



Reasoning-Based Approach with Chain-of-Thought for Alzheimer's Detection Using Speech and Large Language Models

Chanwoo Park¹, Anna Seo Gyeong Choi², Sunghye Cho³, Chanwoo Kim^{1†}

¹Department of Artificial Intelligence, Korea University, South Korea

²Information Science, Cornell University, United States

³Department of Linguistics, University of Pennsylvania, United States

{cksdn1290, chanwcom}@korea.ac.kr

Abstract

Societies worldwide are rapidly entering a super-aged era, making elderly health a pressing concern. The aging population is increasing the burden on national economies and households. Dementia cases are rising significantly with this demographic shift. Recent research using voice-based models and large language models (LLM) offers new possibilities for dementia diagnosis and treatment. Our Chain-of-Thought (CoT) reasoning method combines speech and language models. The process starts with automatic speech recognition to convert speech to text. We add a linear layer to an LLM for Alzheimer's disease (AD) and non-AD classification, using supervised fine-tuning (SFT) with CoT reasoning and cues. This approach showed an 16.7% relative performance improvement compared to methods without CoT prompt reasoning. To the best of our knowledge, our proposed method achieved state-of-the-art performance in CoT approaches.

Index Terms: Chain-of-Thought, dementia detection, cue, speech recognition, large language model

1. Introduction

Dementia is a neurodegenerative disorder characterized by the progressive decline of cognitive functions. It significantly impacts patients' ability to perform daily activities, thereby severely affecting their quality of life. According to the World Health Organization (WHO), approximately 50 million people worldwide are living with dementia, and nearly 10 million new cases are diagnosed each year [1]. Early detection is crucial, as it can substantially reduce treatment costs and mitigate the disease's impact, improving outcomes for patients and caregivers alike. Moreover, dementia not only poses challenges for individuals but also places a significant economic and emotional burden on families and healthcare systems globally. Research suggests that raising awareness about dementia and improving access to diagnostic tools can enhance early intervention efforts. Public health strategies focusing on prevention, such as promoting a healthy lifestyle, managing cardiovascular risk factors, and encouraging cognitive engagement, may also play a vital role in reducing the prevalence of dementia in aging populations.

According to recent research, efforts are being made to solve various medical problems using large language model (LLM). In particular, various methods for the early diagnosis of Alzheimer's Disease (AD) using speech and verbal information have been proposed. Researchers have identified acoustic and linguistic patterns experienced by AD patients, such as

forgetting words, grammatical errors, increased pauses during speech, and repeated use of words. Various studies are being conducted to classify patients using LLM. For comparison with studies dealing with the same problem, the data provided in the Alzheimer's Dementia Recognition through Spontaneous Speech (ADReSS) dataset [2] can be used to effectively classify AD. The ADReSS dataset consists of voice recordings labeled as AD or non-AD, where patients describe a picture of a "cookie thief."

In this study, the pre-trained Whisper model can be used to transcribe speech into text and utilize linguistic information. Fine-tuning is essential because state-of-the-art pre-trained LLMs show very strong performance but have limitations in highly specific tasks. However, fine-tuning all parameters of an LLM poses challenges due to its large capacity. Updating all parameters can lead to the loss of existing knowledge within the LLM and requires significant computational resources and memory.

To address this issue, Parameter-Efficient Fine-Tuning (PEFT) [3] is employed to update only a subset of parameters rather than all of them. This approach helps preserve the existing knowledge of the LLM while reducing memory usage and computational costs. Among PEFT methods, this study selects Low-Rank Adaptation (LoRA) [4] fine-tuning. LoRA works by learning additional low-rank parameters while keeping the existing weights of the model fixed. In other words, instead of modifying the original weights directly, LoRA adds learnable parameters to capture new information.

Previous AD detection studies primarily relied on single-modality approaches using either acoustic or linguistic features, facing limitations in addressing both aspects simultaneously. This study overcomes these constraints by employing a Chain-of-Thought (CoT) [5] methodology that strategically identifies and leverages critical cues within individual modalities, enhancing detection performance without relying on multimodal integration.

