Language-Based Dementia Classification Should Consider Model Cognition for Interpretability

Anonymous Author(s)
Affiliation
Address
email

Abstract

Current approaches to dementia detection in machine learning based on language often treat the task as an end-to-end binary classification problem, directly classifying a person's audio or transcripts to a final diagnostic label (e.g., dementia or cognitively normal). While these prior techniques can be effective for accuracy, because predictions are made in a single step, this binary approach overlooks the progressive nature of cognitive decline and lacks the interpretive analyses that clinicians rely on in clinical settings. Furthermore, these approaches face interpretability limitations, particularly in terms of the linguistic features the models are focusing on. **This position paper argues that language-based dementia detection research should be re-framed to include cognition-based reasoning, specifically by probing model cognition through structured, fine-grained input–output analyses, which allow clearer understanding of how ML systems reason in this task.** This new direction will help advance current ML models towards dementia detection frameworks that are more interpretable and clinically trustworthy.

1 Introduction: Diagnosing Dementia Isn't Just a Classification Task

Machine learning (ML) and deep learning have become important tools in healthcare for assisting in the analysis of medical records, accelerating testing, and predicting disease risks [46, 42]. ML could be particularly useful for screening for cognitive impairment, including dementia, which is estimated to affect 57 million people worldwide [39]. Dementia research using ML has assessed input modalities such as demographic data, neuro-imaging, bio-fluid biomarkers, genetic and medical information, and cognitive assessments [8, 14, 52, 43]. Among these, language-based inputs offer a non-invasive and sensitive source of dementia-related cues.

The most common task used in language-based dementia detection is the picture description task [11, 19, 41, 26], wherein a participant describes everything they see in an image (e.g., Cookie Theft Task [16]) as their audio is recorded. The audio is then analyzed through audio signal processing [60, 26, 7, 21] and natural language processing (NLP) of the transcripts [19, 41, 9].

Current ML research on dementia detection is commonly framed as a binary classification task, typically distinguishing Alzheimer's disease (AD) or mild cognitive impairment (MCI) from normal cognition (NC) [54, 44, 4], as seen in Fig. 1. Existing methods range from traditional classifiers built on handcrafted features to large language models (LLMs) applied directly to transcripts, but in both cases models are trained to map inputs to a single diagnostic label [22, 26, 21, 34, 1, 50]. Datasets often reinforce this binary framing, with recordings annotated only with the final diagnosis [53, 5, 30–32], or with cognitive test score (e.g., Mini Mental State Exam (MMSE) [33])(for an overview of datasets, see [12] and Appendix A). While effective at optimizing categorical or score accuracy, it

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

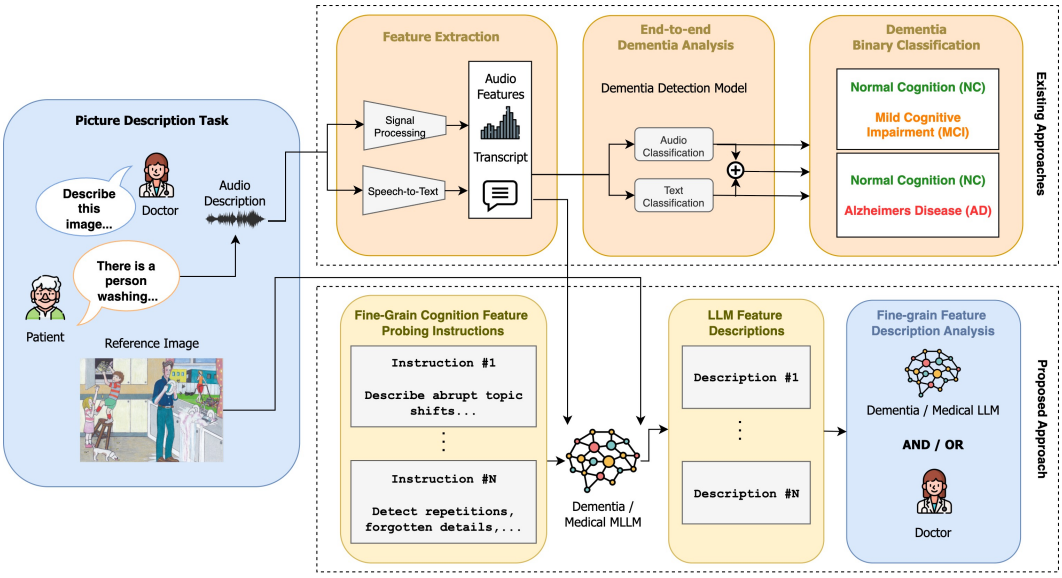

Figure 1: Overview of picture description task for dementia detection. Existing approaches classify audio or text, or both, into NC/MCI or NC/AD, often as a binary task. Our proposed approach probes model cognition through fine-grained feature analysis to assess reasoning and better support clinical diagnosis, offering more interpretable diagnosis. See Appendix B for more details.

collapses the progressive and multidimensional nature of dementia into a single outcome and often lacks interpretability of the model itself.

