



RESEARCH ARTICLE

A large language model based data generation framework to improve mild cognitive impairment detection sensitivity

Yang Han¹ , Jacqueline C.K. Lam¹ , Victor O.K. Li¹ and Lawrence Y.L. Cheung²

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

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

Corresponding authors: Jacqueline C.K. Lam and Victor O.K. Li; Emails: jcklam@eee.hku.hk; vli@eee.hku.hk

Received: 01 April 2024; **Revised:** 14 December 2024; **Accepted:** 30 January 2025

Keywords: biases reduction; counterfactual data generation; large language models; linguistic markers-based early-stage disease detection; mild cognitive impairment

Abbreviations: AD, Alzheimer's Disease; GAI, Generative Artificial Intelligence; LLM, Large Language Model; MCI, Mild Cognitive Impairment; SHAP, SHapley Additive exPlanations; XGBoost, eXtreme Gradient Boosting

Abstract

Recent studies utilizing AI-driven speech-based Alzheimer's disease (AD) detection have achieved remarkable success in detecting AD dementia through the analysis of audio and text data. However, detecting AD at an early stage of mild cognitive impairment (MCI), remains a challenging task, due to the lack of sufficient training data and imbalanced diagnostic labels. Motivated by recent advanced developments in Generative AI (GAI) and Large Language Models (LLMs), we propose an LLM-based data generation framework, leveraging prior knowledge encoded in LLMs to generate new data samples. Our novel LLM generation framework introduces two novel data generation strategies, namely, the cross-lingual and the counterfactual data generation, facilitating out-of-distribution learning over new data samples to reduce biases in MCI label prediction due to the systematic underrepresentation of MCI subjects in the AD speech dataset. The results have demonstrated that our proposed framework significantly improves MCI Detection Sensitivity and F1-score on average by a maximum of 38% and 31%, respectively. Furthermore, key speech markers in predicting MCI before and after LLM-based data generation have been identified to enhance our understanding of how the novel data generation approach contributes to the reduction of MCI label prediction biases, shedding new light on speech-based MCI detection under low data resource constraint. Our proposed methodology offers a generalized data generation framework for improving downstream prediction tasks in cases where limited and/or imbalanced data have presented significant challenges to AI-driven health decision-making. Future study can focus on incorporating more datasets and exploiting more acoustic features for speech-based MCI detection.

Policy Significance Statement

Early-stage detection of Alzheimer's disease (AD) is critical in ensuring timely treatment and improved patient outcomes. However, detecting AD at an early stage, during Mild Cognitive Impairment (MCI), has been challenging due to the lack of training data and imbalanced diagnosis labels. This study leverages prior knowledge encoded in LLMs to generate new data and facilitate out-of-distribution learning, and reduce biases in MCI label prediction due to the systematic underrepresentation of MCI subjects in the AD speech dataset. The results have demonstrated that the proposed methodology significantly improves MCI detection sensitivity, highlighting the potential of LLM-based data generation for improving early-stage AD detection. Our novel LLM-based data generation framework can be used in other disease diagnostic domains where limited and/or imbalanced speech data present the bottlenecks to AI-driven health decision-making, making it an invaluable GAI-driven tool to improve the accuracy, and interpretability of speech-based disease diagnostics and beyond.

1. Introduction

Alzheimer's disease (AD) has become a global public health concern. Low-cost and non-invasive approaches to AD detection, such as speech tests, are promising for population screening at large scale. Throughout this article, prediction refers to classifying diagnosis labels. Previous studies have demonstrated the utility of speech tests in detecting neurodegenerative diseases (Patel et al., 2022; Henderson et al., 2023). Recent studies have highlighted the application of artificial intelligence (AI) technologies in speech-based AD detection using audio and text data (Li et al., 2021). Agbavor and Liang (2022) showed that text embeddings from GPT-3 can distinguish AD subjects from normal controls (NCs). Wang et al. (2021) and Ilias and Askounis (2022) further demonstrated that text embeddings combined with audio features can improve the accuracy of AD classification. For a more comprehensive review of recent AI-driven speech-based AD detection studies, see De la Fuente Garcia et al. (2020), Petti et al. (2020), Yang et al. (2022a). While these AI-driven speech-based AD studies have achieved outstanding performance in detecting AD dementia (near 90% accuracy on average), the accuracy of detecting its early stages, that is, mild cognitive impairment (MCI), is still significantly lower and often overlooked by previous studies (Petti et al., 2020). MCI stages are often undetected, but their detection is critical for early treatment and intervention (Shankle et al., 2005). Unlocking the potential of AI for accurate detection of AD in its early stages (i.e., MCI onset) remains underexplored.

Several datasets have been created to support speech-based AD diagnostics. These datasets often include connected speech data, demographics (e.g., age, gender, and education), and cognitive assessments (De la Fuente Garcia et al., 2020). One of the most widely used public datasets in speech-based AD research is DementiaBank (Lanzi et al., 2023). It contains audio recordings from the Pitt corpus (Becker et al., 1994), derived via a longitudinal study with connected speech samples collected from AD, MCI, and normal control subjects. In the Pitt corpus, participants were asked to perform the Cookie Theft picture description task designed to elicit spontaneous speech. However, despite the increasing availability of speech-based datasets and applications, limited attention has been paid to early-stage diagnostics due to the lack of longitudinal speech data and labels, particularly those data samples collected during the MCI stage of AD development.

Detecting MCI from one's connected speech has presented significant challenges due to the subtle changes in early cognitive decline and the lack of longitudinal AD speech data, particularly those samples collected at the MCI stage of the AD continuum. While many speech-based studies have focused on detecting AD patients, few studies have been done to distinguish language differences between MCI and normal controls (Mueller et al., 2018). Speech-based MCI detection requires high-quality datasets, which are often limited and imbalanced. With limited samples and high-dimensional features, that is, having much fewer samples than features, AI-driven models are more likely to overfit the training data (Ning et al., 2021), leading to poor generalization to unseen data, especially from different AD cohort data (Wang et al., 2022). Furthermore, class imbalance, where critical disease states/labels (such as MCI) are underrepresented in the dataset, can introduce biases in predictive models because the learned model tends to minimize the overall error rate. For example, in the Pitt corpus from the DementiaBank database (Boller and Becker, n.d.), MCI samples are significantly fewer compared to normal control samples. This imbalance can introduce biases in predictive models, skewing predictions toward the majority class (i.e., normal control) and reducing diagnostic performance (Dubey et al., 2014). Consequently, current speech-based MCI detection models face significant challenges in generalizability and robustness.

Addressing these challenges requires innovative strategies to tackle data scarcity and imbalance. Existing AD research has employed imputation techniques to address missing and incomplete data (Jung et al., 2021; Xu et al., 2022; Maheux et al., 2023; Mirabnahrzazam et al., 2023). More advanced learning techniques have been utilized to address the limited data issue. On the one hand, self-supervised learning aims to learn useful representations from data without explicit labels (such as whether AD or not), allowing AI models to learn from auxiliary tasks related to the main prediction task, such as input data reconstruction along with AD prediction (Wang et al., 2022). On the other hand, transfer learning, which aims to apply representations learned from one task to related tasks, can improve AD prediction limited by

low-resource data. For example, capitalizing on shared text embeddings from BERT, a pre-trained language model, English corpus data has facilitated speech-based AD detection using Chinese AD corpus data (Guo et al., 2020). Nevertheless, most existing studies have focused on AD detection using clinical and brain imaging data instead of speech data. How to address the limited and imbalanced speech data for AD detection in its early stages, that is, MCI detection, remains to be fully explored.

Data augmentation and generation have become the key strategies used to address data scarcity issues and enrich the diversity of training data, improving the generalizability and robustness of AI models (Shorten and Khoshgoftaar, 2019; Feng et al., 2021). Traditional data augmentation techniques have been developed to perform simple transformations on various data modalities, such as rotation and flipping (imaging) (Fan et al., 2021), noise addition (speech) (Hlédiková et al., 2022; Yang et al., 2022b), and synonym replacement (text) (Cai et al., 2023), creating some variations of existing data. However, these methods often preserve the original task labels and fail to capture subtle yet complex variations in data. Data generation techniques allow for the creation of synthetic samples beyond simple transformations. Deep generative models, such as variational autoencoders (VAEs) and generative adversarial networks (GANs), can generate realistic and novel data samples, enabling AI models to learn from small datasets (Moreno-Barea et al., 2020).