2. Related Works

Existing research on dementia diagnosis emphasizes the widespread use of cognitive screening tools like the mini-mental state examination (MMSE) and clinical dementia rating (CDR) in primary care settings, particularly in high-income countries [6]. Diagnostic criteria from the DSM-IV/V and ICD-10 remain foundational, though middle-income countries often lack resources for advanced neuroimaging or biomarker testing [6]. Recent advances highlight the potential of AI-driven retinal imaging to detect amyloid plaques for early diagnosis and autophagy-mediated tau protein degradation as a novel therapeutic target. Biomarker frameworks like the ATN

[†]Corresponding author

The source code is available at GitHub by <https://doi.org/10.5281/zenodo.15511013>

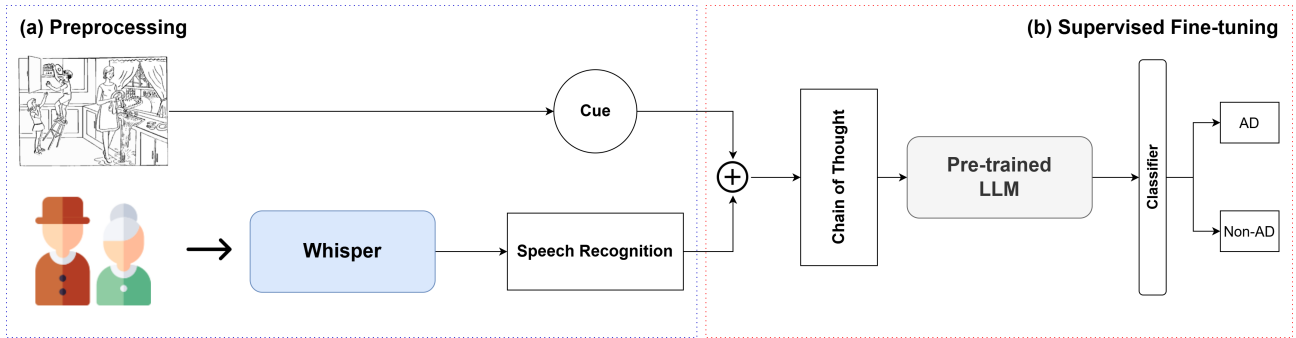


Figure 1: The process of LLM for dementia detection. (a) The preprocessing stage involves speech recognition of participant responses using a speech model and generation of important visual cues. (b) In the supervised fine-tuning (SFT) stage, utilize the pre-trained language model Llama to classify between AD and non-AD cases.

system (Amyloid- β , Tau, Neurodegeneration) are reshaping Alzheimer’s disease diagnostics toward preclinical detection [7, 8]. Studies consistently note significant underdiagnosis in primary care, exacerbated by inadequate clinician training and caregiver communication challenges [6, 9].

Recent advancements in machine learning, speech processing, and natural language processing (NLP) have enabled the development of innovative approaches for detecting AD. These methods leverage speech, language models, and multimodal data to identify early biomarkers of cognitive decline. Below is a comprehensive summary of the two key approaches: speech-based methods and language model approaches.

2.1. Speech-Based Approaches

Speech analysis has emerged as a promising non-invasive tool for early detection of AD due to its strong correlation with cognitive health. Researchers have identified several acoustic and linguistic features that differentiate individuals with AD from healthy controls. Studies have shown that individuals with AD exhibit slower speech rates, longer pauses, increased hesitations, and more frequent grammatical errors compared to healthy individuals [10, 11]. Feature sets like eGeMAPS (extended Geneva Minimalistic Acoustic Parameter Set) have proven effective in capturing these acoustic anomalies. For instance, eGeMAPS [12] achieved a classification accuracy of 74.1% in distinguishing AD patients from healthy controls using leave-one-subject-out (LOSO) evaluation. Advanced deep learning models such as Wav2Vec2 [13] and HuBERT [14] have further improved performance by learning robust speech representations from raw audio data [15].