This raises a fundamental question for the field: *can we improve interpretability of dementia detection models if they reason through the linguistic cues?* Clinical cognitive assessment is inherently a staged process, and clinicians assess narrative completeness and structure, evaluate semantic richness, and integrate these with other observations before reaching a conclusion. When we reduce the problem to a single binary decision, *what aspects of cognition are we ignoring, and what biases are introduced—both by our task design and by the model's own reasoning?*

We argue that detection of cognitive impairment should be reformulated as a structured cognitive inference problem. Rather than collapsing information into a single prediction, LLMs can serve as a cognitive analysis tool: first extracting fine-grained linguistic features from transcripts, then assessing narrative coherence and task faithfulness, and separately analyzing information with acoustic biomarkers before any diagnostic decision for a description of the patient's cognition. Producing these intermediate outputs provides clinicians, who are ultimately responsible for making a diagnosis, the full context and rationale for why the model identified an individual as having cognitive impairment. The clinicians can then decide to include – or not include – aspects of this information in their assessment. To provide more clinically meaningful information, this will require datasets with richer annotations and evaluation methods that include clinically relevant features, to produce dementia detection systems that are interpretable and aligned with real-world healthcare applications.

2 Related Work: The Gap Between ML and Medical Experts

Traditional Methods. Traditional ML and NLP techniques have been widely adopted for language-based dementia detection tasks [26, 7]. Signal processing techniques are used to extract acoustic properties such as pitch, silent pauses, speech rate, harmonic-to-noise ratio and intensity from audio recordings [60, 26, 21]. Traditional NLP methods (e.g., part-of-speech tagging, syntactic constituent length) have been used to extract semantic, lexical, and syntactic properties from transcriptions of the speech [19, 41, 45]. Combined with the rise of traditional classifiers like Random Forest [21] and deep learning models such as convolutional neural networks (CNNs) to extract acoustic embeddings [26, 7], these approaches can help identify complex patterns in detecting cognitive impairment. Furthermore, feature importance (e.g., Random Forest, Support Vector Machines) can shed light on the features the model is relying on [40, 21]. However, while these techniques may achieve strong prediction accuracy, they are limited in their clinical interpretability. For example, [40] found

that Mel-Frequency Cepstral Coefficients (MFCCs) were the most important feature in dementia prediction models, but critically MFCCs currently do not have clinical meaning. Interpretability barriers and challenges with selecting clinically relevant features are important limitations in ML dementia models' ability to contribute in medical diagnostic processes (for more discussion on explainable AI for medical applications, see [3] and [36]).

LLM-Based Methods. With their capability for language understanding, LLMs have introduced a new wave of research in ML for dementia detection. Early works have utilized LLMs to extract embeddings from transcribed speech for downstream classification [62, 48, 23, 35]. More recently, LLMs have been explored through fine-tuning and transfer learning for direct inference of dementia detection tasks [60, 61]. Many of these approaches have shown promise through high classification accuracy, but face notable challenges such as bias [24] and hallucinations, where LLMs generate outputs that sound plausible but are not grounded with accurate information [24, 13, 63]. Recent efforts to develop medical-domain LLMs [57, 56, 27, 28, 51, 49], offer progress, but these models are often trained on general clinical text and are not yet specialized in understanding human cognition for dementia [63]. Currently, state-of-the-art LLMs also lack the embedded clinical knowledge of established clinical diagnostic standards required to understand the dementia task [63, 6]. As a result, they do not properly rely on medically validated markers of cognitive impairment [15, 55]. This causes LLMs to overlook subtle but critical linguistic features that are required to align model cognition with clinical reasoning.

3 Position: Bridging the Gap

We argue that language-based dementia detection research should be re-framed to include cognition-based reasoning, specifically by probing model cognition through structured, fine-grained input-output analyses, which will enable clearer understanding of how ML models reason in this task. This can be achieved by focusing on three key points: **(1) Probing Model Cognition to Better Understand Input Cues:** systematically examining which linguistic features models rely on and how these align with clinical reasoning. **(2) Moving Beyond Binary Classification to Progression-Aware Modeling:** shifting from single-step prediction toward fine-grained, multi-stage outputs that reflect the progressive nature of cognitive decline. **(3) Developing Fine-Grained Datasets for Deeper Understanding:** developing richer, expert-annotated datasets that capture clinically meaningful features to better support interpretable models that are clinically applicable. We believe that these considerations will help design dementia detection ML models that are more aligned with human cognition and clinically trustworthy.