Recent advancements in AI, particularly generative AI (GAI) and large language models (LLMs), such as ChatGPT, have offered new opportunities for medical research (Moor et al., 2023; Thirunavukarasu et al., 2023). For example, general-purpose LLMs pre-trained on massive amounts of data from the internet can encode biomedical knowledge and pass medical exams (Singhal et al., 2023). Some models also allow for multimodal inputs, such as electronic health records (EHRs) and imaging (Thirunavukarasu et al., 2023). One key feature of GAI is generating new data samples that are synthetic yet realistic, making it promising to tackle the data scarcity challenge (Bansal et al., 2022). Specifically, some studies have investigated the use of LLMs for data generation in low-data scenarios, leveraging the prior knowledge of LLMs pre-trained on a large amount of data (Borisov et al., 2023; Seedat et al., 2023). Given that data generation may also lead to noisy samples, the reliability and utility of newly generated data have also been investigated to determine if they benefit the downstream prediction tasks (Seedat et al., 2023). These studies have provided new insights into building more accurate, unbiased, and reliable AI models in low-data settings with the help of GAI.

Specifically, foundation models pre-trained on massive amounts of data have enabled novel data generation techniques. On the one hand, counterfactual data generation, which leverages causal knowledge and introduces hypothetical what-if scenarios often not observed in the training data, has shown promising results in improving generalization performance in imaging classification (Chang et al., 2021) and question-answering (Sachdeva et al., 2023) tasks. From a causality perspective, such data generation techniques can improve generalization performance by reducing spurious correlations between the observed features and their labels (Ilse et al., 2021). On the other hand, cross-lingual data generation represents another powerful avenue to improve generalization performance by synthesizing speech samples in multiple languages to increase linguistic diversity, which is often underrepresented in training datasets. The multi-lingual capabilities of LLMs have been utilized for translation-based text data generation across different language populations (Cahyawijaya et al., 2024). By combining counterfactual and cross-lingual generation techniques, LLMs can create more novel and relevant data samples, advancing the performance of predictive models under limited and imbalanced data.

Despite these advancements, the application of LLM-based data generation in speech-based AD diagnostics, particularly in detecting MCI from speech patterns using limited and imbalanced data, remains to be explored. This study investigates two closely related research questions to address this limited and imbalanced data challenge, reduce potential biases in MCI prediction, and improve MCI detection sensitivity: (1) Can LLM-based data generation reduce biases in MCI prediction due to limited and imbalanced data? (2) Why LLM-based data generation may or may not work in reducing biases in MCI prediction? In this study, bias specifically refers to the systematic underrepresentation of MCI subjects in the AD speech dataset. This underrepresentation is primarily due to the scarcity of longitudinal AD studies, which can more effectively capture the progression of speech changes starting at the MCI

stage rather than the AD stage (Mueller et al., 2018). The limited inclusion of MCI samples in the AD speech dataset can skew the speech-based diagnostic model's training process, reducing detection sensitivity for MCI subjects. As the diagnostic model may struggle to identify the subtle patterns associated with MCI, its effectiveness in supporting early diagnosis and intervention of AD is significantly limited. By leveraging GAI, we aim to enhance the sensitivity and generalizability of speech-based MCI detection models, ultimately supporting earlier diagnosis and intervention for AD patients.

The main contributions of this study are three-fold. First, we develop an LLM-based data generation framework to alleviate the limited and imbalanced data problem. Two novel data generation strategies are proposed, including cross-lingual and counterfactual data generation, facilitating out-of-distribution learning, and reducing biases in MCI prediction. Second, our experimental results based on the Pitt corpus from the DementiaBank database have shown that the proposed LLM-based data generation framework can significantly improve MCI detection sensitivity by up to 38%. Third, we reveal the potential reasons why the proposed data generation methodology can reduce biases in MCI prediction by investigating the key speech markers predicting MCI, highlighting the new markers that emerged after incorporating new data generation. The proposed methodology is a general data generation framework for improving downstream prediction tasks. It can also be transferred to other datasets and areas to strengthen AI-driven decision-making using limited and imbalanced data. The rest of this article is organized as follows. Section 2 describes the study data and proposed methodology. Section 3 presents the results and discusses the implications. Section 4 concludes this study.

2. Data and methodology

This section details the proposed methodology in four steps: (1) speech data collection and pre-processing, (2) LLM-based text data generation using different strategies: (a) observational generation, (b) cross-lingual generation, and (c) counterfactual generation, (3) text-based MCI classification model training and evaluation, and (4) feature importance analysis to identify the most important speech markers predicting MCI without and with new data generation incorporated. Figure 1 gives an overview of the proposed methodology.

2.1. Data collection and pre-processing

We downloaded the Pitt dataset from the DementiaBank database (Lanzi et al., 2023). The Pitt dataset is based on a longitudinal AD study where subjects were followed up in multiple years (Becker et al., 1994). Given the focus of this study on speech-based MCI detection, we included all subjects with speech data available and labeled with normal control and MCI. According to the description of the Pitt dataset (Boller and Becker, n.d.), normal control is defined by the diagnosis category coded 8 (800 or 821), and MCI is defined by the diagnosis category coded 6 or 7 (600, 610, 611, 720, or 740).

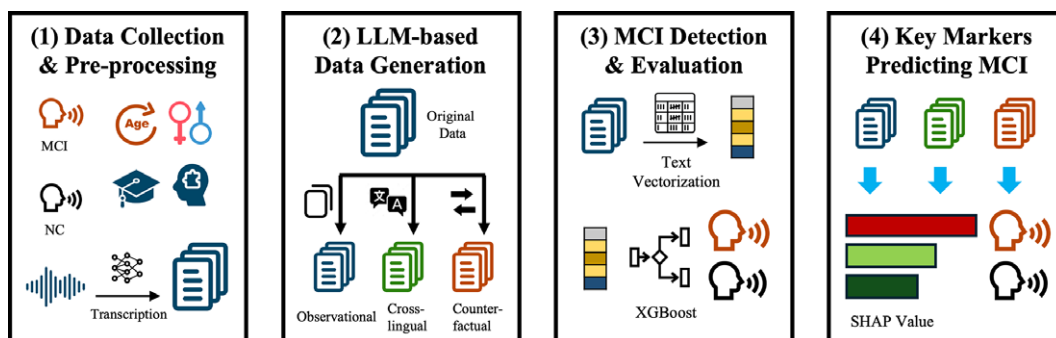


Figure 1. Overall framework.

We extracted baseline demographics, including age, gender, race, and education (in years), from the Pitt dataset. We also extracted the baseline and follow-up cognitive scores derived from the Mini-Mental State Examination (MMSE). Further, and most importantly, we obtained the baseline and follow-up audio recordings of the Cookie Theft test, the most used picture description test in clinical settings (Eyigoz et al., 2020), from the Pitt dataset. We selected subjects with normal control and MCI labels, including 89 normal controls and 18 MCI subjects. Since the longitudinal observations were sparse and limited, we considered the data points cross-sectional samples to enlarge the sample size, resulting in 129 audio samples labeled with normal control and 24 with MCI. Among these 153 samples, all demographic information was complete, and the missing ratio of MMSE was 0.7%. Table 1 summarizes the descriptive statistics of the selected audio samples. As shown in Table 1, the difference in MMSE was small, suggesting that MMSE may not be a good indicator to distinguish MCI from normal control and detect the subtle changes in cognitive function in the early stages of AD.

We transcribed the collected audio recordings in English to text data using OpenAI's Whisper model (Radford et al., 2023), a text-to-speech model pre-trained on a large amount of audio data. In addition to the audio input, we provided a simple text prompt, "Umm, let me think like, hmm... Okay, here's what I'm, like, thinking,," to the model to capture the filler words in the transcript according to the official documentation for speech-to-text prompting (OpenAI, n.d.-c). We removed non-English characters in the resulting transcriptions before further analysis. Finally, we generated a tabular dataset with seven columns, including age, gender, race, education, MMSE, transcription text, and diagnosis label.