2.2. Language Model Approaches

Models like Llama2-7B [16] and Mistral-7B [17] have demonstrated impressive performance when fine-tuned for AD detection tasks [18]. ChatGPT [19] has been used to analyze transcriptions of spontaneous speech, extracting linguistic patterns such as repetition, topic-switching, and fluency irregularities that are characteristic of AD [20].

3. Methodology

3.1. Dataset

Our experiment utilized data from the cookie theft picture description task of the Boston Diagnostic Aphasia Exam (BDAE)

[21–23], which is part of DementiaBank’s Pitt Corpus [2]. The transcripts, annotated using the CHAT coding system [24], were acoustically enhanced through static noise removal. Audio volume was normalized across all speech segments. The dataset consists of speech samples from AD and non-AD English-speaking participants for the cookie theft picture description task, balanced by age, gender, and disease status, comprising a total of 156 participants with 1,955 speech segments from 78 non-AD participants and 2,122 speech segments from 78 AD participants. The data was divided into a training set of 108 participants, consisting of 48 AD participants (55 minutes 46 seconds) and 48 non-AD participants (1 hour 14 minutes), and a test set of 48 participants (24 AD and 24 non-AD participants, 1 hour 6 minutes). In speech processing, cookie theft detection (CTD) involves evaluating cognitive function by analyzing a person’s ability to describe a specific image.

3.2. Frameworks

Our proposed dementia detection algorithm, which utilizes language and speech models, respectively, is illustrated in Figure 1. First, we use Whisper [25], a speech recognition model, to perform text transcription, converting spoken language into written text from participant’s voice data. Next, we generate relevant cues from the given cookie theft picture that could serve as important indicators. Finally, we apply LoRA to a pre-trained LLM and perform supervised fine-tuning to classify whether an individual has AD or non-AD.

The data contains audio conversations between the investigator and participants. Since the investigator tends to prompt responses when participants hesitate, we aimed to eliminate this influence. As we wanted to focus solely on the participants’ speech, we utilized the normalized audio chunks from each participant. For speech recognition, we employed the large-v2 version of the Whisper model.

3.3. Supervised Fine-tuning

Supervised fine-tuning (SFT) is a crucial technique for adapting pre-trained LLMs to specific domains or tasks. Using Llama3.2-1B-Instruct [26], as shown in Figure 1, the LLM takes text input to generate rich contextual embeddings, and a classification head, which is a linear layer, is added on top of the pre-trained LLM to distinguish between AD and non-AD cases.

3.3.1. Chain-of-Thought

Language models have shown various benefits when scaled up, including improved performance and sample efficiency. However, scaling model size alone has proven insufficient to achieve high performance on challenging tasks like arithmetic, common sense, and symbolic reasoning. Traditional few-shot prompting [27] often proves ineffective for tasks requiring reasoning abilities and may not significantly improve even with larger language models. Chain-of-thought (CoT) [5] explores language models’ ability to perform few-shot prompting for reasoning tasks. CoT consists of intermediate reasoning steps that lead to the final output. This reasoning can be applied to tasks such as mathematical problems, common sense reasoning, and symbolic manipulation, and theoretically can potentially be applied to any task that humans can solve through language. Since LLM generate sentences sequentially like filling in blanks, CoT prompting enables complex reasoning through intermediate reasoning steps. When combined with direct answer prompting, it can yield better results in complex tasks that require reasoning before responding.

It then generates important cues from the CTD and integrates them into CoT prompts. The cues entered at the CoT prompts are listed in Table 1. Specifically, we calculate the proportion of these listed cues expressed by participants in the text obtained from speech recognition, and provide prompts to infer whether the case is AD or non-AD.

stool, sink, dish, wash, jar, cookie, child, mother, window, cabinet, kitchen, water

Table 1: Cues for Chain-of-Thought (CoT) prompts.

3.3.2. Parameter-Efficient Fine-Tuning

Many natural language processing applications depend on adapting a single large pre-trained language model to various downstream applications. One solution to this challenge is PEFT, with LoRA being a notable example. LoRA is a specialized PEFT technique that efficiently fine-tunes LLM by adjusting only a small subset of parameters. This method freezes the weights of the pre-trained model and inserts low-rank decomposition matrices into each layer, significantly reducing the number of trainable parameters. This approach maintains performance comparable to full fine-tuning while decreasing computational and memory requirements.

