phrases
Description Quality	– Description/descriptive – Details/detailed – Elaboration – Fluent description – coherent description – General descriptiveness – Action and detail – Simplified language – Uncertain – Conciseness – Rephrasing	**Content Issues**	– Irrelevant/inappropriate – Off-topic/tangential – Extraneous content – Daydreaming – lost in thought – Word-finding difficulties – Disorientation – Inconsistencies – Perseveration – Paraphasia – Limited/minimal content – Semantic errors – Non-sequiturs – Emotional deviations
Coherence	– Coherent/incoherent – Coherence – Cohesive/cohesion – Fluency/fluent/fluidity – Disconnected – Disorganized	**Clarity & Specificity**	– Vagueness/unclear – Ambiguous/ambiguity – Non-specific – Uncertainty/confusion – Hesitation/hesitancy – Clarification needs – Over-explanation – Simplification
Organization	– Disjointed – Transition issues – Abrupt changes – Organization problems – Incomplete thoughts – Completeness concerns		

Note. The attributes above represent the key linguistic features identified in discourse analysis. Each attribute may manifest in different versions.

racy in AD detection, their potential to enhance the interpretability and diagnostic process is invaluable. By leveraging these linguistic markers, healthcare professionals can gain a deeper understanding of the linguistic changes that occur in AD cohorts, enabling the clinicians to make more informed healthcare decisions for the AD patients and provide better patient care.

Second, from the medical and public health perspective, speech-based AD diagnostic offers a great potential as the non-invasive and low-cost solution for AD diagnosis/prognosis, especially when some of these diagnostics may invite scrutiny from a particular stakeholder group (e.g. the patient group) due to time/cost/invasiveness (e.g. MRI/CSF), even though the benefits are huge (Vanderschaeghe et al., 2019;[31] Potsteinssen et al., 2021[29]). It has been well recognized that linguistic markers are effective

markers for both prediction of AD onset and progression,[30 2]). Speech-based non-invasive AD diagnosis generally involves relatively simple collection of short speech samples via smartphone apps. They are spontaneous, non-invasive, low-cost and relatively accurate. This means that these apps can be easily deployed for screening and longitudinal speech-based health monitoring. It is in sharp contrast with complicated and expensive traditional AD diagnostics such as MRI, PET imaging, and CSF. As a result, the use of LLMs to identify AD markers is effective, and will substantially benefit diagnosis and prognosis of AD and related clinical/health-based decision making.

The key challenge in summarising accurate and reliable linguistic markers from LLMs is speech data scarcity. With a small amount of speech transcript samples, LLMs may not be able to summarize linguistic markers at high enough accuracy, to make reliable diagnostic decisions and convincing interpretations. To address this challenge, we propose using LLM-based data generation/augmentation techniques to improve the quality of AD speech dataset. Although the LLM-summarised linguistic markers may not be accurate enough for individual AD detection, we can use these markers to personalize AD speech data based on AD-specific linguistic characteristics. By controlling the linguistic patterns and features underlying the AD synthetic speech data, we can create a more diverse and representative dataset covering a broader range of AD-related languistic variations. This approach can help mitigate the limitations of the original dataset (poor in data quality and volume, especially longitudinal data scarcity) and improve the overall performance of AD detection. For instance, we can generate new AD speech samples for facilitating the in-context learning of LLMs.[25] By creating new speech examples that mimic the linguistic patterns and characteristics of AD, we can provide LLMs with larger and more diverse (e.g. longitudinal) datasets. This new AD speech samples can help LLMs learn and extract more accurate linguistic markers and improve their diagnostic performance. Moreover, we can also generate speech transcript samples for supervised multimodal model learning and training. In addition to enhancing the performance of LLMs, the new AD speech data can also be used to train the multimodal model by combining linguistic markers with other modalities. By generating a larger dataset with pre-defined characteristics, we can explore a wider sample space and potentially improve the training performance of multimodal model training.