2.2. LLM-based data generation framework

After data collection and pre-processing, an LLM-based data generation framework was developed to enlarge the pre-processed text dataset, which was of small sample size and highly imbalanced (24 MCI samples out of 153 samples), by generating more text samples that are synthetic yet realistic, leveraging on the prior knowledge encoded in LLMs pre-trained on the massive amount of data. The newly generated samples were used for model development only, not model evaluation. Specifically, three data generation strategies were investigated, including (1) observation data generation, (2) cross-lingual data generation, and (3) counterfactual data generation. At a high level, observation data generation aims to generate new MCI-like text samples based on the observed MCI samples from the existing dataset. Building on top of the observational data generation strategy, two novel generation strategies were proposed. On the one hand, cross-lingual data generation further translates the new MCI-like samples into another language (Chinese in this study), introducing cross-linguistic diversity in the training data and facilitating out-of-distribution learning. On the other hand, counterfactual data generation aims to generate new MCI-like text samples based on the observed normal control samples from the existing dataset while controlling for other variables as much as possible. The counterfactual generation process requires a deeper understanding of the disease mechanism, exploiting the causal knowledge encoded in LLMs, to answer a what-if question, such as what the speech observed from the normal control subject would be if he/she was an MCI subject while holding other information unchanged.

Table 1. Data descriptive statistics

Variable	MCI ($n = 24$)	Normal Control ($n = 129$)	Total ($n = 153$)
Age Mean (SD)	66.8 (9.2)	63.1 (8.1)	63.6 (8.3)
Gender = Female (%)	33.3%	62.0%	57.5%
Gender = Male (%)	66.7%	38.0%	42.5%
Race = White (%)	100%	99.2%	99.3%
Race = Others (%)	0.0%	0.8%	0.7%
Education Mean (SD)	15.3 (3.0)	13.9 (2.4)	14.1 (2.5)
MMSE Mean (SD)	27.9 (1.5)	29.1 (1.2)	28.9 (1.3)

Each LLM-based data generation strategy was implemented through prompting engineering using OpenAI's text generation application programming interfaces (APIs) (OpenAI, n.d.-b). The prompt consists of two parts: a system message and a user message. The system message is optional and contains instructions on how the AI-driven conversation system should behave, for example, how to rephrase the text. The user message gives a specific request for the AI system to respond, for example, a text example (OpenAI, n.d.-b). In this study, the input to the OpenAI API was formatted with a two-part system message first, followed by a user message. Specifically, the first part of the system message provides background information for the data generation task.

2.2.1. Prompting: Background Information

Use the following step-by-step instructions to respond to user inputs. The user inputs are related to the transcription of one test subject describing the Cookie Theft picture from the Boston Diagnostic Aphasia Exam. Other information of the test subject is provided, including, age, gender, race, education level (number of years), and Mini Mental State Examination (MMSE) score. Before the step-by-step instructions, some background information is listed as follows. This Cookie Theft picture description task is used to determine whether one is probable Alzheimer's disease (AD), mild cognitive impairment (MCI), or normal control (NC). The MMSE score measures one's cognitive function but needs adjustment for the education level. The step-by-step instructions are listed as follows.

The second part of the system message provides detailed step-by-step instructions for new data generation. The step-by-step instructions were designed based on the chain-of-thought prompting strategy, that is, a sequence of intermediate reasoning steps between the input and output. This prompting method has enabled significant improvement in the reasoning ability of LLMs on some tasks, such as math problems (Wei et al., 2022). Three data generation strategies were designed. The first prompting strategy, observational generation, produces new samples based on the observed dataset, mimicking the existing samples to generate similar samples. Building on top of the first one, the second prompting strategy, cross-lingual generation, takes a further step by translating the mimicked samples into another language, aiming to introduce more linguistic variations in the monolingual dataset and allow for out-of-distribution learning. The third prompting strategy, counterfactual generation, deviates from the other methods and exploits the causal reasoning ability of LLMs to generate examples opposite to the observed samples. The detailed data generation prompts are listed as follows.

2.2.2. Chain-of-thought prompting: Observational generation steps

Step 1: Explain the characteristics of this text and the reasons behind why this test subject is labeled MCI.

Step 2: Given the explanations from Step 1, rephrase the original transcription to a similar but new transcription in two lines: the first line only outputs the new transcription in no more than 150 words, with a prefix 'Text: '; the second line outputs the explanations, with a prefix 'Explanations:'.

2.2.3. Chain-of-thought prompting: Cross-lingual generation steps

Step 1: Explain the characteristics of this text and the reasons behind why this test subject is labeled MCI.

Step 2: Given the explanations from Step 1, rephrase the original transcription to a similar but new transcription in two lines: the first line only outputs the new transcription in no more than 150 words, with a prefix 'Text: '; the second line outputs the explanations, with a prefix 'Explanations:'.

Step 3: Given Step 2, only translate the text but not explanations into Chinese, with a prefix ‘Chinese:’.

2.2.4. Chain-of-thought prompting: Counterfactual generation steps

Step 1: Explain the characteristics of this text and the reasons behind why this test subject is labeled NC.

Step 2: Given the explanations from Step 1, imagine what characteristics a subject labeled with MCI would have while keeping the subject’s age, gender, race, and education information unchanged.

Step 3: Given the reasons from Step 2, write a new counterfactual transcription labeled with MCI in two lines: the first line only outputs the new transcription in no more than 150 words, with a prefix ‘Text:’; the second line outputs the explanations, with a prefix ‘Explanations:’.

Next, the user message was provided as the input for data generation. The input described an example, including transcription text, diagnosis label, age, gender, race, education, and MMSE information. Given LLM’s capability in handling missing tabular inputs (Borisov et al., 2023), missing values were not imputed during data pre-processing but represented by the string “MISSING.”

2.2.5. Prompting: Input data

The original transcription of the test subject is given as follows: <transcription text>.” The label of this transcription is: <diagnosis label>. The test subject’s age is <age>, gender is <gender>, race is <race>, education level (number of years) is <education>, and MMSE score is <MMSE score > .

GPT-3.5-Turbo and GPT-4 models were tested for data generation using the default parameters listed in OpenAI’s API documentation (OpenAI, n.d.-a). Two open-source LLMs, including Gemma-2 (with 9 billion parameters) (Mesnard et al., 2024) and Llama-3.1 (with 8 billion parameters) (Touvron et al., 2023), were tested for data generation using the recommended configurations that can improve both the quality and diversity of generated text (temperature = 1.5; min- p = 0.1) (Nguyen et al., 2024). The data generation targeted a new set of MCI samples with a similar size to the normal control samples. For observational and cross-lingual data generation based on existing MCI samples, the data generation process (i.e., one iteration of all samples targeting for generation) was repeated five times to match the number of normal control samples. In contrast, the data generation process only needed to be performed once for counterfactual generation based on existing normal control samples.

After each type of LLM-based data generation, unsupervised clustering-based outlier detection was performed, given that low-quality data samples could have negative impacts on downstream prediction tasks (Seedat et al., 2023). The local outlier factor algorithm (Cheng et al., 2019), which computes outlier scores based on the deviation of a data point with respect to its nearest K neighbors, was adopted using the default parameters (K = 20; threshold = 0.1). The identified outliers were removed before MCI predictive modeling.

2.3. Text classification model training and evaluation

The original and generated text samples were first vectorized before model training and evaluation. The term frequency-inverse document frequency (TF-IDF) method was adopted for text vectorization. Unlike advanced text embedding methods, which generate a high-dimensional vector for each text sample based on deep learning models such as transformers (Reimers and Gurevych, 2019), the traditional TF-IDF method was selected to obtain more explainable features to be investigated in feature importance analysis, allowing for a deeper understanding of how data generation can help alleviate biases in MCI prediction.

Irrelevant keywords were removed, and filler words were unified before TF-IDF vectorization. Specifically, irrelevant keywords included “okay,” “alright,” “tell,” “see,” “describe,” “picture,” “action,” and “happening.” These words were often observed at the beginning of the recording, asking the test subject to perform the picture description task, for example, “*I want you to tell me all of the action that you see, okay?*” Punctuations and filler words, including “...,” “uh,” “um” (“umm” and “ummm”), and “hm” (“hmm” and “hmmm”), were converted to a unified representation (a single keyword “PAUSE”) to capture the pauses and hesitations in speech data. Moreover, Chinese text (from cross-lingual generation) was tokenized using Chinese text segmentation software (Sun, n.d.).