In our experiments using PEFT, we configured LoRA with the following parameters: a rank value of 16 in the low-dimensional space, an adapter scaling value of 16, and a dropout probability of 0.01 prevent overfitting.

Method	Acc (%)	F1 (%)
Baseline	75.00	74.83
Zero-shot	47.92	32.39
Few-shot	54.17	44.54
CoT	83.33	83.22

Table 2: Accuracy and F1-score results of our proposed method.

4. Experiment results and analysis

4.1. Experimental Setup

For our experiments, we trained the model using the original split of 108 participants for the SFT training set, and conducted evaluations using the remaining 48 participants in the test set. We trained the model on a NVIDIA GeForce RTX 4090 GPU. For the hyperparameter configuration, we used a batch size of 8 and fine-tuned the model with a learning rate of $1e-4$ over three phases using the AdamW optimizer [32]. We applied a linear learning rate scheduler that incorporates a sustain phase—keeping the learning rate constant after the warm-up period before it decays linearly—and set the weight decay to 0.001.

4.2. Results

We conducted experiments comparing the baseline model trained with SFT without CoT prompts against zero-shot, few-shot, and our proposed CoT approaches. The CoT prompts we used are shown in Figure 2. The experimental results of the proposed LLM model are shown in Table 2. The baseline results showed comparable performance to previous studies [2], while both zero-shot and few-shot settings demonstrated poor performance. In contrast, we observed superior performance when incorporating CoT prompts that guide the reasoning process.

Reasoning Guidance

You are an expert neurologist specializing in dementia diagnosis. Analyze the participant's description of the kitchen scene image for signs of cognitive impairment.

Analysis Steps:

1. Check for mention of key cue elements (e.g., stool, sink, dish, etc.).
2. Evaluate awareness of safety hazards (e.g., stool, water, window).
3. Assess logical flow and narrative coherence using connecting words.
4. Integrate all findings to determine the likelihood of dementia.

Figure 2: Example of the reasoning guidance Chain-of-Thought (CoT) prompts provided to the model for diagnosing dementia.

4.3. Comparison of results with other studies

The Table 3 compares the performance of various models across different modalities and datasets for the ADReSS [2] benchmark. Traditional methods [2, 28] like linear discriminant analysis (LDA) [33], which rely on acoustic and linguistic features, achieve accuracy and F1-scores of 76.8% and 74.5%, respectively, indicating moderate performance [2]. This study [28], utilizing a multimodal system, identified linguistic and paralinguistic characteristics of dementia through automated screening tools. The researchers achieved an accuracy of 81.3% using support vector machines (SVM) [34], demonstrating that deep neural network embeddings and ensemble learning are viable approaches for objective dementia assessment.

This study [15] utilizing spontaneous speech in AD detection demonstrates a non-invasive and cost-effective approach by implementing self-supervised learning (SSL) [35] models with joint fine-tuning strategies, combined with multitask learning and data augmentation. Fine-tuning pre-trained SSL models, along with multitask learning and data augmentation, enhances the effectiveness of universal speech representations in AD detection. Speech-based models [15] such as Wav2Vec2 [13] and HuBERT [14] show slightly lower results, with Wav2Vec2 achieving 73.7% accuracy and 72.8% F1-score, while HuBERT

Model	Pre-trained Model		Acc (%)	F1 (%)
	Speech	LM		
LDA [2]	acoustic & linguistic features		76.8	74.5
SVM [28]	acoustic & linguistic features		81.3	-
Wav2Vec2 [15]	✓	✗	73.7	72.8
HuBERT [15]	✓	✗	74.2	73.6
BERT [29]	✗	✓	83.3	83.9
Bi-LSTM [30]	✗	✓	81.3	81.2
BERT + SpeechBERT [31]	✓	✓	82.9 ± 1.6	82.9 ± 1.9
Ours (Llama3.2-1B)	✓	✓	83.3	83.2
	✗	✓	87.5	87.5

Table 3: Experimental performance comparison using models of different modalities on the ADRess Challenge set.

scores 74.2% accuracy and 73.6% F1-score, suggesting limited improvement over traditional approaches.

