On Large Foundation Models and Alzheimer's Disease Detection

Chuyuan Li[1], Giuseppe Carenini[1], Thalia Field[2]
[1] Department of Computer Science
[2] Vancouver Stroke Program and Division of Neurology
The University of British Columbia
V6T 1Z4, Vancouver, BC, Canada
`{chuyuan.li, thalia.field}@ubc.ca, carenini@cs.ubc.ca`

Abstract

Large Foundation Models such as Llama and LLaVA have displayed incredible capabilities in a wide range of domains and tasks. However, it is unclear whether these models match specialist capabilities without special training or fine-tuning. In this paper, we investigate the innate ability of foundation models as neurodegenerative disease specialists, particularly for detecting the Alzheimer's Disease. Precisely, we use a language model, Llama-3.1, and a visual language model, Llama3-LLaVA-NeXT, to detect language specificity between Alzheimer's Disease patients and healthy controls through a well-known Picture Description task. Results show that Llama is comparable to supervised classifiers, while LLaVA, despite its additional "vision", lags behind.

1 Introduction

Large Foundation Models such as Llama have demonstrated surprising capabilities in the field of Natural Language Processing (NLP). Recent work seems to indicate that these generalist models can be used in specialized domains, such as clinical medicine, with proprietary Large Language Models (LLMs) such as GPT-4 achieving impressive performance on professional benchmarks in health domain (Bubeck et al., 2023; Cui et al., 2024; Belyaeva et al., 2023; Jin et al., 2024). Other work, however, suggests that GPT-4 does not outperform traditional AI tools and cannot replace them at current stage (Wang et al., 2023).

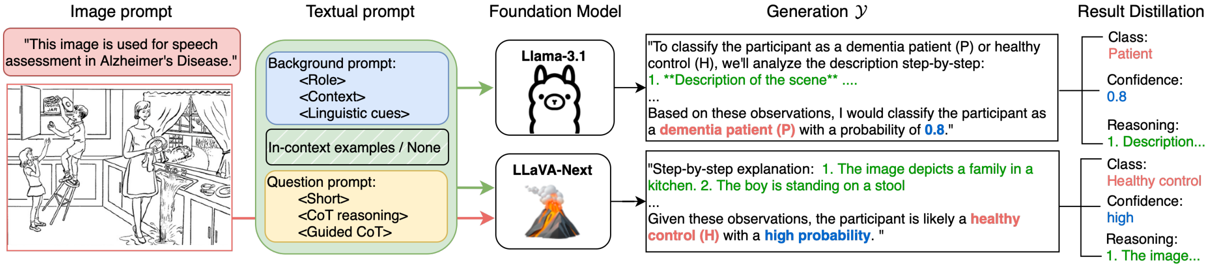
The healthcare sector often prefers open LLMs that can be deployed in local environments, especially since relying on third-party commercial LLMs is not always feasible due to concerns about traceability, privacy, and security. Taking into account the complexity of real-world applications, in this paper, we explore the use of small (e.g., less than 10B), cost-effective open-source LLMs for Alzheimer's Disease detection.

Alzheimer's Disease (AD) is a insidious progressive neurodegenerative disease resulting in impaired cognition and dementia, and eventual death (Scharre, 2019). Since there is no effective cure for dementia, early intervention is essential. Modern deep learning approaches utilize data from various modalities, such as speech (Berube et al., 2019; Ilias and Askounis, 2022), eye-tracking (Sriram et al., 2023; Sheng et al., 2022), facial (Chou et al., 2025), and neuroimaging (Sarraf et al., 2023). However, some modalities require invasive and costly screening tools. In contrast, language data is easy to collect—a speech recording takes no more than 10 minutes—and involves no invasive procedure, making it an ideal resource for early disease detection.