Taking the TF-IDF vectors as the inputs, text classification models were constructed based on eXtreme Gradient Boosting (XGBoost), a tree-boosting algorithm with 100 estimators using an objective function for binary logistic regression. MCI classification models were trained across various scenarios: (1) the original data, (2) the original data plus the three proposed data generation strategies, and (3) the original data plus different combinations of data generation strategies. For the XGBoost model trained on the original data, the parameter to control the balance of positive and negative weights for imbalanced labels was set to the ratio of the original normal control and MCI samples. This parameter was set to the default value of one for other scenarios with newly generated MCI samples to overcome the limited MCI samples in the dataset.

We performed five-fold cross-validation to evaluate the performance of MCI classification models using text data (see [Supplementary Figure S1](#) for an overview). Each fold was a 20% subsample of the original whole dataset. Four folds were selected as the training set and the remaining one as the testing set. All newly generated samples for model training were obtained from LLM-based data generation using the original samples from the training set (the four folds) but not from the testing set (the remaining fold). This cross-validation procedure was repeated five times to ensure that all folds were used for testing. As a result, each MCI sample was tested precisely once, thus providing a better understanding of MCI classification performance, even though the number of MCI samples was limited. The same cross-validation folds were applied to each MCI classification model under different data generation strategies.

Two performance evaluation metrics, including sensitivity and F1-score, were adopted in this study. Sensitivity (also known as recall or true positive rate) measures the true positive rate. F1-score provides a more balanced view of predictive accuracy given imbalanced labels. However, the traditional accuracy metric, which measures the percentage of correctly labeled samples, was not used in this study because it can be misleading when dealing with imbalanced labels. For example, given the imbalanced distribution of positive and negative samples in this study (24 MCI and 129 normal control samples), a naive classifier that always predicts normal control can still give an 84% accuracy but fails to predict MCI every time, making it not useful for AD screening at all.

The two metrics used in this study are further elaborated in [Equations \(1\) and \(2\)](#). Average sensitivity and F1-score across the five-fold cross-validation were reported in this study. The higher the sensitivity and F1-score values, the better the predictive performance.

$$\text{Sensitivity} = TP / (TP + FN) \quad (1)$$

$$\text{F1-score} = \frac{TP}{TP + \frac{1}{2}(FP + FN)} \quad (2)$$

where TP , FN , and FP are obtained from the confusion matrix calculated based on the predicted and ground truth labels, representing the number of true positives (MCI labels that are correctly predicted), false negatives (MCI labels that are incorrectly labeled), and false positives (normal control labels that are incorrectly labeled), respectively.

2.4. SHAP analysis for identifying the most important speech markers

After model evaluation, feature importance analysis was performed to investigate the key speech features predicting MCI compared to normal control, allowing for a better understanding of the new insights brought by LLM-based text data generation and why the biases in MCI prediction can or cannot be reduced via new data generation. The feature importance score was calculated using SHapley Additive exPlanations (SHAP) analysis (Lundberg and Lee, 2017). In general, SHAP analysis uses a cooperative game theoretic approach to allocate credits for a model's output among its input features. Here, game theory is connected to AI by matching input features with players in a game and aligning the model function with the game's rules. A feature participates in a model when information is available about its value. The importance of each feature is quantified by its contribution to the model prediction across all possible orderings of feature coalitions (Lundberg and Lee, 2017).

Three models were selected for SHAP analysis, including (1) the baseline model trained on the original data, (2) the best model trained based on one of the three data generation strategies, and (3) the best model trained based on the best combination of data generation strategies. For XGBoost models, tree-based SHAP analysis was performed (Lundberg et al., 2020). The SHAP value of each feature was calculated, utilizing both training and testing sets. Features were ordered by their SHAP values, highlighting their high impacts on individual samples using the maximum absolute SHAP value, that is, features having significant positive or negative effects on MCI onset.

3. Results and discussions

This section first presents the evaluation result of different data generation strategies and the most important speech markers distinguishing MCI compared to normal control, followed by discussions on the strengths and limitations of this study with future study directions.

3.1. Evaluation of data generation strategies

Table 2 shows the data generation statistics using different strategies and LLM models. The expected number of new MCI samples obtained from observational and cross-lingual generation is 120, given that these two strategies were repeated five times. The expected number of new MCI samples obtained from counterfactual generation is 129 (i.e., the number of normal control samples). However, a few samples were dropped, mainly due to the following reasons: (1) some generated text samples failed to use the specific format provided in the prompt, and no new transcription data could be extracted, and (2) some were automatically blocked due to the potentially sensitive content according to the content filtering of Microsoft Azure OpenAI service.

For each data generation strategy across different LLM models (including Gemma-2, Llama-3.1, GPT-3.5, and GPT-4), we calculated two mean text vectors based on the original and newly generated samples using the TF-IDF method. The average distance was calculated using the Euclidean distance between the two mean text vectors. Results show that the average distance varied across different data generation strategies. Observational data generation, which aims to mimic patterns observed in the original samples without introducing substantial variations, has yielded the smallest distances on average, especially with GPT-4, indicating that the generated samples are more consistent with the original samples. One exception is Gemma-2, which shows a relatively higher distance compared to other models, suggesting that it may struggle to replicate the original data distribution closely. Cross-lingual data generation has yielded the largest distances likely due to the variations introduced by cross-lingual transformations. This finding is consistent across different models, suggesting that such variations are more likely to be driven by the data generation strategy rather than the model's generation capabilities. In terms of counterfactual data generation, GPT-4 and Llama-3.1 have generated samples with higher distances from the original ones, suggesting that these models are more capable of generating complex and diverse samples in hypothetical scenarios.

Table 2. Data generation statistics across different strategies and models^a

Data generation strategy and model	Original normal control sample size	Original MCI sample size	New MCI sample size	New MCI sample size (after outlier removal)	Euclidean distance between the original and new samples
OBG (Gemma-2)	129	24	120	110	0.54
CLG (Gemma-2)	129	24	120	110	0.67
CFG (Gemma-2)	129	24	129	116	0.45
OBG (Llama-3.1)	129	24	120	108	0.30
CLG (Llama-3.1)	129	24	120	108	0.55
CFG (Llama-3.1)	129	24	129	116	0.38
OBG (GPT-3.5)	129	24	117	105	0.35
CLG (GPT-3.5)	129	24	119	107	0.64
CFG (GPT-3.5)	129	24	122	109	0.30
OBG (GPT-4)	129	24	120	108	0.28
CLG (GPT-4)	129	24	118	106	0.67
CFG (GPT-4)	129	24	123	110	0.60

^aData generation strategy abbreviation: OBG, observational generation; CLG, cross-lingual generation; CFG, counterfactual generation. The Euclidean distance measures the average distance between the original and newly generated samples in the high-dimensional TF-IDF vector space, where a lower value indicates a higher similarity between the original and new samples.

Table 3. Evaluation of different data generation strategies in improving MCI detection sensitivity and F1-score based on five-fold cross-validation^a

Training data	Test sensitivity	Test F1-score
Original (Baseline)	4%	4%
Original + OBG (Gemma-2)	12%	15%
Original + OBG (Llama-3.1)	20%	21%
Original + OBG (GPT-3.5)	20%	21%
Original + OBG (GPT-4)	24%	24%
Original + CLG (Gemma-2)	4%	6%
Original + CLG (Llama-3.1)	20%	26%
Original + CLG (GPT-3.5)	8%	11%
Original + CLG (GPT-4)	8%	11%
Original + CFG (Gemma-2)	8%	11%
Original + CFG (Llama-3.1)	37%	30%
Original + CFG (GPT-3.5)	21%	15%
Original + CFG (GPT-4)	30%	30%

^aData generation strategy abbreviation: OBG, observational generation; CLG, cross-lingual generation; CFG, counterfactual generation. Higher sensitivity and F1-score values indicate better predictive performance. Bold value indicates the best performance.

