

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

Tingyu Mo¹, Jacqueline C. K. Lam^{* 1}, Victor O. K. Li^{* 1}, and Lawrence Y. L. Cheung²

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

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

Abstract

Alzheimer's disease (AD) is a progressive and irreversible neurodegenerative disorder. Early detection of AD is crucial for timely disease intervention. This study proposes a novel LLM framework, which extracts interpretable linguistic markers from LLM models and incorporates them into supervised AD detection models, while evaluating their model performance and interpretability. Our work consists of the following novelties: First, we design in-context few-shot and zero-shot prompting strategies to facilitate LLMs in extracting high-level linguistic markers discriminative of AD and NC, providing interpretation and assessment of their strength, reliability and relevance to AD classification. Second, we incorporate linguistic markers extracted by LLMs into a smaller AI-driven model to enhance the performance of downstream supervised learning for AD classification, by assigning higher weights to the high-level linguistic markers/features extracted from LLMs. Third, we investigate whether the linguistic markers extracted by LLMs can enhance the accuracy and interpretability of the downstream supervised learning-based models for AD detection. Our findings suggest that the accuracy of the LLM-extracted linguistic markers-led supervised learning model is less desirable as compared to their counterparts that do not incorporate LLM-extracted markers, highlighting the tradeoffs between interpretability and accuracy in supervised AD classification. Although the use of these interpretable markers may not immediately lead to improved detection accuracy, they significantly improve medical diagnosis and trustworthiness. These interpretable markers allow healthcare professionals to gain a deeper understanding of the linguistic changes that occur in individuals with AD, enabling them to make more informed decisions and provide better patient care.

*Corresponding authors: V.O.K. Li (vli@eee.hku.hk), J.C.K. Lam (jcklam@eee.hku.hk).
NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

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 intervention and optimal management strategies to be implemented, potentially improving patient outcomes and quality of life. Recent advancement in natural language processing (NLP)^{22,24} and deep learning has opened up new possibilities for developing automated tools to assist in the detection and diagnosis of AD^{1, 2,25,26}. LLMs have emerged as a powerful tool in the field of NLP. LLMs are deep learning models trained on vast amounts of text data, enabling them to capture rich linguistic information and generate human-like text. These models, such as GPT (Generative Pre-trained Transformer)³ and BERT (Bidirectional Encoder Representations from Transformers),⁴ have achieved remarkable performance in various NLP tasks, including language understanding, generation, and classification. Moreover, the traditional NLP methods for AD detection faces several challenges. One major challenge is the limited availability of labeled speech data from individuals with AD. Traditional supervised learning approaches require large amounts of annotated data, which can be difficult and time-consuming to be obtained in the healthcare domain. To address this challenge, few-shot and zero-shot learning techniques have been proposed, enabling models to learn from limited labeled and unlabeled data. Specifically, LLMs, pretrained models, can be suitable for in-context few-shot or zero-shot learning. Applying LLMs for medical analysis is promising. Another challenge lies in the interpretability of the speech data and its transcripts. While these the traditional NLP methods for AD detection can achieve satisfactory performance in AD or non-AD classification, they struggle to capture the underlying complex linguistic patterns. The extracted features may not always be easily understandable or explainable to healthcare professionals. As interpretability is crucial in healthcare applications, enabling clinicians to make informed decisions and trust the model's predictions, it is critical to evaluate these new techniques that are used for extracting interpretable linguistic biomarkers. One promising approach is the use of LLMs to identify interpretable linguistic biomarkers that can serve as indicators of cognitive decline. LLMs have demonstrated remarkable performance in various NLP tasks, including language understanding, generation, and classification.