Picture description, such as the one shown in Figure 1, has been widely used to capture deficits or abnormalities in **language** (Yorkston and Beukelman, 1980). Over the years, clinicians have assessed a variety of measures, such as grammaticality (Ash and Grossman, 2015), vocabulary (Forbes-McKay and Venneri, 2005), frequency of noun-verb ratio (Bird et al., 2000), and the percentage and change of information units (IUs), e.g., "mom", "girl" (Giles et al., 1996; Bouazizi et al., 2023). In the pre-LLM era, NLP practitioners manually craft linguistic features and use machine learning algorithms to train supervised models (Fraser et al., 2016, 2019; Barral et al., 2020; Jang et al., 2021). While these experiments yield promising results, they require the collection of training data, a time-consuming and labor-intensive process. Additionally, the variability in datasets and recording conditions also makes it challenging for supervised models to generalize well (Favaro et al., 2024).

With the advent of LLMs, NLP has shifted from developing task-specific representations and architectures to using task-agnostic foundation models (Radford et al., 2019; Brown, 2020), which are pretrained on vast, cross-disciplinary data. These models not only streamline the process but also

Proceedings of the Second Workshop on Patient-Oriented Language Processing (CL4Health), pages 158–168
May 4, 2025 ©2025 Association for Computational Linguistics



"To classify the participant as a dementia patient (P) or healthy control (H), we'll analyze the description step-by-step:
1. **Description of the scene** ....
...
Based on these observations, I would classify the participant as a dementia patient (P) with a probability of 0.8."

"Step-by-step explanation: 1. The image depicts a family in a kitchen. 2. The boy is standing on a stool
...
Given these observations, the participant is likely a healthy control (H) with a high probability. "

Figure 1: The roadmap of our approach. Textual prompt are provided to language model Llama; image and textual prompts are provided to vision-language model LLaVA-NeXT. We extract class, probability, and analysis from $\mathcal{Y}$.

offer interpretable explanations, providing clinical doctors with valuable insights into their reasoning (Perlis, 2023; Nori et al., 2023a,b). Our approach leverages the powerful open-source LLM, Llama-3.1-8B (Dubey et al., 2024), and explores its potential for AD detection through carefully designed zero-shot and few-shot prompting strategies. At the time of our experiments, Llama-3.1 models offered state-of-the-art performance in the open-source LLM landscape.

Given that our task involves describing a **picture** using language, a Vision-Language Model (VLM) should offer a clear advantage. We choose a VLM from the LLaVA family (Liu et al., 2024b), a pioneering work in visual instruction tuning, while also considering its base language model and comparable size. LLaVA is pretrained on image-caption data and designed to provide detailed descriptions (e.g., position of objects in an image) and perform complex reasoning (e.g., "*What is unusual in an image, and explain.*"). These capabilities align well with our experimental setup. Based on these considerations, we use Llama3-LLaVA-NeXT (Liu et al., 2024a), one of the latest LLaVA models, which is built on Llama-3-8B-Instruct and integrates a vision encoder for image processing.

Our experimental results show that Llama-3.1-8B model can match or even surpass traditional supervised methods with minimal supervision. The key factor is the effective combination of Background and Question prompts. LLaVA, on the other hand, is not yet suited for this task, as intriguingly it may itself exhibit symptoms akin to "neurodegenerative" issues.

2 The Power of Prompting

Prompt engineering is a popular and effective way for using LLMs without altering their parameters. Empirical studies have shown that a model's performance on specific tasks can be significantly affected by the prompt, often in surprising ways (Feng et al., 2024; Sivarajkumar et al., 2024; Salinas and Morstatter, 2024; Sclar et al., 2024). For instance, by adding "*Let's think step by step*" can greatly improve model performance (Kojima et al., 2022). To date, there is not yet a consensus on how to formulate the most effective prompts for a specific task.

To unleash the inner specialist capabilities of LLMs and gain a better understanding of the crucial components in a prompt, we design our prompts in a systematic way. For language-only model, we divide our prompt into three parts: (1) Background prompt, (2) Question prompt, and (3) Example prompt. For vision-language model, we include the image and prepend the cue phrase, "*This image is used for speech assessment in Alzheimer's Disease.*", to the beginning of the textual prompt, as shown in Figure 1. The textual input for VLM is the same as the language-only model. Precisely:

(1) **Background Prompt** aims to place LLMs in a specific knowledge graph where the information is closely related to the target domain. For instance, prompts starting with "*You are an intelligent AI assistant*" or "*You are an expert in clinical NLP*" use Persona pattern as guiding cues (Sivarajkumar et al., 2024). In our experiments, we test three cue phrases, including `Role`—"*You are a medical expert in Alzheimer's Disease*", `Context`— a brief introduction of the Cookie Theft picture description task, and `Linguistic`—clinical observations of linguistic features from AD and healthy controls.