The study [29], which utilized the effectiveness of speech transcript representations obtained from the BERT natural language processing model, compared it with more clinically interpretable language feature-based methods. Both the feature-based approach and fine-tuned BERT model achieved an accuracy of 83.3% and an F1-score of 83.9% using a limited number of linguistic features.

Subsequently, in a text-based study [36], they experimented with a hierarchical neural network equipped with an attention mechanism trained on linguistic features, along with an acoustic-based system. Using bidirectional long short-term memory (bi-LSTM) with attention [30], they achieved performance scores of 81.3% and 81.2% for accuracy and F1-score, respectively.

A study [31] using dual-modality BERT and SpeechBERT [37] models explored transfer learning techniques for AD classification and MMSE regression tasks. The transfer learning models were pre-trained on general large-scale datasets and fine-tuned and tested using the ADRess dataset, achieving 82.9 ± 1.56 and 82.9 ± 1.86 for accuracy and F1-score, respectively.

Finally, the proposed model using the ground truth (GT) text format recorded by the CHAT coding system and fine-tuned with CoT prompts achieved competitive results with an accuracy of 87.50% and an F1-score of 87.48%. Our approach demonstrates robust performance across various input types, highlighting its versatility in handling complex data integration tasks. Additionally, by employing the lightweight Llama3.2-1B model with fewer parameters, the approach ensures efficiency without compromising performance.

4.4. Ablation Study

The results of ablation study are presented in Table 4. We fine-tuned the model using cues and CoT prompts, then compared the speech-to-text transcriptions with the ground truth content recorded using the CHAT coding system. While the performance of the ASR model is crucial, considering the challenges in recognizing speech from middle-aged and elderly participants, we found that fine-tuning the LLM with the transcribed text yielded positive results.

5. Conclusions

Advancements in AI technology, particularly LLM and multi-modal approaches, are driving transformative changes in dementia research and treatment. Models that combine automatic speech recognition (ASR) have demonstrated a 11.1%

Method	Acc (%)	F1 (%)
Baseline (ASR CoT)	83.33	83.22
ground truth (SFT)	83.33	83.30
ground truth (CoT)	87.50	87.48

Table 4: Comparison of accuracy and F1-score for different training strategies in the ablation study. The Baseline (ASR CoT) uses automatic speech recognition outputs without ground truth supervision. The ground truth supervised fine-tuning (SFT) is performed without Chain-of-Thought (CoT) prompts, while ground truth (CoT) applies CoT prompts during supervised fine-tuning.

improvement in accuracy for early diagnosis and classification of Alzheimer’s disease (AD) compared to traditional methods, underscoring the value of multimodal approaches. The Chain-of-Thought (CoT) [5] reasoning framework enhances AI performance and transparency by breaking down complex diagnostic tasks into structured steps, thereby increasing its clinical applicability. Additionally, speech-based AI models offer a non-invasive and scalable method for monitoring cognitive decline [20], making them a valuable tool for early detection and management in aging societies. AI frameworks that integrate cues and neuroimaging data not only achieve high diagnostic accuracy but also contribute to predicting disease progression.

6. Acknowledgment

This work was supported by the IITP(Institute of Information & Communications Technology Planning & Evaluation)-ITRC(Information Technology Research Center) grant funded by the Korea government(Ministry of Science and ICT)(IITP-2025-RS-2024-00436857), IITP grant funded by the Korea government(MSIT) (No. RS-2019-II190079, Artificial Intelligence Graduate School Program(Korea University)), the ”Leaders in Industry-university Cooperation 3.0” Project, supported by the Ministry of Education and National Research Foundation of Korea, and IITP under the artificial intelligence star fellowship support program to nurture the best talents (IITP-2025-RS-2025-02304828) grant funded by the Korea government(MSIT).