6 Conclusion

In conclusion, this study demonstrates the potential of using LLMs to extract interpretable linguistic markers for speech-based non-invasive AD detection. By leveraging the in-context Few-Shot and Zero-

Shot learning and prompting strategies, we show that LLMs can identify high-level linguistic patterns discriminating AD from NC. Our designed prompting strategies facilitate LLMs by providing interpretation and assessment of the summarised markers, showcasing the strength, reliability, and relevance of LLMs for AD classification. Furthermore, we propose a novel framework that incorporates the linguistic markers summarised by LLMs into a small supervised-learning model. Our comprehensive analyses have shown that while acoustic features alone are highly effective for AD detection, linguistic markers summarised by LLMs in isolation lack the sufficient contexts for accurate linguistic-based classification. When these linguistic markers are integrated with either acoustic features or original speech transcripts through Few-Shot learning approaches, they can significantly enhance AD detection performance, achieving 97% accuracy (maximum) for AD classification. In future studies, we should focus on improving the accuracy of LLM-summarised markers while maintaining interpretability. Investigating speech data generation/augmentation techniques is also a promising direction. Overall, this study takes an important step in harnessing the power of LLMs for interpretable speech-based AD detection, paving the way for future research. By combining the strength of LLMs with supervised learning, we can work towards developing more accurate, interpretable, and trustworthy LLM-based AD detection systems to assist the healthcare professionals for early diagnosis and therapeutic care for the AD patients, helping LLMs fulfil the desirable and meaningful purpose of safeguarding the well-being of our society, by facilitating sound LLM-based clinical decision-making.

7 Appendix-A

Table 5: Example of using LLMs to identify linguistic markers

Few-Shot Prompting Example:

Input Prompt:

Designed prompt for summarizing linguistic markers in Section 3

Original Transcript:

Continued on next page

Table 5 – *Continued from previous page*

[There's something has to be where the water goes down over. There's probably something that, or they don't have it open or something. I don't know. When the water goes down, what do you call that? This here. Right here. This. What do you call that? What is that? What is that? What would you call that? I don't know. That's what I'm saying. I don't know what that is. I would call that the U-pipe. What? What? A U-pipe? A U-bend? No, it's just a pipe. A pipe? Oh, oh, oh, water pipe. Yeah. Oh, yeah. Okay, well then maybe the water pipe is broke, or not broke, but there must be things in there that the water will not go down. I don't know....o what's happening to the water? Huh? What's happening to the water? What's happening to the water? Yeah. The water is going down in the... What would you call this? On the floor? Floor, floor, floor. Yeah. Okay, okay. Okay. And what's happening over this side of the picture then? What's happening over here? Well, down on this side of the picture. Well this thing here is turning over. Uh huh. Yeah. Can you tell me what's going on there? No. I don't know what's going on. Well he's probably getting, what's this here? Cocoa jar. What's this cocoa? Say it again. Can you read it again for me? C-O-O-K-I-E. What does that spell? I don't know. Cookie. Cookie. Cookie. Cookie. Oh, okay. And what are the kids doing? Well, he's getting it out, and he's going to give it to the girl down here. Okay. Is there anything else going on in the picture that you want to tell me about? Going on in the picture? Well, the boy is giving him the girl, the cocoa. This probably is broke so the water will not go down in and it's coming up and going in here. And what's she doing? Well, it looks like she's what she was gonna wash What she what they eat with yeah, what do you call it? What do you call this a plate Dishes plate a plate. Oh, oh, yeah. Yeah. Yeah. Yeah what you eat in Is that is that what you call them a plate cup? Oh, this is a cup. It might be a saucer. Oh Oh maybe. It might be a saucer or a plate. Okay.]

LLM's Response:

"Given the transcript, potential language-based indications of cognitive impairment include:"

Identified Linguistic Markers:

"1. Struggling with word finding and lexical memory: The individual frequently appears unable to recall common terms, such as "cookie," "floor," "plate," and "cup."

2. Exhibiting higher levels of repetition: The individual repeatedly asks the same question or makes the same comment multiple times.

3. Lack of coherence and confusion: The person struggles to provide a systematic description of the picture and shifts focus between points."

Continued on next page

Table 5 – *Continued from previous page*

Diagnosis Decision:

"Based on the identified linguistic indicators, a diagnosis of potential Alzheimer's Disease is tentatively suggested but cannot be conclusively determined."

Explanation for Diagnosis Decision:

"The individual's language patterns indicate certain difficulties consistent with Alzheimer's Disease-like cognitive impairment, such as word-finding struggles and incoherence. However, these indicators alone are not strong enough to definitively diagnose Alzheimer's. This suggests a need for further cognitive testing and medical consultation to avoid overdiagnosing based purely on a limited linguistic evaluation."

8 Appendix-B

Table 6: Example of LLMs categorizing linguistic markers

Few-Shot Prompting Example:

Task:

[Categorize the given linguistic markers into predefined categories, calculate the frequency of each category, and prepare data for visualization.]

Context:

[You will be provided with a list of linguistic markers and a set of predefined categories. Each marker should be assigned to one category only.]