(2) **Question Prompt** directs LLMs to produce desirable output, whether text generation, classification, or resolution. It has been shown that by simply adding "*Let's think step by step*" before each answer, LLMs can become decent zero-shot reasoners (Kojima et al., 2022). Here, we compare short answer and Chain of Thought (CoT) prompting answer in the question prompt. `Short answer`

simply asks the LLM to predict a class (i.e., normal *vs.* patient) without any explanation, while `CoT answer` asks the model to first analyze step by step and then give an answer. We also offer a `Guided CoT` (`G. CoT`) version to direct the model to reason from specific linguistic perspectives, such as "vocabulary richness" and "syntactic complexity".

Pre-trained language models are often inherently calibrated to different extents (Jiang et al., 2021; Liang et al., 2023), with token probabilities might be employed off-the-shelf. Since they could verbalize confidence scores (Tian et al., 2023), we use the cue phrase "*Give a prediction with a probability*" to directly ask for prediction probabilities. In preliminary experiments, we found that using this cue phrase yielded better results than omitting it.

(3) **Example Prompt** is positioned between Background Prompt and Question Prompt. It aims to examine whether In-Context Learning (ICL) with demonstrations further improves LLMs' performance in comparison to zero-shot prompting. Practically, we employ *fixed* and *dynamic few-shot* ICL. The fixed examples are selected randomly in the held-out set to be broadly representative and relevant to a wide distribution of text examples. The dynamic examples are instead chosen with a kNN-based approach ($k = 2$) (Nori et al., 2023b), where we embed all texts in the held-out set using OpenAI's latest embedding model `text-embedding-3-small`[1]. For each test example, we identify its nearest neighbor in both Patient and Control classes by computing their cosine similarity scores. In our experiments, we use a small k value, i.e., one positive and one negative examples.

We combine different types of Background and Question prompts in both zero-shot and 2-shot ICL. For instance, (`Role`; `Short`) uses `Role` in the background prompt and requires a short answer in the question prompt; (`Context+Role+Ling`; `G. CoT`) uses a combination of all background prompts and a guided CoT answer. Detailed prompting templates are provided in Appendix 5.

3 Data and Experiment Settings

Dataset. We use the dataset from Jang et al. (2021), which comprises 63 patients recruited from a specialty memory clinic and 67 healthy controls from the community. Patients are either diagnosed with Alzheimer's Disease (AD) or exhibiting initial symptoms of Mild Cognitive Impairments (MCI), potentially progressing to AD. Participants completed four tasks—pupil calibration, picture description, paragraph reading, and memory recall—during which both language and eye movement data were collected. In this study, we focus on the picture description task. Demographic and clinical data is provided in Table 1.

Group	#	Age	Gender	MoCA
Patient	63	72 ± 9	31M / 34F	18 ± 7
Control	67	62 ± 15	22M / 45F	27 ± 3

Table 1: Dataset demographic and clinical statistics. MoCA stands for Montreal Cognitive Assessment score.

Data Processing. The original speech data is transcribed and timestamped using WhisperX (Bain et al., 2023). Following automatic transcriptions, a human transcriber manually verified each transcript for word spelling and speaker diarization accuracy. Task instructions from the instructor were removed to include only participant speech.

Hyper-Parameters. We use the 8B checkpoint of Llama-3.1 and LLaVA-NeXT from Huggingface (Wolf et al., 2020). We use a low temperature (0.1) and set top_k sampling to 50. The maximum new tokens are 16 and 512 for `Short answer` and `CoT answer`, respectively. To investigate the potential *non-determinism* of LLMs (Ouyang et al., 2023; Song et al., 2024), each prompt configuration (e.g., (`Role`; `Short answer`)) is executed on two different servers with three runs per server. We report micro-averaged scores with standard deviation over 6 runs per setup.