Five-fold cross-validation was performed based on the original and newly generated data. The generated data were used only for model training. Table 3 lists the performance of MCI prediction using different data generation strategies, demonstrating their effects on facilitating model training and improving MCI prediction performance in terms of sensitivity and F1-score. All data generation methods have achieved higher predictive performance than the baseline model using the original data only. GPT-4 consistently delivered robust performance across all strategies, particularly excelling in observational generation while maintaining a high F1-score in counterfactual generation. Among open-source LLMs, Llama-3.1 demonstrated exceptional performance in counterfactual generation, while Gemma-2 performed less well than other models across all data generation strategies. Specifically, among LLMs employing observational generation, GPT-4 achieved the best sensitivity (24%) and F1-score (24%), outperforming GPT-3.5 and open-source models, including Gemma-2 and Llama-3.1. In terms of cross-lingual generation, performance varied across LLMs. Llama-3.1 outperformed all other models, achieving a sensitivity of 20% and the highest F1-score of 26%. Gemma-2 showed minimal improvement, while GPT models were tied at modest values of 8% sensitivity and 11% F1-score. The counterfactual generation strategy demonstrated the greatest potential for improving sensitivity. Llama-3.1 achieved the highest sensitivity (37%) and an F1-score of 30%, matching GPT-4's F1-score, although GPT-4 trailed slightly in sensitivity (30%). GPT-3.5 and Gemma-2 performed less competitively, with GPT-3.5 reaching a sensitivity of 21% and an F1-score of 15%.

Moreover, as shown in Table 3, counterfactual generation has demonstrated superior performance compared to other data generation methods, especially when using Llama-3.1 and GPT-4 models (which have generated counterfactual samples with higher distances from the original ones, as shown in Table 2). This advantage can be attributed to its reliance on causal reasoning, which enables the introduction of hypothetical “what-if” scenarios and reduces spurious correlations between observed features and their labels (Ilse et al., 2021). This process has been shown to effectively address dataset biases, enhancing out-of-distribution generalization performance in computer vision tasks such as imaging classification (Chang et al., 2021) and language processing tasks such as question-answering (Sachdeva et al., 2023). In contrast, the cross-lingual generation has only achieved the lowest improvement on the baseline. One possible reason is that the simple TF-IDF method may only partially capture the cross-lingual text representations. Although the TF-IDF vector is a more interpretable representation, it may fail to capture the contexts and semantic relationships compared to more advanced methods, such as deep learning-

Table 4. Evaluation of different data generation combinations in improving MCI detection sensitivity and F1-score based on five-fold cross-validation^a

Training data	Test sensitivity	Test F1-score
Original + OBG + CLG (Gemma-2)	8%	12%
Original + OBG + CLG (Llama-3.1)	20%	17%
Original + OBG + CLG (GPT-3.5)	24%	23%
Original + OBG + CLG (GPT-4)	28%	27%
Original + OBG + CFG (Gemma-2)	8%	9%
Original + OBG + CFG (Llama-3.1)	20%	14%
Original + OBG + CFG (GPT-3.5)	37%	26%
Original + OBG + CFG (GPT-4)	37%	31%
Original + CLG + CFG (Gemma-2)	16%	20%
Original + CLG + CFG (Llama-3.1)	33%	25%
Original + CLG + CFG (GPT-3.5)	29%	23%
Original + CLG + CFG (GPT-4)	26%	30%
Original + OBG + CLG + CFG (Gemma-2)	12%	14%
Original + OBG + CLG + CFG (Llama-3.1)	33%	25%
Original + OBG + CLG + CFG (GPT-3.5)	37%	27%
Original + OBG + CLG + CFG (GPT-4)	42%	35%

^aData generation strategy abbreviation: OBG, observational generation; CLG, cross-lingual generation; CFG, counterfactual generation. Higher sensitivity and F1-score values indicate better predictive performance. Bold value indicates the best performance.

based multilingual text representation learning (Cahyawijaya et al., 2024). Nonetheless, as shown in Table 2, it can still provide new samples that are different from the original monolingual samples.

Furthermore, Table 4 evaluates predictive performance using different combinations of data generation strategies. These combinations generally led to improvements over single-strategy approaches, with notable variations among models. The best performance (sensitivity and F1-score) out of all settings has been achieved using all three data generation methods based on GPT-4. In contrast, for the Llama model, combining multiple strategies did not surpass the performance achieved with counterfactual generation alone. Meanwhile, the Gemma model achieved its best performance when combining cross-lingual and counterfactual generation. These findings highlight the additive effect of data generation strategies on improving predictive performance, particularly with GPT models. Compared to Gemma and Llama models, GPT models are better equipped to capture diverse aspects of speech-based MCI detection, introducing beneficial variations into the training data that enhance model generalization and robustness.

3.2. Key speech markers distinguishing MCI from normal control

In addition to performance evaluation, SHAP analysis was performed to understand which speech markers are most important in MCI prediction before and after data generation. This feature importance analysis can facilitate understanding why LLM-based data generation can improve predictive performance and reduce biases, bringing new insights into the key speech markers predicting MCI. Figure 2 shows the top 10 speech markers predicting MCI compared to normal control based on the baseline model trained on the original data. According to Figure 2, the top marker is PAUSE. The corresponding red circles are mainly distributed across the right part of the x-axis with larger positive SHAP values. This means that filler words with higher TF-IDF values significantly contribute to MCI label prediction, suggesting that pauses are more frequently observed among the MCI transcription samples. Similarly, “reaching,” “onto,” and “dish” with higher TF-IDF values are significant contributors to MCI label prediction, making them the distinguishing markers of MCI compared to normal control.

Moreover, Figures 3 and 4 illustrate the top 10 speech markers predicting MCI compared to normal control based on (1) the best model using the original data and the counterfactual data generation using



Figure 2. The top 10 speech markers predicting MCI compared to normal control based on the baseline model trained on the original data. Each circle in the plot represents one sample. The color of the circle indicates the speech marker's TF-IDF value (see the color bar on the right). The higher the TF-IDF value, the darker the red color. The lower the TF-IDF value, the lighter the blue color. The x-axis represents the SHAP value (i.e., the feature importance score). A higher positive value indicates a higher contribution to the prediction of the positive label (i.e., MCI). A lower negative value indicates a higher contribution to the prediction of the negative label (i.e., normal control).

GPT-4 (with an F1-score of 30%; see Table 3) and (2) the best model using the original data and all data generation (i.e., observational, cross-lingual, and counterfactual generation) using GPT-4 (with an F1-score of 35%; see Table 4). Supplementary Table S1 further shows the top 50 speech markers predicting MCI compared to normal control. After incorporating the generated data, PAUSE remains the top marker in different data generation scenarios, suggesting that interruptions in speech or delays between words could be a significant indicator of MCI. MCI subjects may show more frequent pauses, potentially reflecting cognitive processing difficulties in the early stage of AD (Pistono et al., 2019). However, as shown in Figure 3, new markers, such as “something” and “might,” appear more prominently after incorporating counterfactual data generation. These words, which were insignificant in the original dataset, could reflect language deficiencies common in MCI subjects, such as uncertainty or vagueness in speech (Gosztolya et al., 2019). The increase in the frequency of these markers in the newly generated data implies that counterfactual generation, with the help of prior knowledge encoded in LLMs when generating MCI-like samples, may help highlight subtle cognitive issues that were less significant initially. Further, the top distinctive speech markers separating the MCI samples from normal control samples have been highlighted in Figure 4, suggesting that normal control samples are more likely to use verbs, such as “running,” “falling,” and adjectives, such as “little,” compared to MCI samples, which is also evident in previous research where impaired verb fluency is considered a sign of MCI (Östberg et al., 2005).