7. References

- [1] G. Livingston, J. Huntley, A. Sommerlad, D. Ames, C. Ballard, S. Banerjee, and N. Mukadam, “Dementia prevention, intervention, and care: 2020 report of the lancet commission,” *The Lancet*,

- vol. 396, no. 10248, pp. 413–446, 2020.
- [2] S. Luz, F. Haider, S. de la Fuente Garcia, D. Fromm, and B. MacWhinney, “Alzheimer’s dementia recognition through spontaneous speech,” *Frontiers in Computer Science*, vol. 3, p. 780169, 2021.
 - [3] N. Ding, Y. Qin, G. Yang, F. Wei, Z. Yang, Y. Su, and M. Sun, “Parameter-efficient fine-tuning of large-scale pre-trained language models,” *Nature Machine Intelligence*, vol. 5, no. 3, pp. 220–235, 2023.
 - [4] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, and W. Chen, “Lora: Low-rank adaptation of large language models,” *arXiv preprint arXiv:2106.09685*, 2021.
 - [5] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, and D. Zhou, “Chain-of-thought prompting elicits reasoning in large language models,” in *Advances in Neural Information Processing Systems*, vol. 35, 2022, pp. 24 824–24 837.
 - [6] L. N. Pelegrini, G. M. Mota, C. F. Ramos, E. Jesus, and F. A. Vale, “Diagnosing dementia and cognitive dysfunction in the elderly in primary health care: a systematic review,” *Dementia & Neuropsychologia*, vol. 13, pp. 144–153, 2019.
 - [7] F. M. Elahi and B. L. Miller, “A clinicopathological approach to the diagnosis of dementia,” *Nature Reviews Neurology*, vol. 13, no. 8, pp. 457–476, 2017.
 - [8] A. S. Abdul Manap, R. Almadodi, S. Sultana, M. G. Sebastian, K. S. Kavani, V. E. Lyenouq, and A. Shankar, “Alzheimer’s disease: A review on the current trends of the effective diagnosis and therapeutics,” *Frontiers in Aging Neuroscience*, vol. 16, p. 1429211, 2024.
 - [9] L. Stokes, H. Combes, and G. Stokes, “The dementia diagnosis: a literature review of information, understanding, and attributions,” *Psychogeriatrics*, vol. 15, no. 3, pp. 218–225, 2015.
 - [10] X. Qi, Q. Zhou, J. Dong, and W. Bao, “Noninvasive automatic detection of alzheimer’s disease from spontaneous speech: a review,” *Frontiers in Aging Neuroscience*, vol. 15, p. 1224723, 2023.
 - [11] F. Haider, S. De La Fuente, and S. Luz, “An assessment of paralinguistic acoustic features for detection of alzheimer’s dementia in spontaneous speech,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 2, pp. 272–281, 2019.
 - [12] J. Chen, J. Ye, F. Tang, and J. Zhou, “Automatic detection of alzheimer’s disease using spontaneous speech only,” in *Interspeech*, vol. 2021. NIH Public Access, August 2021, p. 3830.
 - [13] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” in *Advances in neural information processing systems*, vol. 33, 2020, pp. 12 449–12 460.
 - [14] W. N. Hsu, B. Bolte, Y. H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, “Hubert: Self-supervised speech representation learning by masked prediction of hidden units,” *IEEE/ACM transactions on audio, speech, and language processing*, vol. 29, pp. 3451–3460, 2021.
 - [15] M. Chen, C. Miao, J. Ma, S. Wang, and J. Xiao, “Exploring multi-task learning and data augmentation in dementia detection with self-supervised pretrained models,” in *Proc. INTERSPEECH*, vol. 2023, 2023, pp. 5037–5041.
 - [16] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, and T. Scialom, “Llama 2: Open foundation and fine-tuned chat models,” *arXiv preprint arXiv:2307.09288*, 2023.
 - [17] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. D. L. Casas, and W. E. Sayed, “Mistral 7b,” *arXiv preprint arXiv:2310.06825*, 2023.