In this paper, we propose a novel approach for AD detection based on interpretable linguistic biomarkers extracted from LLMs. Our approach leverages few-shot and zero-shot learning techniques to enable LLMs to identify high-level linguistic patterns from limited transcription data. We design a set of prompting strategies to guide LLMs in extracting key linguistic features and generating explanations. Furthermore, we investigate the incorporation of LLM-extracted linguistic biomarkers into downstream supervised learning models for assessing AD/NC classification performance. While LLMs have shown impressive results in various NLP tasks, their performance in AD classification lags behind traditional supervised learning approaches given the limited data constraint. To bridge this gap, we propose a novel AD detection architecture that capitalizes on the strengths of both LLMs and supervised learning techniques. The main contributions and novelties of our work are as follows:

- We use Large Language Models (LLMs) to extract interpretable linguistic markers for AD detection using in-context few-shot and zero-shot prompting strategies.
- We 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 classification.
- We incorporate linguistic markers extracted by LLMs into a supervised-learning model to enhance the performance of downstream supervised learning for AD classification, by assigning higher weights to the high-level linguistic markers/features extracted from LLMs.

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

2 Related Work

In recent years, there has been a growing interest in using linguistic markers for AD detection. Linguistic markers refer to the distinctive features and patterns in language production that can serve as indicators of cognitive impairment. 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.¹⁷ investigated the use of verbal fluency tests as a screening tool for frontotemporal lobar degeneration and demonstrated their effectiveness in detecting cognitive impairment. Kavé and Goral⁷ explored the semantic and phonemic fluency deficits in AD and highlighted their potential as linguistic markers of the disease.

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.^{11,12} proposed interpretable mental health analysis frameworks using LLMs to identify linguistic patterns and generate explanations from text data. Xu et al.¹³ introduced ExpertPrompting, instructing LLMs to behave as domain experts, enhancing performance and reliability in mental health applications. Luo et al.¹⁴ explored ChatGPT's capability in evaluating factual inconsistencies, highlighting the potential of LLMs in assessing the reliability of generated insights. Jeon et al.¹⁵ proposed a dual-prompting approach to improve the interpretability and reliability of mental health analysis using LLMs. These studies demonstrate the growing interest in leveraging LLMs for accurate, interpretable, and trustworthy mental health analysis. 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 Large Language Models (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 markers discriminative of AD, assessing their strength and reliability, and providing interpretations concerning their relevance for AD detection. We use novel linguistic markers extracted by LLMs to enhance the performance of downstream supervised learning, by proposing a novel framework that incorporates the LLM-identified markers into a multi-modal AD detection model. Our work contributes significantly to the growing body of research that advances AD diagnostics capitalizing on LLM technologies.

3 Methodology

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

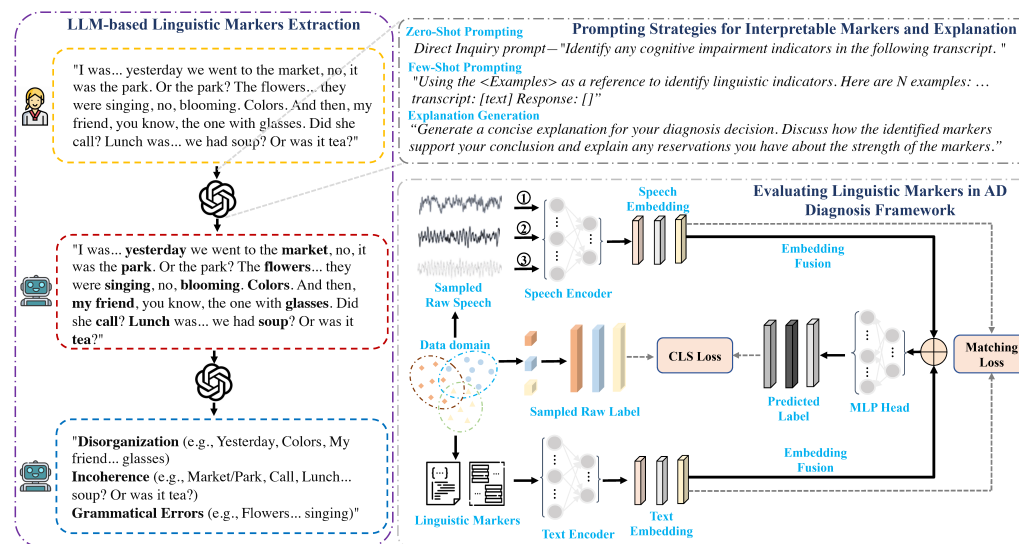


Figure 1: A framework for leveraging large language models for identifying interpretable linguistic markers and enhancing Alzheimer's Disease Diagnostics

3.1 LLM-based Linguistic Markers Extraction

In this study, we explore the capabilities and applications of several leading LLMs, including ChatGPT-3.5, ChatGPT-4, to extract linguistic patterns and markers relevant to Alzheimer's disease (AD) diagnosis. These models are representatives in large language technology, each with unique characteristics and strengths in generating and analyzing human language.