3.3. Discussions

The strengths of this study are three-fold. First, unlike prior speech-based methods focused on AD dementia, this study targets early detection at the MCI stage, aiming to reduce prediction biases caused by

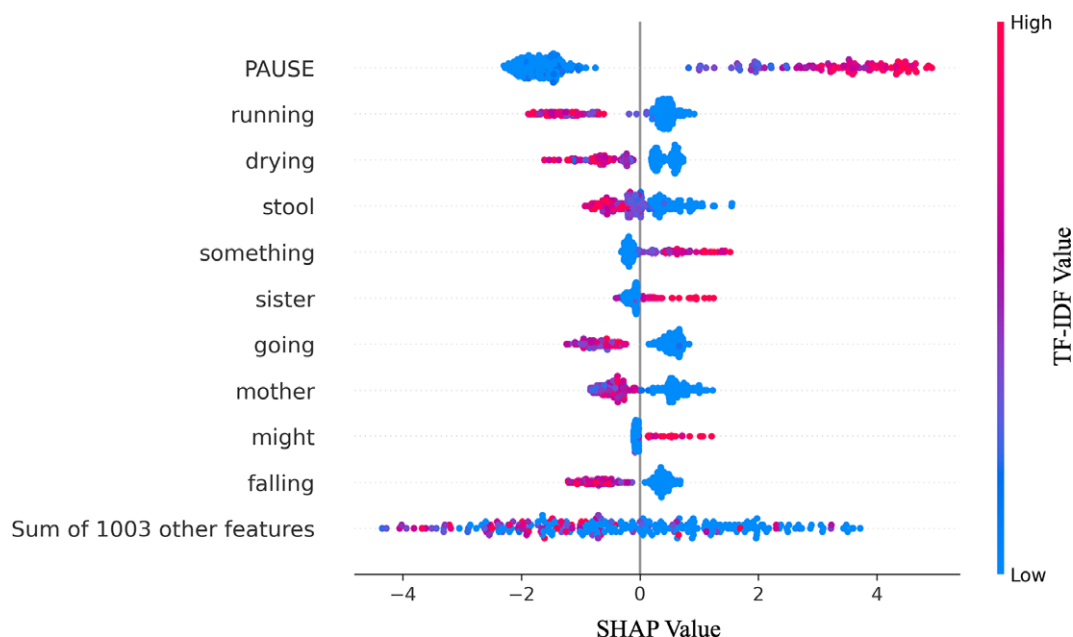


Figure 3. The top 10 speech markers predicting MCI compared to normal control based on the best model using the original data and the counterfactual data generation. Each circle in the plot represents one sample. The color of the circle indicates the speech marker's TF-IDF value (see the color bar on the right). The higher the TF-IDF value, the darker the red color. The lower the TF-IDF value, the lighter the blue color. The x-axis represents the SHAP value (i.e., the feature importance score). A higher positive value indicates a higher contribution to the prediction of the positive label (i.e., MCI). A lower negative value indicates a higher contribution to the prediction of the negative label (i.e., normal control).

limited and imbalanced data and improve detection sensitivity. Second, it introduces a novel LLM-based data generation framework, utilizing cross-lingual and counterfactual strategies to enhance out-of-distribution learning and improve prediction performance. Experimental results based on the Pitt corpus have shown that the proposed LLM-based data generation framework can significantly improve MCI detection sensitivity by up to 38% and F1-score by up to 31%. Third, SHAP analysis provides explainability, revealing key speech markers before and after data generation and highlighting how LLM-based data generation improves MCI prediction.

The proposed new data generation strategies, including cross-lingual and counterfactual generation, can reduce biases in MCI prediction due to the limited and imbalanced data because these two strategies can generate out-of-distribution samples either (1) with the same labels but with more linguistic variations from cross-lingual generation or (2) with opposite labels through counterfactual generation, which requires causal reasoning about the data generation process. These samples are not in the original dataset, making them helpful in model training under low-data settings. This study has particularly attempted to inject causality in LLMs using chain-of-thought prompting. The counterfactual generation has achieved the best improvement compared to other generation strategies (see Table 3), suggesting certain levels of causal reasoning capability of LLMs, which can be further verified in different scenarios.

This study, however, has several limitations that can be addressed in future work. First, the dataset for model training and evaluation is small. With the help of LLM-based data generation, the current study has demonstrated significant relative improvement (more than 30% for MCI detection sensitivity and F1-score) over the baseline, but the absolute predictive performance remains low. The trained model may remain overfit to the limited samples from the Pitt study and fail to generalize to entirely new samples, especially from a different cohort. Importantly, LLM-based data generation should not be used to

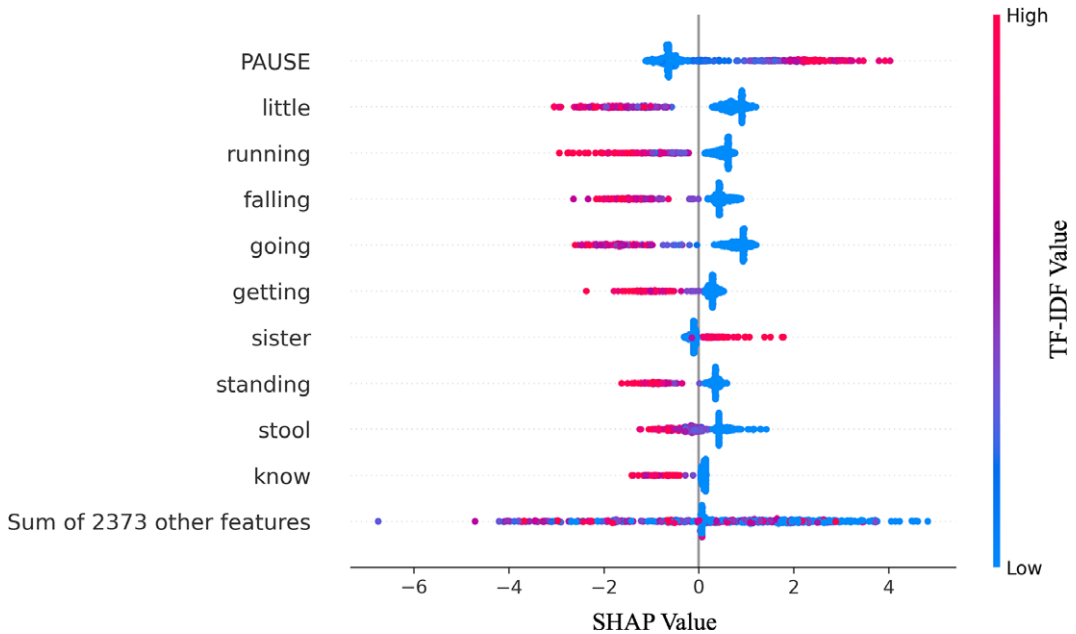


Figure 4. The top 10 speech markers predicting MCI compared to normal control based on the best model using the original data and all data generation. Each circle in the plot represents one sample. The color of the circle indicates the speech marker's TF-IDF value (see the color bar on the right). The higher the TF-IDF value, the darker the red color. The lower the TF-IDF value, the lighter the blue color. The x-axis represents the SHAP value (i.e., the feature importance score). A higher positive value indicates a higher contribution to the prediction of the positive label (i.e., MCI). A lower negative value indicates a higher contribution to the prediction of the negative label (i.e., normal control).

replace collecting more real samples but to assist speech-based disease diagnosis in low-resource settings. Future study can incorporate more related speech datasets to improve MCI detection performance, for example, from the TAUkADIAL Challenge with MCI and normal control speech samples in English and Chinese (Luz et al., 2024). Future study can also investigate the progression modeling of AD, such as MCI-to-AD conversion, using longitudinal speech data (Xue et al., 2021). Second, some audio samples were noisy, with sounds from the examiner (though often very short and brief). The quality of audio samples may affect the performance and accuracy of speech-to-text transcription, even though this study has taken several post-hoc measures for text quality control, such as removing noisy characters that were non-English and filtering out words related to the examiner's speech but unrelated to the picture description task itself. Future study can investigate more audio quality control techniques, for example, noise reduction and speaker diarization (Fujita et al., 2019). Third, this study utilized the text data obtained from audio transcripts without modeling audio data directly. Given that the experimental results have shown the importance of acoustic-related features, such as pauses, in predicting MCI, integrating both text and audio data using a multimodal approach can potentially improve speech-based MCI detection. Finally, while LLMs are trained on massive amounts of data, they may inadvertently encode societal biases embedded within their training datasets (Schramowski et al., 2022). These biases, such as those related to gender or ethnicity, can influence the generation of synthetic data and compromise the fairness of downstream tasks (Li et al., 2023). Addressing these concerns is particularly crucial in the context of real-world clinical decision-making, where biased training data can lead to inequitable outcomes (Cross et al., 2024). Future study can investigate advanced LLM debiasing techniques to improve the fairness of LLM-based data generation, ultimately enhancing the reliability and equity of medical diagnostics.