3.1 Probing Alignment between Clinician and Model Cognition to Understand Input Cues

We propose that by systematically probing LLMs and ML models by testing feature importance in the language-based dementia detection task, we can reveal what cues models actually rely on. Clinicians pay attention to features in patients' speech such as pauses, narrative coherence, speech rate, word choice, and limited descriptions of images or recent events [58, 20, 37]. Recent work [63] shows that structured probing in language-based dementia detection by varying prompt types (both human-defined and LLM-generated) can reveal how different inputs affect model accuracy. They demonstrate that by directing LLMs to focus on linguistic features such as places, objects, and actions, and even generate intermediate steps in the detection process, we can understand which aspects of cognition the model is focusing on. Building on this idea, we argue that systematically testing inputs (e.g. masking speech features or removing demographic data), can help serve as interpretability probes to test whether models perform detection through the same cues that clinicians focus on. In parallel, recent trends in multi-agent systems, visual question answering (VQA) [10, 29, 18], and human-in-the-loop [2, 59] show how LLMs can be guided through intermediate steps, either by auxiliary models or by human experts, to probe what the model is focusing on when producing an output. For example, in the case of a picture description task, the model can first analyze the input image and then answer structured questions about the accompanying transcription (see Fig. 1). This approach can enable us to directly test how well model cognition aligns with clinical reasoning, since clinicians themselves can understand the intermediate reasoning steps linking the reference image, patient speech, and model outputs. Understanding model cognition through careful input-output analysis can help make the internal decision-making process of models more transparent and clinically useful by showing cognitive reasoning and process.

3.2 Moving Beyond Binary Classification to Progression-Aware Modeling

To better align model cognition with clinical reasoning, we argue that language-based dementia detection should move beyond end-to-end binary classification, to a continuum that is more consistent with the neurodegenerative process. While many existing datasets include graded measures such as MMSE scores, prior work commonly collapse these into binary categories (e.g. Healthy vs Dementia or NC vs. MCI), losing clinically meaningful granularity. For example, the DAIC-WOZ dataset for depression [17] uses Patient Health Questionnaire-8 (PHQ-8) scores ranging from 0 to 23 in the train set to model severity, showing how scaled outputs can reveal deeper insights into model responses than binary labels. This progression-aware framing for the dementia task will help produce more clinically aligned predictions and also provide understanding of model cognition, showing whether internal representations capture clinically meaningful details in a way clinicians may observe. Furthermore, we suggest using LLMs to provide a clinician-friendly summary to convey these features to improve interpretability and utility in making a clinical evaluation. Relevant demographic and contextual information can then be incorporated into LLM summaries of the intermediate steps (e.g., "For an 81-year-old woman from Sacramento, California reading a passage her speech rate range, (1.8-2.3 syllables/second) is slower than average"). To support this framing, we propose building on existing studies that leverage ML and NLP features by incorporating them into interpretability frameworks that examine how models capture clinically meaningful cues semantic-pragmatic features, narrative coherence, and acoustic features across the full spectrum from cognitively normal to dementia. Rather than focusing solely on accuracy metrics such as accuracy, ROC AUC, or F-1 score, this combined approach enables us to examine how models weigh specific cues across the NC-MCI-AD spectrum. Techniques such as reinforcement learning, fine-tuning of LLMs, and Chain-of-Thought prompting can further help reveal how models understand cognition and can be applied to focus on clinically relevant features and structured reasoning steps in diagnosis.

3.3 Annotating Datasets with Clinically Relevant Features for Deeper Understanding

Progress in dementia detection from audio is also limited in the availability of datasets due to the sensitive nature of medical data. Most publicly available datasets of this modality contain only basic demographic information (e.g., sex, gender), clinical scores such as MMSE, and raw audio recordings, and are only sometimes accompanied by ground truth transcriptions [33]. While this minimal data is valuable, it lacks the depth needed for models to understand the nuanced reasoning of medical experts. We believe that significant advances could be made with the addition of fine-grain annotations in datasets, including annotating existing datasets with expert-labeled features such as markers of emotion in voice, semantic-pragmatic cues, lexico-syntactic features, and other fine-grained indicators a medical professional might consider during diagnosis. These enriched datasets would not only support model training but also serve as interpretability probes, which will enable exploration on whether models observe the same clinically meaningful signals that guide expert reasoning. In this way, dataset development can become integral in understanding model cognition, which further provides the foundation for interpretable and trustworthy dementia detection models.

4 Conclusion: Call for a Fresh Outlook

In this work, we argue that despite advances in applications of ML models for language-based dementia detection, prior techniques have largely optimized for accuracy at the expense of interpretability and alignment with clinical reasoning. Current classification techniques tend to lose the intermediate cognitive assessment steps that human clinicians rely on, which limits the reliability of LLM usage in clinical settings. We assert the need to re-frame dementia detection as a cognitive inference process – one that systematically probes model cognition to understand what lingustic cues models focus on and how these align with clinical reasoning. Proper integration of ML frameworks in real-world clinics will require fine-grained, expert-annotated datasets that capture clinically meaningful features, guiding models to focus on cues that matter most in practice. By combining these ideas with the strengths of existing approaches, we can develop diagnostic tools that are more clinically trustworthy and interpretable.

Impact Statement: This position paper calls for a fundamental shift in language-based dementia detection research toward probing model cognition through structured, fine-grained input–output analyses, enabling clearer understanding of how ML systems reason in this task and paving the way for more interpretable, clinically trustworthy models. See Appendix C for a discussion on limitations.

References

[1] Randa Ben Ammar and Yassine Ben Ayed. Speech processing for early alzheimer disease diagnosis: machine learning based approach. In *2018 IEEE/ACS 15th International Conference on Computer Systems and Applications (AICCSA)*, pages 1–8. IEEE, 2018.

[2] Suzanne Bakken. Ai in health: keeping the human in the loop, 2023.