- **ChatGPT-3.5:** As one of the versions of the GPT (Generative Pre-trained Transformer) series developed by OpenAI, ChatGPT-3.5 making it one of the most powerful language models available at its time of release. It excels in generating human-like text, completing given text prompts with high coherence, and understanding subtle nuances in language. We leverage ChatGPT-3.5's capabilities to identify linguistic patterns and anomalies that may indicate cognitive impairment or AD.

- **ChatGPT-4:** An advancement over its predecessor, ChatGPT-4 is a more sophisticated model that has been 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. We evaluate ChatGPT-4's advanced language understanding to extract more refined and contextually-aware linguistic markers.

3.2 Prompting Strategies for Linguistic Markers Extraction

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 identity¹³ 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 learning⁹ 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 and blows:

"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.”

• **Few-shot Prompting:** Few-shot learning⁸ involves training the model with extremely limited labeled data to perform a task. In our work, we aim to help the LLMs to quickly adapt to identifying AD-specific linguistic markers with limited speech transcripts. We design a few-shot prompting for general LLM without being specified to specific domain. Few-shot prompting contains extremely limited speech transcripts collected from patients with AD, NC, followed by a question asking for identification of similar linguistic patterns in unseen data. For instance, a Few-shot prompting based on Chain of thought is similar to blows:

”<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]”.

In this prompt, the LLM 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 these zero-shot and few-shot prompting strategies, we aim evaluate whether LLMs can leverage the knowledge and adapt to extract linguistic markers specific to AD and NC. The extracted markers will be utilized in downstream supervised learning models for examining the performance on AD diagnosis.

3.3 LLM-based Interpretable Explanation Generation

To further analyze the identified linguistic markers and provide interpretable explanations, we assign two crucial tasks to the LLM. As shown in Line 13 in Algorithm.1, the task (b) focuses on making a diagnosis decision based on the identified linguistic indicators and their assessed strength. The prompt content is blows:

”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 individual who provided the Cookie Theft picture description as either having Alzheimer’s disease (AD) or being a normal control (NC). However, we emphasize the importance of caution in this decision-making process. If the identified linguistic indicators are not strong or conclusive enough to confidently determine the presence of AD, we guide the LLM to lean towards a diagnosis of ”Not Sure.” This approach aims to prevent overdiagnosis and ensures that the LLM considers

the reliability and significance of the linguistic markers before making a definitive classification.

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

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. This explanation should highlight the specific linguistic features that contributed to the classification decision, providing insights into the reasoning process of the LLM. Additionally, we instruct the LLM to express 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 markers and communicate any limitations or ambiguities in its decision-making process. This step emphasizes the importance of a cautious approach to avoid overdiagnosis and promotes transparency in the interpretation of the results.

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

Input:

Designed prompt for Linguistic markers Extraction; Transcripts from dataset.

Output:

Linguistic markers extracted from LLM and Exploration for extraction results;

1 Assigning an expert identity:

- 2 You are a medical professional tasked with analyzing linguistic patterns in descriptions of the Cookie Theft picture from the Boston Diagnostic Aphasia Examination.

3 Assigning task:

- 4 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.

5 In-context Few-shot Learning

- 6 <Examples> Transcripts from people with Alzheimer's disease describing the Cookie Theft picture:

7 Script: [example1]...[exampleK]

- 8 <Examples> Transcripts from people with normal control describing the Cookie Theft picture:

9 Script: [example1]...[exampleK]

²¹⁰ **10 Assigning more specific tasks:**

- 11 Using the <Examples> as a reference, please analyze the following transcript [text]. please complete the following tasks:
- 12 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.
- 13 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.
- 14 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.