4. Conclusion

AI-driven speech-based AD detection studies have shown remarkable performance in detecting AD dementia using audio and text data. However, the detection of AD in its early stages, that is, MCI onset, remains a challenging task due to the need for sufficient training data and imbalanced diagnosis labels. Recent advancements in GAI and LLMs have provided new insights into building more accurate, unbiased, and reliable AI models in low-data settings with the help of data generation. This study introduces an LLM-based data generation framework to address the limited and imbalanced data problem, proposing two novel data generation strategies to improve MCI prediction. Experimental results based on the Pitt corpus from the DementiaBank database have demonstrated that the proposed framework can significantly enhance MCI detection sensitivity and F1-score by up to 38% and 31%, respectively. Moreover, new speech markers that emerged from data generation have been identified. These findings can help better understand why new data generation can reduce biases in MCI prediction and shed new light on speech-based MCI detection in low-data settings. Future study will incorporate more datasets and exploit more acoustic features for speech-based MCI detection. The proposed methodology is a general data generation framework that can be used for improving downstream prediction tasks in other datasets and areas where limited and imbalanced data has presented significant challenges to AI-driven decision-making.

Supplementary material. The supplementary material for this article can be found at <http://doi.org/10.1017/dap.2025.8>.

Acknowledgements. The authors are grateful for the Pitt corpus provided by the DementiaBank database. The authors would also like to acknowledge the following grant support for the Pitt corpus: NIA AG03705 and AG05133. The authors are grateful for the administrative support provided by Information Technology Services, the University of Hong Kong, for allowing them to use the Microsoft Azure OpenAI Service. The authors would also like to thank Tingyu Mo, Department of Electrical and Electronic Engineering, the University of Hong Kong, for his helpful discussions on this study. The authors 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 contribution. Conceptualization: JCK Lam; VOK Li. Methodology: Y Han; JCK Lam; VOK Li; LYL Cheung. Data curation: Y Han. Formal analysis: Y Han. Visualization: Y Han. Writing original draft: Y Han; JCK Lam; VOK Li; LYL Cheung. Funding acquisition: JCK Lam; VOK Li. All authors approved the final submitted draft.

Data availability statement. The data associated with this study can be obtained from the DementiaBank database upon request (<https://dementia.talkbank.org/>). The code associated with this study is available at <https://github.com/yanhangit/LLM-MCI-detection>.

Funding statement. This study was supported in part by the United States National Academy of Medicine Healthy Longevity Catalyst Award (Grant No. HLCA/E-705/24), administered by the Research Grants Council of Hong Kong, awarded to V.O.K.L. and J.C.K.L., and by The Hong Kong University Seed Funding for Collaborative Research 2023 (Grant No. 109000447), awarded to V.O.K.L. and J.C.K.L. The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interest. The authors declare none.

References

- Agbavor F and Liang H (2022) Predicting dementia from spontaneous speech using large language models. *PLOS Digital Health* 1(12), e0000168.
- Bansal MA, Sharma DR and Kathuria DM (2022) A systematic review on data scarcity problem in deep learning: Solution and applications. *ACM Computing Surveys (CSUR)* 54(10s), 1–29.
- Becker JT, Boiler F, Lopez OL, Saxton J and McGonigle KL (1994) The natural history of Alzheimer's disease: Description of study cohort and accuracy of diagnosis. *Archives of Neurology* 51(6), 585–594.
- Boller F and Becker J (n.d.) *Dementia Bank English Pitt Corpus*. Available at <https://dementia.talkbank.org/access/English/Pitt.html> (accessed 31 March 2024).
- Borisov V, Seßler K, Leemann T, Pawelczyk M and Kasneci G (2023) Language models are realistic tabular data generators. In *The Eleventh International Conference on Learning Representations*. Kigali, Rwanda. ICLR.
- Cahyawijaya S, Chen D, Bang Y, Khalatbari L, Wilie B, Ji Z, Ishii E and Fung P (2024) High-dimension human value representation in large language models. Preprint, [arXiv:2404.07900](https://arxiv.org/abs/2404.07900).
- Cai H, Huang X, Liu Z, Liao W, Dai H, Wu Z, Zhu D, Ren H, Li Q and Liu T (2023) Multimodal approaches for alzheimer's detection using patients' speech and transcript. In *International Conference on Brain Informatics*. Hoboken & New Jersey, USA. Springer.

- Chang C-H, Adam GA and Goldenberg A** (2021) Towards robust classification model by counterfactual and invariant data generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Virtual Event. Institute of Electrical and Electronics Engineers (IEEE).
- Cheng Z, Zou C and Dong J** (2019) Outlier detection using isolation forest and local outlier factor. In *Proceedings of the Conference on Research in Adaptive and Convergent Systems*. Chongqing, China. Association for Computing Machinery (ACM).
- Cross JL, Choma MA and Onofrey JA** (2024) Bias in medical AI: Implications for clinical decision-making. *PLOS Digital Health* 3(11), e0000651.
- De la Fuente Garcia S, Ritchie CW and Luz S** (2020) Artificial intelligence, speech, and language processing approaches to monitoring Alzheimer's disease: A systematic review. *Journal of Alzheimer's Disease* 78(4), 1547–1574.
- Dubey R, Zhou J, Wang Y, Thompson PM, Ye J and Alzheimer's Disease Neuroimaging Initiative** (2014) Analysis of sampling techniques for imbalanced data: A n= 648 ADNI study. *NeuroImage* 87, 220–241.
- Eyigoz E, Mathur S, Santamaria M, Cecchi G and Naylor M** (2020) Linguistic markers predict onset of Alzheimer's disease. *EClinicalMedicine* 28, 100583.
- Fan Y, Lam JCK and Li VOK** (2021) Demographic effects on facial emotion expression: An interdisciplinary investigation of the facial action units of happiness. *Scientific Reports* 11(1), 5214.
- Feng SY, Gangal V, Wei J, Chandar S, Vosoughi S, Mitamura T and Hovy E** (2021) A survey of data augmentation approaches for NLP. Preprint, [arXiv:2105.03075](https://arxiv.org/abs/2105.03075).
- Fujita Y, Kanda N, Horiguchi S, Xue Y, Nagamatsu K and Watanabe S** (2019) End-to-end neural speaker diarization with self-attention. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. Singapore. Institute of Electrical and Electronics Engineers (IEEE).
- Gosztolya G, Vincze V, Tóth L, Pákási M, Kálmán J and Hoffmann I** (2019) Identifying mild cognitive impairment and mild Alzheimer's disease based on spontaneous speech using ASR and linguistic features. *Computer Speech & Language* 53, 181–197.
- Guo Z, Liu Z, Ling Z, Wang S, Jin L and Li Y** (2020) Text classification by contrastive learning and cross-lingual data augmentation for alzheimer's disease detection. In *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain (Online). International Committee on Computational Linguistics (ICCL).
- Henderson SK, Peterson KA, Patterson K, Lambon Ralph MA and Rowe JB** (2023) Verbal fluency tests assess global cognitive status but have limited diagnostic differentiation: evidence from a large-scale examination of six neurodegenerative diseases. *Brain Communications* 5(2), ead042.
- Hlédiková A, Woszczyk D, Akman A, Demetriou S and Schuller B** (2022) Data augmentation for dementia detection in spoken language. Preprint, [arXiv:2206.12879](https://arxiv.org/abs/2206.12879).
- Ilias L and Askounis D** (2022) Multimodal deep learning models for detecting dementia from speech and transcripts. *Frontiers in Aging Neuroscience* 14, 830943.
- Ilse M, Tomczak JM and Forré P** (2021) Selecting data augmentation for simulating interventions. In *International Conference on Machine Learning*. Virtual Event. Proceedings of Machine Learning Research (PMLR).
- Jung W, Jun E, Suk H-I and Alzheimer's Disease Neuroimaging Initiative** (2021) Deep recurrent model for individualized prediction of Alzheimer's disease progression. *NeuroImage* 237, 118143.
- Lanzi AM, Saylor AK, Fromm D, Liu H, MacWhinney B and Cohen ML** (2023) DementiaBank: Theoretical rationale, protocol, and illustrative analyses. *American Journal of Speech-Language Pathology* 32(2), 426–438.
- Li VOK, Lam JCK, Han Y, Cheung LYL, Downey J, Kaistha T and Gozes I** (2021) Designing a protocol adopting an artificial intelligence (AI)-driven approach for early diagnosis of late-onset Alzheimer's disease. *Journal of Molecular Neuroscience* 71(7), 1329–1337.
- Li Y, Du M, Song R, Wang X and Wang Y** (2023) A survey on fairness in large language models. Preprint, [arXiv:2308.10149](https://arxiv.org/abs/2308.10149).
- Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, Katz R, Himmelfarb J, Bansal N and Lee S-I** (2020) From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence* 2(1), 56–67.
- Lundberg SM and Lee S-I** (2017) A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems* 30, 4768–4777.
- Luz S, Garcia SDF, Haider F, Fromm D, MacWhinney B, Lanzi A, Chang Y-N, Chou C-J and Liu Y-C** (2024) Connected speech-based cognitive assessment in Chinese and English. Preprint, [arXiv:2406.10272](https://arxiv.org/abs/2406.10272).
- Maheux E, Koval I, Ortholand J, Birkenbihl C, Archetti D, Bouteloup V, Epelbaum S, Dufouil C, Hofmann-Apitius M and Durrleman S** (2023) Forecasting individual progression trajectories in Alzheimer's disease. *Nature Communications* 14(1), 761.
- Mesnard T, Hardin C, Dadashi R, Bhupatiraju S, Pathak S, Sifre L, Rivière M, Kale MS, Love J, Tafti P, Hussenot L and Sessa PG** (2024) Gemma: Open models based on gemini research and technology. Preprint, [arXiv:2403.08295](https://arxiv.org/abs/2403.08295).
- Mirabnabhazam G, Ma D, Beaulac C, Lee S, Popuri K, Lee H, Cao J, Galvin JE, Wang L and Beg MF** (2023) Predicting time-to-conversion for dementia of Alzheimer's type using multi-modal deep survival analysis. *Neurobiology of Aging* 121, 139–156.
- Moor M, Banerjee O, Abad ZSH, Krumholz HM, Leskovec J, Topol EJ and Rajpurkar P** (2023) Foundation models for generalist medical artificial intelligence. *Nature* 616(7956), 259–265.
- Moreno-Barea FJ, Jerez JM and Franco L** (2020) Improving classification accuracy using data augmentation on small data sets. *Expert Systems with Applications* 161, 113696.