[3] Shahab S Band, Atefeh Yarahmadi, Chung-Chian Hsu, Meghdad Biyari, Mehdi Sookhak, Rasoul Ameri, Iman Dehzangi, Anthony Theodore Chronopoulos, and Huey-Wen Liang. Application of explainable artificial intelligence in medical health: A systematic review of interpretability methods. *Informatics in Medicine Unlocked*, 40:101286, 2023.

[4] Benjamin Barrera-Altuna, Daeun Lee, Zaima Zarnaz, Jinyoung Han, and Seungbae Kim. The interspeech 2024 taukadial challenge: Multilingual mild cognitive impairment detection with multimodal approach. *group*, 25:26, 2024.

[5] James T Becker, François Boiler, Oscar L Lopez, Judith Saxton, and Karen L McGonigle. The natural history of alzheimer's disease: description of study cohort and accuracy of diagnosis. *Archives of neurology*, 51(6):585–594, 1994.

[6] Mondher Bouazizi, Chuheng Zheng, Siyuan Yang, and Tomoaki Ohtsuki. Dementia detection from speech: What if language models are not the answer? *Information*, 15(1):2, 2023.

[7] Karol Chlasta and Krzysztof Wołk. Towards computer-based automated screening of dementia through spontaneous speech. *Frontiers in Psychology*, 11:623237, 2021.

[8] Hongyoon Choi, Kyong Hwan Jin, Alzheimer's Disease Neuroimaging Initiative, et al. Predicting cognitive decline with deep learning of brain metabolism and amyloid imaging. *Behavioural brain research*, 344:103–109, 2018.

[9] Natasha Clarke, Thomas R Barrick, and Peter Garrard. A comparison of connected speech tasks for detecting early alzheimer's disease and mild cognitive impairment using natural language processing and machine learning. *Frontiers in Computer Science*, 3:634360, 2021.

[10] Fuze Cong, Shibiao Xu, Li Guo, and Yinbing Tian. Caption-aware medical vqa via semantic focusing and progressive cross-modality comprehension. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 3569–3577, 2022.

[11] Patricia V Cooper. Discourse production and normal aging: Performance on oral picture description tasks. *Journal of gerontology*, 45(5):P210–P214, 1990.

[12] Kewen Ding, Madhu Chetty, Azadeh Noori Hoshyar, Tanusri Bhattacharya, and Britt Klein. Speech based detection of alzheimer's disease: a survey of ai techniques, datasets and challenges. *Artificial Intelligence Review*, 57(12):325, 2024.

[13] Chengfeng Dou, Ying Zhang, Yanyuan Chen, Zhi Jin, Wenpin Jiao, Haiyan Zhao, and Yu Huang. Detection, diagnosis, and explanation: A benchmark for chinese medial hallucination evaluation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4784–4794, 2024.

[14] Xinyang Feng, Zachary C Lipton, Jie Yang, Scott A Small, Frank A Provenzano, Alzheimer's Disease Neuroimaging Initiative, Frontotemporal Lobar Degeneration Neuroimaging Initiative, et al. Estimating brain age based on a uniform healthy population with deep learning and structural magnetic resonance imaging. *Neurobiology of aging*, 91:15–25, 2020.

[15] Ravindra Kumar Garg, Vijeth L Urs, Akshay Anand Agarwal, Sarvesh Kumar Chaudhary, Vimal Paliwal, and Sujita Kumar Kar. Exploring the role of chatgpt in patient care (diagnosis and treatment) and medical research: A systematic review. *Health Promotion Perspectives*, 13 (3):183, 2023.

[16] Harold Goodglass, Edith Kaplan, and Sandra Weintraub. *BDAE: The Boston diagnostic aphasia examination*. Lippincott Williams & Wilkins Philadelphia, PA, 2001.

[17] Jonathan Gratch, Ron Artstein, Gale Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David DeVault, Stacy Marsella, David Traum, Skip Rizzo, and Louis-Philippe Morency. The distress analysis interview corpus of human and computer interviews. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3123–3128, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA). URL https://aclanthology.org/L14-1421/.

[18] Kaisi Guan, Zhengfeng Lai, Yuchong Sun, Peng Zhang, Wei Liu, Kieran Liu, Meng Cao, and Ruihua Song. Etva: Evaluation of text-to-video alignment via fine-grained question generation and answering. *arXiv preprint arXiv:2503.16867*, 2025.

[19] Melisa Gumus, Morgan Koo, Christa M. Studzinski, Aparna Bhan, Jessica Robin, and Sandra E. Black. Linguistic changes in neurodegenerative diseases relate to clinical symptoms. *Frontiers in Neurology*, 15, March 2024. ISSN 1664-2295. doi: 10.3389/fneur.2024.1373341. URL https://www.frontiersin.org/journals/neurology/articles/10.3389/fneur.2024.1373341/full.

[20] Phillip Hamrick, Victoria Sanborn, Rachel Ostrand, and John Gunstad. Lexical speech features of spontaneous speech in older persons with and without cognitive impairment: reliability analysis. *JMIR aging*, 6(1):e46483, 2023.