15 Extra output regulation:

- 16 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
 - 17 Collect response from LLM;
-

3.4 Evaluating Linguistic Markers in AD Diagnosis Framework

In this section, we propose an evaluation framework to assess the effectiveness and interpretability of the linguistic markers extracted by the LLMs for AD diagnosis. We aim to investigate whether these markers can replace the original transcripts in training AD diagnostic models and evaluate the performance of different prompting strategies and the interpretability of the LLMs. We train AD diagnostic models using these markers and explanations as input features and compare their performance to models trained on the original transcripts. We employ various evaluation metrics, such as accuracy, precision, recall, and F1 score, to assess the performance of the models. If the models trained on the linguistic markers achieve comparable or better performance than those trained on the original transcripts, it indicates that the extracted markers capture relevant information for AD diagnosis and can potentially replace the need for the less interpretable transcripts.

Referring to Fig. 1, our framework consists of two main components: the speech embedding branch and the text embedding branch. The speech embedding branch processes the sampled raw speech using a speech encoder to generate speech embeddings. On the other hand, the text embedding branch utilizes the extracted linguistic markers and their corresponding explanations. The linguistic markers serve as an alternative representation of the original speech, capturing the relevant information for AD diagnosis. These markers are then fed into the text encoder, which converts them into a text embedding for the subsequent processing steps, including embedding fusion and classification using the MLP head.

To train the AD diagnostic model, we employ two loss functions: cross-entropy loss and 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 make accurate AD diagnosis predictions. The matching loss, on the other hand, is designed to align the speech embeddings and 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:

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

$$\operatorname{argmin}_{\Theta_e, \Theta_f} \mathcal{L}_{\text{matching}} = \frac{1}{\|D_T\|} \sum_{\|D_T\|}^{\|D_T\|} \sum_{\text{es}, \text{et}} \mathcal{L}_{\text{cosine}}(\text{es}, \text{et}; \Theta_e, \Theta_f) \quad (2)$$

$$\operatorname{argmin}_{\Theta} \mathcal{L}_{\text{total}} = \mathcal{L}_{\text{cls}} + \beta \mathcal{L}_{\text{matching}} \quad (3)$$

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. $\mathcal{L}_{\text{cosine}}$ is the cosine distance based loss. β is the coefficient weight for balancing these objects.

4 Experiments

4.1 Evaluation Dataset

In our experiments, we utilize the ADReSSo Challenge 2021 dataset¹⁸ to investigate the feasibility of using LLMs for identifying interpretable linguistic markers of AD and enhancing the performance of detection. The ADReSSo Challenge 2021 dataset has been widely used in the research community, serving as a benchmark for developing and evaluating methods for AD detection through speech analysis. The dataset focuses on spontaneous speech only and its diverse participant cohort make it a valuable resource for advancing our understanding of the linguistic markers associated with cognitive decline in AD. More specific, the speech recordings are elicited using the Cookie Theft picture description task from the Boston Diagnostic Aphasia Examination. In this task, participants are shown an image depicting a scene where a boy is stealing cookies from a jar while his mother is distracted by washing dishes. The scene also includes other elements such as a sink overflowing with water and a girl reaching for a cookie jar on a high shelf. Participants are asked to describe what they observe in the picture, providing a spontaneous speech sample that captures their language abilities and cognitive functioning. These spontaneous speech sample are divided into two main categories based on the cognitive labels: Alzheimer’s Disease (AD) and Normal Control (NC). AD category includes speech recordings and transcripts from individuals diagnosed with Alzheimer’s Disease. The speech samples in this category exhibit linguistic patterns and features associated with AD. NC category includes speech recordings and transcripts from individuals who are cognitively normal, meaning they do not show signs of cognitive impairment or dementia. The speech samples in this category serve as a control group 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.