- Mueller KD, Hermann B, Mecollari J and Turkstra LS (2018) Connected speech and language in mild cognitive impairment and Alzheimer's disease: A review of picture description tasks. *Journal of Clinical and Experimental Neuropsychology* 40(9), 917–939.
- Nguyen M, Baker A, Neo C, Roush A, Kirsch A and Schwartz-Ziv R (2024) Turning up the heat: Min-p sampling for creative and coherent LLM outputs. Preprint, [arXiv:2407.01082](https://arxiv.org/abs/2407.01082).
- Ning Z, Xiao Q, Feng Q, Chen W and Zhang Y (2021) Relation-induced multi-modal shared representation learning for Alzheimer's disease diagnosis. *IEEE Transactions on Medical Imaging* 40(6), 1632–1645.
- OpenAI (n.d.-a) Chat API Reference. Available at <https://platform.openai.com/docs/api-reference/chat> (accessed 31 March 2024).
- OpenAI (n.d.-b) Text Generation Models. Available at <https://platform.openai.com/docs/guides/text-generation> (accessed 31 March 2024).
- OpenAI (n.d.-c) Text-to-Speech Prompting. Available at <https://platform.openai.com/docs/guides/speech-to-text/prompting> (accessed 31 March 2024).
- Östberg P, Fernaeus S-E, Hellström Å, Bogdanović N and Wahlund L-O (2005) Impaired verb fluency: A sign of mild cognitive impairment. *Brain and Language* 95(2), 273–279.
- Patel N, Peterson KA, Ingram RU, Storey I, Cappa SF, Catricala E, Halai A, Patterson KE, Lambon Ralph MA and Rowe JB (2022) A 'Mini Linguistic State Examination' to classify primary progressive aphasia. *Brain Communications* 4(2), fcab299.
- Petti U, Baker S and Korhonen A (2020) A systematic literature review of automatic Alzheimer's disease detection from speech and language. *Journal of the American Medical Informatics Association* 27(11), 1784–1797.
- Pistono A, Pariente J, Bézy C, Lemesle B, Le Men J and Jucla M (2019) What happens when nothing happens? An investigation of pauses as a compensatory mechanism in early Alzheimer's disease. *Neuropsychologia* 124, 133–143.
- Radford A, Kim JW, Xu T, Brockman G, McLeavey C and Sutskever I (2023) Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*. Honolulu, Hawaii, USA. Proceedings of Machine Learning Research (PMLR).
- Reimers N and Gurevych I (2019) Sentence-bert: Sentence embeddings using siamese bert-networks. Preprint, [arXiv:1908.10084](https://arxiv.org/abs/1908.10084).
- Sachdeva R, Tutek M and Gurevych I (2023) CATFOOD: Counterfactual augmented training for improving out-of-domain performance and calibration. Preprint, [arXiv:2309.07822](https://arxiv.org/abs/2309.07822).
- Schramowski P, Turan C, Andersen N, Rothkopf CA and Kersting K (2022) Large pre-trained language models contain human-like biases of what is right and wrong to do. *Nature Machine Intelligence* 4(3), 258–268.
- Seedat N, Huynh N, van Breugel B and van der Schaar M (2023) Curated LLM: Synergy of LLMs and Data Curation for tabular augmentation in ultra low-data regimes. Preprint, [arXiv:2312.12112](https://arxiv.org/abs/2312.12112).
- Shankle WR, Romney AK, Hara J, Fortier D, Dick MB, Chen JM, Chan T and Sun X (2005) Methods to improve the detection of mild cognitive impairment. *Proceedings of the National Academy of Sciences* 102(13), 4919–4924.
- Shorten C and Khoshgoftaar TM (2019) A survey on image data augmentation for deep learning. *Journal of Big Data* 6(1), 1–48.
- Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, Scales N, Tanwani A, Cole-Lewis H and Pfohl S (2023) Large language models encode clinical knowledge. *Nature* 620(7972), 172–180.
- Sun A (n.d.) "Jieba" (Chinese for "to stutter") Chinese text segmentation. Available at <https://github.com/fxsjy/jieba> (accessed 31 March 2024).
- Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF and Ting DSW (2023) Large language models in medicine. *Nature Medicine* 29(8), 1930–1940.
- Touvron H, Lavril T, Izacard G, Martinet X, Lachaux M-A, Lacroix T, Rozière B, Goyal N, Hambro E and Azhar F (2023) Llama: Open and efficient foundation language models. Preprint, [arXiv:2302.13971](https://arxiv.org/abs/2302.13971).
- Wang C, Li Y, Tsuboshita Y, Sakurai T, Goto T, Yamaguchi H, Yamashita Y, Sekiguchi A, Tachimori H and Alzheimer's Disease Neuroimaging Initiative (2022) A high-generalizability machine learning framework for predicting the progression of Alzheimer's disease using limited data. *NPJ Digital Medicine* 5(1), 43.
- Wang N, Cao Y, Hao S, Shao Z and Subbalakshmi K (2021) Modular multi-modal attention network for Alzheimer's disease detection using patient audio and language data. In *Interspeech 2021. Brno, Czechia. International Speech Communication Association (ISCA)*.
- Wei J, Wang X, Schuurmans D, Bosma M, Xia F, Chi E, Le QV and Zhou D (2022) Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems* 35, 24824–24837.
- Xu L, Wu H, He C, Wang J, Zhang C, Nie F and Chen L (2022) Multi-modal sequence learning for Alzheimer's disease progression prediction with incomplete variable-length longitudinal data. *Medical Image Analysis* 82, 102643.
- Xue C, Karjadi C, Paschalidis IC, Au R and Kolachalama VB (2021) Detection of dementia on voice recordings using deep learning: A Framingham Heart Study. *Alzheimer's Research & Therapy* 13, 1–15.
- Yang L, Wei W, Li S, Li J and Shinozaki T (2022a) Augmented adversarial self-supervised learning for early-stage Alzheimer's speech detection. In *Interspeech 2022. Incheon, Korea. International Speech Communication Association (ISCA)*.
- Yang Q, Li X, Ding X, Xu F and Ling Z (2022b) Deep learning-based speech analysis for Alzheimer's disease detection: A literature review. *Alzheimer's Research & Therapy* 14(1), 186.