[21] Lior Hason and Sri Krishnan. Spontaneous speech feature analysis for alzheimer's disease screening using a random forest classifier. *Frontiers in Digital Health*, 4:901419, 2022.

[22] Emtiaz Hussain, Mahmudul Hasan, Syed Zafrul Hassan, Tanzina Hassan Azmi, Md Anisur Rahman, and Mohammad Zavid Parvez. Deep learning based binary classification for alzheimer's disease detection using brain mri images. In *2020 15th IEEE Conference on Industrial Electronics and Applications (ICIEA)*, pages 1115–1120. IEEE, 2020.

[23] Zerin Jahan, Surbhi Bhatia Khan, and Mo Saraee. Early dementia detection with speech analysis and machine learning techniques. *Discover Sustainability*, 5(1):65, 2024.

[24] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12):1–38, 2023.

[25] Edith Kaplan. *Boston diagnostic aphasia examination booklet*. Lea & Febiger Philadelphia, PA, 1983.

[26] M Rupesh Kumar, Susmitha Vekkot, S Lalitha, Deepa Gupta, Varasiddhi Jayasuryaa Govindraj, Kamran Shaukat, Yousef Ajami Alotaibi, and Mohammed Zakariah. Dementia detection from speech using machine learning and deep learning architectures. *Sensors*, 22(23):9311, 2022.

[27] Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. Biomistral: A collection of open-source pretrained large language models for medical domains. *arXiv preprint arXiv:2402.10373*, 2024.

[28] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36: 28541–28564, 2023.

[29] Yunyi Liu, Zhanyu Wang, Dong Xu, and Luping Zhou. Q2atransformer: Improving medical vqa via an answer querying decoder. In *International conference on information processing in medical imaging*, pages 445–456. Springer, 2023.

[30] Saturnino Luz, Fasih Haider, Sofia de la Fuente Garcia, Davida Fromm, and Brian MacWhinney. Alzheimer's dementia recognition through spontaneous speech, 2021.

[31] Saturnino Luz, Fasih Haider, Davida Fromm, Ioulietta Lazarou, Ioannis Kompatsiaris, and Brian MacWhinney. Multilingual alzheimer's dementia recognition through spontaneous speech: a signal processing grand challenge. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–2. IEEE, 2023.

[32] Saturnino Luz, Sofia De La Fuente Garcia, Fasih Haider, Davida Fromm, Brian MacWhinney, Alyssa Lanzi, Ya-Ning Chang, Chia-Ju Chou, and Yi-Chien Liu. Connected speech-based cognitive assessment in chinese and english. *arXiv preprint arXiv:2406.10272*, 2024.

[33] Saturnino Luz, Fasih Haider, Davida Fromm, Ioulietta Lazarou, Ioannis Kompatsiaris, and Brian MacWhinney. An overview of the adress-m signal processing grand challenge on multilingual alzheimer's dementia recognition through spontaneous speech. *IEEE open journal of signal processing*, 5:738–749, 2024.

[34] Pranav Mahajan and Veeky Baths. Acoustic and language based deep learning approaches for alzheimer's dementia detection from spontaneous speech. *Frontiers in Aging Neuroscience*, 13: 623607, 2021.

[35] Amit Meghanani, Chandran Savithri Anoop, and Angarai Ganesan Ramakrishnan. Recognition of alzheimer's dementia from the transcriptions of spontaneous speech using fasttext and cnn models. *Frontiers in Computer Science*, 3:624558, 2021.

[36] Munib Mesinovic, Peter Watkinson, and Tingting Zhu. Explainability in the age of large language models for healthcare. *Communications Engineering*, 4(1):128, 2025.

[37] Shamila Nasreen, Morteza Rohanian, Julian Hough, and Matthew Purver. Alzheimer's dementia recognition from spontaneous speech using disfluency and interactional features. *Frontiers in Computer Science*, 3:640669, 2021.

[38] Linda E Nicholas and Robert H Brookshire. A system for quantifying the informativeness and efficiency of the connected speech of adults with aphasia. *Journal of Speech, Language, and Hearing Research*, 36(2):338–350, 1993.

[39] Emma Nichols, Jaimie D Steinmetz, Stein Emil Vollset, Kai Fukutaki, Julian Chalek, Foad Abd-Allah, Amir Abdoli, Ahmed Abualhasan, Eman Abu-Gharbieh, Tayyaba Tayyaba Akram, et al. Estimation of the global prevalence of dementia in 2019 and forecasted prevalence in 2050: an analysis for the global burden of disease study 2019. *The Lancet Public Health*, 7(2): e105–e125, 2022.

[40] Marko Niemelä, Mikaela von Bonsdorff, Sami Äyrämö, and Tommi Kärkkäinen. Dementia classification using acoustic speech and feature selection. *arXiv preprint arXiv:2502.03484*, 2025.

[41] Rachel Ostrand and John Gunstad. Using automatic assessment of speech production to predict current and future cognitive function in older adults. *Journal of Geriatric Psychiatry and Neurology*, 34(5):357–369, September 2021. ISSN 0891-9887. doi: 10.1177/0891988720933358. URL `https://doi.org/10.1177/0891988720933358`.