4.2 Experimental Settings

In this section, we provide a detailed description of the experimental setup used to evaluate our proposed method. Our experiments are conducted using the Python 3.8 programming language on an Ubuntu 22.04 LTS operating system. We utilize the PyTorch 1.9 deep learning framework along with additional libraries such as Huggingface for establishing our model. The experiments are 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 employ the AdamW optimizer with an initial learning rate of $2e-5$ and a weight decay of 0.05. The batch size is set to 4, and we train the model for a total of 100 epochs. Gradient accumulation is 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 objects is set to 0.1. We apply a linear learning rate scheduler with a warmup parameter of 0.05 to gradually increase the learning rate during the initial stages of training.

4.3 Evaluation Metrics

To assess the performance of our proposed method for detecting Alzheimer's Disease using linguistic markers, we employ two primary evaluation metrics: accuracy and F1 score. Accuracy is a widely used metric that measures the overall correctness of the model's predictions. 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. The accuracy is computed using the following formula:

4.3.1 Evaluation Metric

Accuracy is used as the major metric to measure the model performance and defined as follows:

$$\text{Accuracy} = \frac{\text{label}(x_t) = l \cap \text{predict}(x_t) = l}{\text{len}(\text{TestSet})} \quad (4)$$

where, x_t is the test data, $\text{len}(\text{TestSet})$ is the total counts of testset samples, $\text{label}(x_t)$ is the ground truth label of x_t and $\text{predict}(x_t)$ is the predicted label of x_t by the trained classifier for AD diagnosis. 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 between classes. Therefore, we

also utilize the F1 score, which is the harmonic mean of precision and recall.

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (5)$$

These evaluation metrics will be reported for each fold in our stratified 5-fold cross-validation setup, and the final performance will be determined by averaging the results across all folds.

4.4 Results & Analysis

4.5 Performance Analysis

Table 1 presents the performance comparison of various methods on the ADReSSo dataset. The evaluation metrics used are accuracy and F1 score, which provide insights into the overall correctness and balance between precision and recall of the models. Among the single models, Wave2vec¹⁹ achieves the highest accuracy of 0.85, closely followed by Bert⁴ with an accuracy of 0.84. These results indicate that acoustic and linguistic features captured by Wave2vec and Bert are effective in distinguishing between AD and non-AD samples. Wave2vec also achieves higher accuracy than Bert, which verifies audio data contain more information related to AD diagnosis than transcript. When considering the F1 score, Wave2vec and Bert maintain their strong performance with scores of 0.85 and 0.83, respectively. This result means that these models are able to identify AD samples accurately while minimizing false positives and false negatives.

Table 1: Performance comparison of different methods on the ADReSSo dataset

Methods	Use GPT	Prompting Strategy	Accuracy	F1 Score
GPT3.5-Only	Yes	Zero-Shot	0.58	0.68
GPT4.0-Only	Yes	Zero-Shot	0.41	0.38
Bert	No	No	0.84	0.83
Wave2vec	No	No	0.85	0.85
Wave2vec-Bert	No	No	0.89	0.88
Bert-GPT3.5-ZS	Yes	Zero-Shot	0.69	0.75
Bert-GPT4.0-ZS	Yes	Zero-Shot	0.62	0.56
Wave2vec-Bert-GPT3.5-ZS	Yes	Zero-Shot	0.83	0.83
Wave2vec-Bert-GPT3.5-FS10	Yes	Few-Shot	0.85	0.85
Wave2vec-Bert-GPT4.0-ZS	Yes	Zero-Shot	0.71	0.73
Wave2vec-Bert-GPT4.0-FS10	Yes	Few-Shot	0.73	0.74
Wave2vec-Bert-GPT4.0-FS20	Yes	Few-Shot	0.76	0.79

However, the performance of GPT3.5-Only and GPT4.0-Only is significantly lower compared to Wave2vec and Bert. GPT3.5-Only achieves an accuracy of 0.58 and an F1 score of 0.68, while GPT4.0-Only obtains an even lower accuracy of 0.41 and an F1 score of 0.38.¹⁶ These results suggest that these large language models may not be sufficient for accurate AD detection under zero-shot setting. The poor performance

of GPT3.5-Only and GPT4.0-Only raises concerns about the feasibility of using linguistic markers extracted by LLMs to replace the original transcript. The low accuracies and F1 scores indicate that the linguistic attributes produced by these LLMs may contain less informative features or introduce false information compared to the original transcript. This could be attributed to the limitations of the LLMs in capturing the nuances and complexities of language patterns associated with AD.