Predefined categories:

[1. Repetition 2. Description Quality 3. Coherence 4. Language Structure 5. Content Issues 6. Clarity and Specificity 7. Organization]

Input linguistic markers:

[Find it in the uploaded txt file.]

Instructions:

[1. Categorize each linguistic marker into one of the predefined categories. 2. Count the frequency of markers in each category. 3. Prepare a JSON object with the following structure: [Predefined Output Format]

LLM's Response:

Continued on next page

Table 6 – *Continued from previous page*

"1. Categorizing the linguistic markers: Repetition: [Repetition, repeated, repeating...] Description Quality: [Description, descriptive...] ... Organization: [Disjointed, Transition issues...] "
"2. Counting the frequency of markers in each category: Repetition: 9 Description Quality: 9 Coherence: 7 Language Structure: 5 Content Issues: 7 Clarity and Specificity: 8 Organization: 4 Overall Characteristic: 4 "

Acknowledgements

We gratefully acknowledge Yang Han for the valuable discussions and insights that greatly benefited our work. We would also like to thank Prof. James Rowe and Prof. David Rubinstein, the University of Cambridge, for their helpful insights and suggestions for this study.

Author Contributions

Conceptualization,V.O.K.L.,J.C.K.L; Methodology,T.Y.M.,J.C.K.L., V.O.K.L; Formal Analysis,T.Y.M, J.C.K.L., L.Y.L.C; Data Curation, T.Y.M.; Writing—Original Draft, T.Y.M., Writing—Review and Editing, J.C.K.L. and V.O.K.L; Supervision, J.C.K.L. and V.O.K.L.; Funding Acquisition, J.C.K.L. and V.O.K.L.

References

[1] **Young AL, Oxtoby NP, Garbarino S, Fox NC, Barkhof F, Schott JM, Alexander DC.** (2024) Data-driven modelling of neurodegenerative disease progression: thinking outside the black box.*Nature Reviews Neuroscience. 8*,1–20.

[2] **Yang Q, Li X, Ding X, Xu F, Ling Z.** (2022) Deep learning-based speech analysis for Alzheimer's disease detection: A literature review.*Alzheimer's Research & Therapy. 14*(1), 186.

[3] **Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I.** (2017) Attention is all you need. *JAdvances in neural information processing systems 30*(2).

[4] **Devlin J, Chang MW, Lee K, Toutanova K.** (2018) Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL*, 4171–-4186.

[5] **Fraser, K.C., Meltzer, J.A. and Rudzicz, F.** (2016) Linguistic features identify Alzheimer's disease in narrative speech, *Journal of Alzheimer's Disease 49*(2), 407–422.

[6] **Orimaye, S. O., Wong, J. S. M., Golden, K. J.** (2014) TLearning predictive linguistic features for Alzheimer's disease and related dementias using verbal utterances, *Jn Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From linguistic signal to clinical reality.* 78–87.

[7] **Kavé, G. and Goral, M.** (2016) Word retrieval in picture descriptions produced by individuals with Alzheimer's disease, *Journal of clinical and experimental neuropsychology 38*(9), 958–966.

[8] **Snell J, Swersky K, Zemel R.** (2017) Prototypical networks for few-shot learning. *Advances in neural information processing systems. 30.*

[9] **Socher R, Ganjoo M, Manning CD, Ng A.** (2013) Zero-shot learning through cross-modal transfer. *Advances in neural information processing systems 26.*

[10] **Rasmy L, Xiang Y, Xie Z, Tao C, Zhi D.** (2021) Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease predictionr. *NPJ digital medicine 20*(4), 86.

[11] **Amini, S., Hao, B., Yang, J., Karjadi, C., Kolachalama, V.B., Au, R. and Paschalidis, I.C.** (2024). Prediction of Alzheimer's Disease progression within 6 years using speech: A novel approach leveraging language models. *Alzheimer's & Dementia*, *20(8)*,5262-5270.

[12] **Kucherenko, T., Hasegawa, D., Henter, G.E., Kaneko, N. and Kjellström, H.** (2019). Analyzing input and output representations for speech-driven gesture generation. *In Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*, 97-104.

[13] **Yang K, Ji S, Zhang T, Xie Q, Kuang Z, Ananiadou S.** (2023) Towards interpretable mental health analysis with large language models. *In The 2023 Conference on Empirical Methods in Natural Language Processing*.

[14] **Yang K, Zhang T, Kuang Z, Xie Q, Ananiadou S.** (2023) Mentalllama: Interpretable mental health analysis on social media with large language models. *arXiv preprint* , 2309.13567.