[42] Dong Jin Park, Min Woo Park, Homin Lee, Young-Jin Kim, Yeongsic Kim, and Young Hoon Park. Development of machine learning model for diagnostic disease prediction based on laboratory tests. *Scientific reports*, 11(1):7567, 2021.

[43] Zhao Pei, Zhiyang Wan, Yanning Zhang, Miao Wang, Chengcai Leng, and Yee-Hong Yang. Multi-scale attention-based pseudo-3d convolution neural network for alzheimer's disease diagnosis using structural mri. *pattern recognition*, 131:108825, 2022.

[44] Paula Andrea Pérez-Toro, Tomás Arias-Vergara, Philipp Klumpp, Tobias Weise, Maria Schuster, Elmar Noeth, Juan Rafael Orozco-Arroyave, and Andreas Maier. Multilingual speech and language analysis for the assessment of mild cognitive impairment: Outcomes from the taukadial challenge. In *Proc. Interspeech 2024*, pages 982–986, 2024.

[45] Aurélie Pistono, Mélanie Jucla, Emmanuel J. Barbeau, Laure Saint-Aubert, Béatrice Lemesle, Benjamin Calvet, Barbara Köpke, Michèle Puel, and Jérémie Pariente. Pauses during autobiographical discourse reflect episodic memory processes in early alzheimer's disease. *Journal of Alzheimer's disease: JAD*, 50(3):687–698, 2016. ISSN 1875-8908. doi: 10.3233/JAD-150408.

[46] Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M Dai, Nissan Hajaj, Michaela Hardt, Peter J Liu, Xiaobing Liu, Jake Marcus, Mimi Sun, et al. Scalable and accurate deep learning with electronic health records. *NPJ digital medicine*, 1(1):18, 2018.

[47] N Rockwell. Going and coming (oil on canvas). *Norman Rockwell Art Collection Trust, Indianapolis, IN, United States*, 1947.

[48] Yamanki Santander-Cruz, Sebastián Salazar-Colores, Wilfrido Jacobo Paredes-García, Humberto Guendulain-Arenas, and Saúl Tovar-Arriaga. Semantic feature extraction using sbert for dementia detection. *Brain sciences*, 12(2):270, 2022.

[49] Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. Medgemma technical report. *arXiv preprint arXiv:2507.05201*, 2025.

[50] Zehra Shah, Jeffrey Sawalha, Mashrura Tasnim, Shi-ang Qi, Eleni Stroulia, and Russell Greiner. Learning language and acoustic models for identifying alzheimer's dementia from speech. *Frontiers in Computer Science*, 3:624659, 2021.

[51] Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, et al. Toward expert-level medical question answering with large language models. *Nature Medicine*, 31(3):943–950, 2025.

[52] Xuegang Song, Feng Zhou, Alejandro F Frangi, Jiuwen Cao, Xiaohua Xiao, Yi Lei, Tianfu Wang, and Baiying Lei. Multicenter and multichannel pooling gcn for early ad diagnosis based on dual-modality fused brain network. *IEEE Transactions on Medical Imaging*, 42(2):354–367, 2022.

[53] Muhammad Shehram Shah Syed, Zafi Sherhan Syed, Margaret Lech, and Elena Pirogova. Automated screening for alzheimer's dementia through spontaneous speech. In *Interspeech*, volume 2020, pages 2222–6, 2020.

[54] Khandaker Mohammad Mohi Uddin, Mir Jafikul Alam, Md Ashraf Uddin, and Sunil Aryal. A novel approach utilizing machine learning for the early diagnosis of alzheimer's disease. *Biomedical Materials & Devices*, 1(2):882–898, 2023.

[55] Harriet Louise Walker, Shahi Ghani, Christoph Kuemmerli, Christian Andreas Nebiker, Beat Peter Müller, Dimitri Aristotle Raptis, and Sebastian Manuel Staubli. Reliability of medical information provided by chatgpt: assessment against clinical guidelines and patient information quality instrument. *Journal of Medical Internet Research*, 25:e47479, 2023.

[56] Xidong Wang, Nuo Chen, Junyin Chen, Yidong Wang, Guorui Zhen, Chunxian Zhang, Xiangbo Wu, Yan Hu, Anningzhe Gao, Xiang Wan, et al. Apollo: A lightweight multilingual medical llm towards democratizing medical ai to 6b people. *arXiv preprint arXiv:2403.03640*, 2024.

[57] Qianqian Xie, Qingyu Chen, Aokun Chen, Cheng Peng, Yan Hu, Fongci Lin, Xueqing Peng, Jimin Huang, Jeffrey Zhang, Vipina Keloth, et al. Me-llama: Foundation large language models for medical applications. *Research square*, pages rs–3, 2024.

[58] Anthony Yeung, Andrea Iaboni, Elizabeth Rochon, Monica Lavoie, Calvin Santiago, Maria Yancheva, Jekaterina Novikova, Mengdan Xu, Jessica Robin, Liam D Kaufman, et al. Correlating natural language processing and automated speech analysis with clinician assessment to quantify speech-language changes in mild cognitive impairment and alzheimer's dementia. *Alzheimer's research & therapy*, 13(1):109, 2021.