The combination of Wave2vec and Bert (Wave2vec-Bert) further improves the performance, achieving an accuracy of 0.89 and an F1 score of 0.88. This suggests that leveraging the complementary strengths of different models can lead to enhanced AD detection capabilities. Wave2vec-Bert achieves the highest accuracy of 0.89, followed by Wave2vec with an accuracy of 0.85. This indicates that using speech and transcript both is more informative.

Bert-GPT3.5-ZS/FS and Bert-GPT4-ZS/FS (accuracy lower than 0.76) are inferior to the Bert (0.85). These results indicate that the current performance of LLMs in extracting these markers may not be highly accurate, it is important to recognize the tradeoff between interpretability and accuracy. Despite the linguistic markers are more interpretable as compared to the original transcript, they sacrificed some information to exchange higher interpretability. Furthermore, the combination of Wave2vec, Bert, and GPT3.5 with Ten-shot learning (Wave2vec-Bert-GPT3.5-FS10) and the combination of Wave2vec, Bert, and GPT4.0 with Twenty-shot learning (Wave2vec-Bert-GPT4.0-FS20) also do not show significant improvements over the Wave2vec-Bert combination. Wave2vec-Bert-GPT3.5-FS10 achieves an accuracy of 0.85 and an F1 score of 0.85, while Wave2vec-Bert-GPT4.0-FS20 obtains an accuracy of 0.83 and an F1 score of 0.84 and Wave2vec-Bert-GPT4.0-ZS obtains an accuracy of 0.76 and an F1 score of 0.79. These results suggest that incorporating linguistic markers extracted by GPT3.5 and GPT4.0 with few learning may not provide benefits over using only Wave2vec and Bert feature. The analysis highlights the importance of carefully evaluating the quality and informativeness of linguistic markers extracted by LLMs for each sample before using them to replace the original transcript. The poor performance of GPT3.5-Only and GPT4.0-Only indicates that the linguistic attributes produced by these LLMs may not capture the essential information required for accurate AD detection. Relying on these markers alone could potentially lead to a loss of valuable information present in the original transcript. It is crucial to consider the limitations and potential biases of LLMs and the trade-off between interpretability and performance when utilizing them for extracting linguistic markers.

4.6 Linguistic Marker Analysis

We summarised the linguistic markers extracted by LLMs in Table.2. Then, We asked the LLMs to identify the different linguistic properties found in AD vs NC subjects. The descriptors of AD language issues offered by LLMs can be categorized into four types in Table.3. The features generally match well with what is reported in the literature.^{5,20,21} First, AD subjects display problems in lexical recall, resulting in hesitation, repetition, more frequent occurrence of fillers, vague descriptions or even erroneous use of words. In contrast, NC subjects have no issues with word retrieval and produce fluent speech. Second, AD subjects are weak in organizing thoughts and maintaining thematic coherence of a discourse. Their narration is characterized by unexpected topic switch as well as incoherent picture description. NC subjects, on the other hand, can organize their thoughts coherently and logically, providing consistent and relevant descriptions. Third, AD subjects show difficulties in correcting themselves. NC subjects are able to self-correct their speech when needed. Last, AD subjects show problem commanding grammatical structure while NC subjects can generate sentences with normal grammatical structures.

Table 2: Linguistic markers extracted from AD and NC individuals

Category	Linguistic markers	Index
AD	Repetition and uncertainty	LM1
	Lack of clear descriptions and irrelevant phrases	LM3
	Naming difficulties and misuse of words	LM4
	Unnecessary excessive details and language errors	LM5
	Inappropriate linguistic structure and slight confusion	LM6
	Difficulty generating content and severe paraphasias	LM7
	Excessive use of filler words and irrelevant past memory	LM8
	Narrative inconsistencies and circumlocutory speech	LM9
	Sudden topic switches and exaggerated emotional expressions	LM10
	Disorganized thoughts and difficulty naming tasks	LM11
	Sudden topic switches and exaggerated emotional expressions	LM12
	Use of disconnected and unrelated expressions	LM13
	Nebulous phrases and disjointed narration	LM14
NC	Consistent flow of thought	LM15
	Repetition and uncertainty	LM16
	Ability to self-correct	LM17
	Coherent, logical sentences showing adequate comprehension	LM18
	Clear memory, good attention span, and strong language abilities	LM19