[15] **Xu B, Yang A, Lin J, Wang Q, Zhou C, Zhang Y, Mao Z.** (2023) Expertprompting: Instructing large language models to be distinguished experts. *arXiv preprint* , 2305.14688.

[16] **Luo Z, Xie Q, Ananiadou S.** (2023) Chatgpt as a factual inconsistency evaluator for abstractive text summarization. *arXiv preprint*, 2303.15621.

[17] **Jeon H, Yoo D, Lee D, Son S, Kim S, Han J.** (2023) A Dual-Prompting for Interpretable Mental Health Language Models. *arXiv preprint*, 2402.14854.

[18] **Chen JM and Balamurali B.T.** (2023) Performance Assessment of ChatGPT vs Bard in Detecting Alzheimer's Dementia. *arXiv preprint*, 2402.01751.

[19] **Pakhomov SV, Smith GE, Chacon D, Feliciano Y, Graff-Radford N, Caselli R, Knopman DS.** (2023) Computerized analysis of speech and language to identify psycholinguistic correlates of frontotemporal lobar degeneration. *Cognitive and Behavioral Neurology*, *23*(3), 165–77.

[20] **Luz S, Haider F, de la Fuente S, Fromm D, MacWhinney B.** (2021) Detecting cognitive decline using speech only: The adresso challenge. *arXiv preprint*, 2104.09356.

[21] **Baevski A, Zhou Y, Mohamed A, Auli M.** (2023) wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, *33*, 12449-60.

[22] **Glosser G, Deser T.** (2023) Patterns of discourse production among neurological patients with fluent language disorders. *Brain and language*, *40*1, 67-88.

[23] **Sajjadi SA, Patterson K, Tomek M, Nestor PJ.** (2023) Abnormalities of connected speech in semantic dementia vs Alzheimer's disease. *Aphasiology*, *26*6, 847-66.

[24] **Hao H, Zhou L, Liu S, Li J, Hu S, Wang R, Wei F.** (2024) Boosting large language model for speech synthesis: An empirical study. *arXiv preprint*, 2401.00246.

[25] **Dai H, Liu Z, Liao W, Huang X, Cao Y, Wu Z, Zhao L, Xu S, Liu W, Liu N, Li S.** (2023) Auggpt: Leveraging chatgpt for text data augmentation. *arXiv preprint*, 2302.13007.

[26] **Shi M, Cheung G, Shahamiri SR.** (2023) Speech and language processing with deep learning for dementia diagnosis: A systematic review. *Psychiatry Research*, *10*, 115538.

[27] **Van Der Ende EL, Bron EE, Poos JM, Jiskoot LC, Panman JL, Papma JM, Meeter LH, Dopper EG, Wilke C, Synofzik M, Heller C.** (2022) A data-driven disease progression model of fluid biomarkers in genetic frontotemporal dementia. *Brain*, *10*, 115538.

[28] **Li VO, Lam JC, Han Y, Cheung LY, Downey J, Kaistha T, Gozes I.** (2022) Designing a protocol adopting an artificial intelligence (AI)–driven approach for early diagnosis of late-onset Alzheimer's disease. *Journal of Molecular Neuroscience*, *717*, 1329–37.

[29] **Porsteinsson AP, Isaacson RS, Knox S, Sabbagh MN, Rubino I.** (2022) Diagnosis of early Alzheimer's disease: clinical practice in 2021. *The journal of prevention of Alzheimer's disease*, *8*, 371–86.

[30] **Eyigoz, E., Mathur, S., Santamaria, M., Cecchi, G.** (2020) Linguistic markers predict onset of Alzheimer's disease. *TEClinicalMedicine*, 28.

[31] **Vanderschaeghe G, Vandenberghe R, Dierickx K.** (2019) Stakeholders' views on early diagnosis for Alzheimer's disease, clinical trial participation and amyloid PET disclosure: a focus group study. *Journal of Bioethical Inquiry*, *16*,45–59.

[32] **Rowe J.** (2024) Precision Medicine for Dementia: Thinking Large and Thinking Small. *Symposium on AI for Social Good 2024*.

[33] **Becker, J. T.; Boiler, F.; Lopez, O. L.; Saxton, J.; and McGonigle, K. L.** (1994). The natural history of Alzheimer's disease: description of study cohort and accuracy of diagnosis. *Archives of neurology*, *51(6)*,585–594.

[34] **Baevski A, Zhou Y, Mohamed A, Auli M.** (2023) wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, *33*, 12449-60.