 - [18] F. Casu, E. Grosso, A. Lagorio, and G. A. Trunfio, “Optimizing and evaluating pre-trained large language models for alzheimer’s disease detection,” in *2024 32nd Euromicro International Conference on Parallel, Distributed and Network-Based Processing (PDP)*. IEEE, March 2024, pp. 277–284.
 - [19] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, and B. McGrew, “Gpt-4 technical report,” *arXiv preprint arXiv:2303.08774*, 2023.
 - [20] J. U. Bang, S. H. Han, and B. O. Kang, “Alzheimer’s disease recognition from spontaneous speech using large language models,” *ETRI Journal*, vol. 46, no. 1, pp. 96–105, 2024.
 - [21] I. T. Draper, “The assessment of aphasia and related disorders,” *Journal of Neurology, Neurosurgery, and Psychiatry*, vol. 36, no. 5, p. 894, 1973.
 - [22] H. Goodglass and E. Kaplan, *Boston diagnostic aphasia examination booklet*. Lea & Febiger, 1983.
 - [23] H. Goodglass, E. Kaplan, and S. Weintraub, *BDAE: The Boston diagnostic aphasia examination*. Philadelphia, PA: Lippincott Williams & Wilkins, 2001.
 - [24] B. Macwhinney, “The chldes project part 1: The chat transcription format,” in *The CHILDES Project*, 2009.
 - [25] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, “Robust speech recognition via large-scale weak supervision,” in *International Conference on Machine Learning*. PMLR, July 2023, pp. 28 492–28 518.
 - [26] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, and R. Ganapathy, “The llama 3 herd of models,” *arXiv preprint arXiv:2407.21783*, 2024.
 - [27] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, and D. Amodei, “Language models are few-shot learners,” in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 1877–1901.
 - [28] Z. S. Syed, M. S. S. Syed, M. Lech, and E. Pirogova, “Automated recognition of alzheimer’s dementia using bag-of-deep-features and model ensembling,” *IEEE Access*, vol. 9, pp. 88 377–88 390, 2021.
 - [29] A. Balagopalan and J. Novikova, “Comparing acoustic-based approaches for alzheimer’s disease detection,” *arXiv preprint arXiv:2106.01555*, 2021.
 - [30] N. Cummins, Y. Pan, Z. Ren, J. Fritsch, V. S. Nallanthighal, H. Christensen, and A. Härmä, “A comparison of acoustic and linguistics methodologies for alzheimer’s dementia recognition,” in *Interspeech 2020*. ISCA-International Speech Communication Association, October 2020, pp. 2182–2186.
 - [31] Y. Zhu, X. Liang, J. A. Batsis, and R. M. Roth, “Exploring deep transfer learning techniques for alzheimer’s dementia detection,” *Frontiers in Computer Science*, vol. 3, p. 624683, 2021.
 - [32] I. Loshchilov, “Decoupled weight decay regularization,” *arXiv preprint arXiv:1711.05101*, 2017.
 - [33] P. Xanthopoulos, P. M. Pardalos, and T. B. Trafalis, “Linear discriminant analysis,” in *Robust data mining*, 2013, pp. 27–33.
 - [34] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf, “Support vector machines,” *IEEE Intelligent Systems and their applications*, vol. 13, no. 4, pp. 18–28, 1998.
 - [35] X. Liu, F. Zhang, Z. Hou, L. Mian, Z. Wang, J. Zhang, and J. Tang, “Self-supervised learning: Generative or contrastive,” *IEEE transactions on knowledge and data engineering*, vol. 35, no. 1, pp. 857–876, 2021.
 - [36] P. Zhou, W. Shi, J. Tian, Z. Qi, B. Li, H. Hao, and B. Xu, “Attention-based bidirectional long short-term memory networks for relation classification,” in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, August 2016, pp. 207–212.
 - [37] Y. S. Chuang, C. L. Liu, H. Y. Lee, and L. S. Lee, “Speechbert: An audio-and-text jointly learned language model for end-to-end spoken question answering,” *arXiv preprint arXiv:1910.11559*, 2019.