Table 3: Linguistic features of AD and NC individuals

AD	NC
Lexical recall problems: <ul style="list-style-type: none"> Hesitation, repetition, and frequent use of fillers Vague descriptions and erroneous word usage Difficulty in finding relevant words 	No issues with word retrieval: <ul style="list-style-type: none"> Fluent speech production Accurate and relevant word usage
Weak thought organization and thematic coherence: <ul style="list-style-type: none"> Unexpected topic switches Irrelevant past memories Mismatches between description and source image 	Coherent thought organization and thematic consistency: <ul style="list-style-type: none"> Logical and consistent narration Relevant and accurate descriptions Maintains focus on the given task
Difficulty in self-correction: <ul style="list-style-type: none"> Struggle to correct errors or confusion May persist with incorrect statements 	Ability to self-correct: <ul style="list-style-type: none"> Identifies and corrects mistakes Maintains clarity and accuracy in speech
Weaker command of grammatical structures: <ul style="list-style-type: none"> Inappropriate linguistic structures Difficulty in formulating grammatically correct sentences 	Normal command of grammatical structures: <ul style="list-style-type: none"> Generates sentences with proper grammar Maintains linguistic coherence and structure

5 Discussion and Future Work

In future research, we aim to enhance the use of linguistic markers for AD diagnosis. This will involve addressing potential errors or biases arising from the linguistic markers extracted by the LLMs. We consider linguistic markers crucial for two reasons. First, while the current performance of LLMs in extracting these markers may not be highly accurate, it is important to recognize the tradeoff between interpretability and accuracy (Rowe, 2024³⁰). The linguistic markers extracted 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 these interpretable markers may not immediately lead to improved accuracy in AD detection, their significance lies in their potential to enhance the interpretability and trustworthiness of the diagnostic process. By leveraging these linguistic markers, healthcare professionals can gain a deeper understanding of the linguistic changes that occur in individuals with AD, enabling them to make more informed decisions and provide better patient care.

Second, from the medical and public health perspective, language-based AD diagnosis has great potential as a non-invasive and low-cost solution for AD diagnosis/prognosis in healthcare systems, especially when some of the diagnostics may invite scrutiny by a particular stakeholder group (e.g. the patient group) due to time/cost/invasiveness (e.g. MRI/CSF) and the benefits of using alternative methods are huge

(Vanderschaeghe et al., 2019;²⁹ Potsteinssen et al., 2021²⁷). It has been well recognized that linguistic markers are effective markers for both diagnosis of AD and prognosis of AD onset (e.g. Eyigoz et al.,²⁸ 2020, Yang et al., 2022²). Language-based diagnosis generally involves relatively simple collection of short speech samples that can be done on smartphone apps. It is natural, non-invasive, low-cost and relatively accurate means that can be easily deployed for screening and longitudinal monitoring. It is starkly contrasted 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, it will make a very positive impact on the diagnosis and prognosis of AD at the healthcare policy level.

The key challenge in extracting accurate and reliable linguistic markers from LLMs is the limited size of our dataset. With a small amount of data, the LLMs may not be able to fully activate their potential in extracting linguistic markers with satisfactory accuracy, reliable diagnostic decisions, and convincing explanations. To address this challenge in future, we propose utilizing data augmentation techniques to expand our dataset. Although the LLM-extracted linguistic markers may not be accurate enough for each individual case, we can summarize these markers and utilize them to customize synthetic data with specific characteristics. By controlling the linguistic patterns and features in the synthetic data, we can create a more diverse and representative dataset that covers a broader range of AD-related language variations. This approach can help mitigate the limitations of the original dataset and improve the overall performance of the AD detection system. For instance, we can generate synthetic data for facilitating in-context learning for LLMs.²³ By creating synthetic examples that mimic the linguistic patterns and characteristics of AD, we can provide the LLMs with a larger and more diverse dataset. This augmented data can help the LLMs learn and extract more accurate linguistic markers and improve their diagnostic performance. Moreover, we can also generate samples for supervised learning of the multimodal model. In addition to enhancing the LLMs' performance, the augmented data can also be used to train the multimodal model that combines linguistic markers with other modalities. By generating a larger dataset with predefined characteristics, we can explore a wider sample space and potentially improve the training results of the multimodal model.