[59] Han Yuan, Lican Kang, Yong Li, and Zhenqian Fan. Human-in-the-loop machine learning for healthcare: current progress and future opportunities in electronic health records. *Medicine Advances*, 2(3):318–322, 2024.

[60] Jiahong Yuan, Yuchen Bian, Xingyu Cai, Jiaji Huang, Zheng Ye, and Kenneth Church. Disfluencies and fine-tuning pre-trained language models for detection of alzheimer's disease. In *Interspeech*, volume 2020, pages 2162–6, 2020.

[61] Youxiang Zhu, Xiaohui Liang, John A Batsis, and Robert M Roth. Exploring deep transfer learning techniques for alzheimer's dementia detection. *Frontiers in computer science*, 3: 624683, 2021.

[62] Youxiang Zhu, Abdelrahman Obyat, Xiaohui Liang, John A Batsis, and Robert M Roth. Wavbert: Exploiting semantic and non-semantic speech using wav2vec and bert for dementia detection. In *Interspeech*, volume 2021, page 3790, 2021.

[63] Youxiang Zhu, Nana Lin, Kiran Balivada, Daniel Haehn, and Xiaohui Liang. Adversarial text generation using large language models for dementia detection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21918–21933, 2024.

A Alzheimer's/Dementia Datasets

Picture description datasets are widely used in ML research for dementia detection. Participants are asked to describe reference/stimuli images such as the well-known "Cookie Theft" picture, or in some cases multiple images like "Cat Rescue" and "Coming and Going" (see Fig. 2). Table 1 summarizes key datasets and subject counts. Among these, ADReSSo and TAUKADIAL have become especially popular datasets, with the former derived from the Pitt corpus.

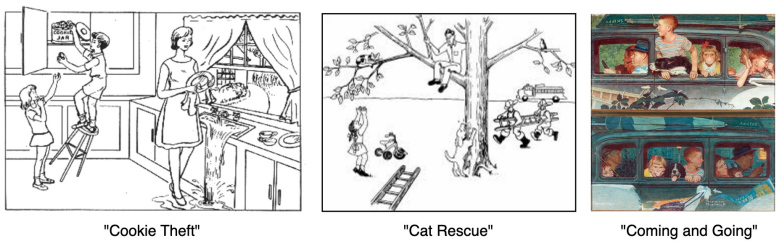

Figure 2: Stimuli/reference images used in the ADReSS [30] and TAUKADIAL [32] datasets: "Cookie Theft" [25], "Cat Rescue" [38], and "Coming and Going" [47]. ADReSS includes only the Cookie Theft task, while TAUKADIAL requires each participant to describe all three images.

Dataset	Language	# of Subjects			
		NC	MCI	AD	Total
Pitt [5]	English	243	-	306	549
ADReSSo [30]	English	115	-	122	237
ADReSS-M [31]	English	115	-	122	237
	Greek	24	-	22	46
TAUKADIAL [32]	English	31	51	-	82
	Chinese	43	44	-	87

Table 1: Overview of existing dementia classification datasets for the picture description task. ADReSSo and TAUKADIAL are widely used datasets [26, 21, 34, 50]. Pitt Corpus and ADReSS rely solely on the Cookie Theft image, while TAUKADIAL employs three images (see Fig. 2), providing three transcripts per subject. ADReSS is a curated subset of the Pitt Corpus. NC: Normal Cognition, MCI: Mild Cognitive Impairment, AD: Alzheimer's Disease.

ADReSSo. The ADReSSo [30] dataset contains speech samples collected from patients diagnosed with *Probable Alzheimer's Disease (AD)* and *non-AD controls*. It captures both conversational and task-oriented speech, with a focus on distinguishing acoustic, lexical, semantic, and syntactic markers of Alzheimer's progression. The dataset provides several hours of audio along with manual transcripts, allowing multimodal dementia detection. Fig. 3 shows an example for the two labels, Probable Alzheimer's Disease (AD) and control.

TAUKADIAL. The TAUKADIAL [32] dataset consists of speech samples from individuals with *Mild Cognitive Impairment (MCI)* and *Normal Cognition* (NC) controls, in both English and Chinese. Participants completed the picture description task by answering questions about three stimuli images (see Fig. 2), providing three transcripts per subject. Fig. 4 shows example transcripts for the two labels, MCI and normal control.

Figure 3: Example samples from the ADReSSo [30] dataset, showing a **probable AD** patient (top) and a cognitively normal **control** (bottom). Speaker roles (patient vs. clinician) are not distinguished in the dataset; they are presented here for readability.

B Fine-Grain Feature Probing for Interpretability

As discussed in Sec. 3, we argue that dementia detection should be repositioned from a direct classification problem to a structured cognitive reasoning process. The picture description task is well-suited for fine-grained analysis, since transcripts can be broken down into clinically meaningful cues such as word repetitions, story coherence, image alignment, narrative flow, etc. Fig. 5 describes some of these fine-grain prompting categories that can be used to assess model cognition. Fig. 6 shows how structured prompting can extract these features explicitly, in contrast to vague outputs from zero-shot prompting. Standardizing such fine-grained extractions, and providing ground-truth annotations from clinicians, would allow future models to reason through intermediate features before arriving at a diagnosis. This process mirrors clinical assessment more closely and offers interpretability by showing intermediate reasoning for features.