6 Conclusion

In conclusion, this study demonstrates the potential of using LLMs to extract interpretable linguistic markers for Alzheimer's disease detection. By leveraging in-context few-shot and zero-shot learning prompting strategies, we show that LLMs can identify high-level linguistic patterns discriminative of

AD and NC. The designed prompting strategies facilitate LLMs in providing interpretation and assessment of the extracted markers, considering their strength, reliability, and relevance to AD classification. Furthermore, we propose a novel approach that incorporates the linguistic markers extracted by LLMs into a small supervised-learning model. By assigning higher weights to these high-level features, we demonstrate the potential to enhance the performance of downstream supervised learning for AD classification. However, the current performance of LLMs in extracting these markers may not be highly accurate, highlighting the tradeoff between interpretability and accuracy. Future research should focus on improving the accuracy of LLM-extracted markers while maintaining their interpretability. Investigating data augmentation techniques is also promising direction. Overall, this study takes an important step towards leveraging the power of LLMs for interpretable AD detection and paves the way for further research in this field. By combining the strengths of LLMs and supervised learning, we can work towards developing more accurate, interpretable, and trustworthy AD detection systems to assist healthcare professionals in early diagnosis and intervention.

7 Appendix. A

Table 4: LLM’s Response Example

Few-shot Prompting Example:

Input Prompt:

Designed prompt for Linguistic markers Extraction in Section.3

Original Transcript:

Continued on next page

Table 4 – 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 4 – 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.”

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. *Advances 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, *In 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 prediction. *NPJ digital medicine* 20(4), 86.
- [11] **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*.
- [12] **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.
- [13] **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.
- [14] **Luo Z, Xie Q, Ananiadou S.** (2023) Chatgpt as a factual inconsistency evaluator for abstractive text summarization. *arXiv preprint*, 2303.15621.
- [15] **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.

- [16] **Chen JM and Balamurali B.T.** (2023) Performance Assessment of ChatGPT vs Bard in Detecting Alzheimer's Dementia. *arXiv preprint*, 2402.01751.
- [17] **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.
- [18] **Luz S, Haider F, de la Fuente S, Fromm D, MacWhinney B.** (2021) Detecting cognitive decline using speech only: The addresso challenge. *arXiv preprint*, 2104.09356.
- [19] **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.
- [20] **Glosser G, Deser T.** (2023) Patterns of discourse production among neurological patients with fluent language disorders. *Brain and language*, 401, 67-88.
- [21] **Sajjadi SA, Patterson K, Tomek M, Nestor PJ.** (2023) Abnormalities of connected speech in semantic dementia vs Alzheimer's disease. *Aphasiology*, 266, 847-66.
- [22] **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.
- [23] **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.
- [24] **Shi M, Cheung G, Shahamiri SR.** (2023) Speech and language processing with deep learning for dementia diagnosis: A systematic review. *Psychiatry Research*, 10, 115538.
- [25] **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.
- [26] **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.
- [27] **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.

- 483 [28] **Eyigoz, E., Mathur, S., Santamaria, M., Cecchi, G.** (2020) Linguistic markers predict onset of
 484 Alzheimer's disease. *TEClinicalMedicine*, 28.
- 485 [29] **Vanderschaeghe G, Vandenberghe R, Dierickx K.** (2019) Stakeholders' views on early diagnosis
 486 for Alzheimer's disease, clinical trial participation and amyloid PET disclosure: a focus group study.
 487 *Journal of Bioethical Inquiry*, 16,45–59.
- 488 [30] **Rowe J.** (2024) Precision Medicine for Dementia: Thinking Large and Thinking Small. *Symposium*
 489 *on AI for Social Good 2024*.