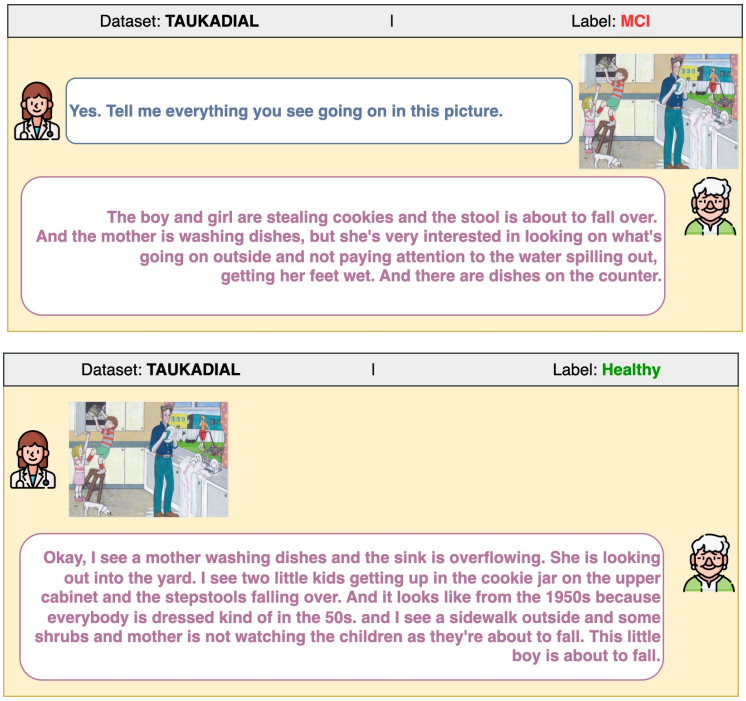

Figure 4: Example samples from the TAUKADIAL [32] dataset, showing an **MCI** patient (top) and a cognitively **normal** control (bottom). Transcripts are ASR-generated; speaker roles are not always distinguished, and clinician speech may be absent.

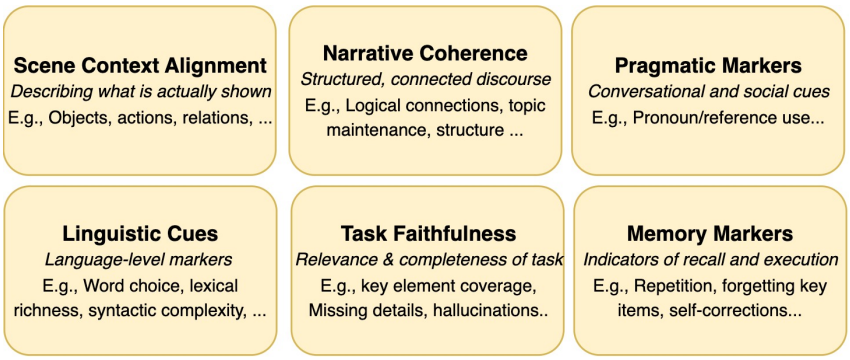

Figure 5: Examples of proposed structured probing tasks applied to dementia models, showing how varying prompt types can reveal attention to fine grain linguistic features and aid explainability.

C Limitations and Ethical Considerations

Dataset sizes and lack of diversity. While picture description datasets have enabled significant progress, their small scale and limited demographic diversity restrict generalizability. Most corpora are dominated by English-speaking participants or a small handful of languages, and subject pools underrepresent age, and socioeconomic variation, potentially embedding systematic biases into downstream models. Individual recordings are often very short, offering only brief snapshots of patient speech that limit the richness of cognitive assessment. Finally, many datasets reflect institutional biases in recruitment and annotation practices, which can influence both the linguistic style of collected samples and the diagnostic framing provided to models.

Privacy and consent issues. The sensitive nature of speech data also raises privacy and consent concerns. Audio might contain personal information that, if misused, could expose individuals beyond

Figure 6: **Top**: patient transcript to be analyzed for dementia. **Middle**: zero-shot prompting with a general-purpose MLLM (ChatGPT-5) and a medical MLLM (MedGemma-4B) produces incorrect diagnoses with vague explanations. **Bottom**: structured prompting elicits explicit feature extraction and transcript analysis, which can be used in a manner similar to clinical notes to support more fine-grained and interpretable diagnosis.

their cognitive status. Responsible use requires strict supervision for anonymization, controlled sharing, and adherence to medical data governance frameworks.

LLMs and Hallucinations. Reliance on LLMs introduces risks of hallucination and misplaced trust. Current models may generate reasonable but clinically invalid explanations. Without grounding in clinician-validated features, hallucinated outputs could reinforce overconfidence and inaccurate information in automated dementia detection models. Our position emphasizes that future work must treat LLMs as tools for structured analysis, not as substitutes for clinical judgment, ensuring interpretability and alignment with real-world diagnostic reasoning.