Supervised Classifiers and Metrics. We compare with three classic supervised algorithms: Logistic Regression (LR), Random Forest (RF), and Gaussian Naive Bayes (GNB). We follow the feature extraction process outlined in Jang et al. (2021) and split the dataset into ten folds for cross-validation. Appendix 5 provides details.

In all experiments, we report three metrics: (1) Area Under the *Receiver Operating Characteristic (ROC)* Curve (**AUC**): the ability to distinguish between Patient and Control under different thresholds; (2) **Sensitivity**: the True Positive rate for Patient detection, and (3) **Specificity**: the True Negative rate for Control detection. The main measure is the AUC score.

---

[1] `https://platform.openai.com/docs/guides/embeddings`.

Background	Question	AUC	Sensitivity	Specificity
Role	Short	60.3 ± 1.1	**96.4** ± 0.8	11.5 ± 0.8
	CoT	65.8 ± 0.5	91.13 ± 1.1	24.6 ± 2.5
	G. CoT	**70.9** ± 0.4	84.7 ± 1.1	**35.4** ± 2.1
Context	Short	69.4 ± 1.5	35.9 ± 2.0	**93.5** ± 1.4
	CoT	68.9 ± 0.6	50.8 ± 1.1	73.9 ± 2.1
	G. CoT	**74.3** ± 1.1	**69.4** ± 2.2	69.3 ± 0.0
Context	Short	71.6 ± 0.5	**72.6** ± 0.0	69.6 ± 1.4
+Role	CoT	72.9 ± 3.8	70.2 ± 3.4	70.8 ± 4.3
+Ling	G. CoT	**76.1** ± 2.0	71.8 ± 3.4	**73.9** ± 2.1
Supervised Classifiers				
GNB	-	72.8 ± 2.2	64.1 ± 2.2	66.5 ± 3.5
LR	-	73.2 ± 1.7	68.5 ± 3.8	70.2 ± 1.6
RF	-	75.2 ± 3.1	67.7 ± 4.6	73.1 ± 3.6

Table 2: Top: zero-shot with different background and question prompts. Scores are averaged across 6 runs. Best score in each sub-section is in **bold**. Bottom: performances using supervised classifiers.

4 Results with Foundation Models

Impact of Background and Question Prompts on Zero-shot Prompting. We present zero-shot results in Table 2. When using the `Role` pattern in the Background prompt, Llama is highly sensitive in detecting Patients (90%), but much less so for Controls. In contrast, when the model is provided only with the `Context` of the picture description task, it predicts Controls more accurately. This suggests that different background prompts shift the model's threshold for identifying *Patient language* in distinct ways. Combining different background prompts (`Context+Role+Ling`) provides more complete information, enabling the LLM to retrieve the most relevant knowledge and deliver optimal performance. With a 76% AUC score, it matches and even surpasses supervised classifiers (GNB 73%, LR 73%, RF 75%). In the Question prompt, `CoT` significantly enhances model performance compared to `Short` answers, and this improvement is consistent across various background settings. Interestingly, we find that simply asking the model to analyze before making predictions helps the LLM to show moderation in its decision making, as evidenced by a more balanced Sensitivity and Specificity rate.

Few-shot *vs.* Zero-shot Prompting. In few-shot in-context learning, we use the same Background and Question prompts as in the zero-shot setting, but add input-output pairs as demonstration in the Example Prompt. Table 3 demonstrates that few-shot prompting consistently enhances AUC scores, particularly when the background prompt lacks sufficient task information, as seen with the `Role` background. Remarkably, we also find that zero-shot prompting can be highly effective in some cases, even surpassing few-shot prompting. This occurs when the background is complete and a `CoT answer` is employed—a trend also observed in other clinical NLP tasks (Sivarajkumar et al., 2024). In most prompt settings, random few-shot outperforms *k*NN, suggesting that a more general and representative set of examples leads to better performance than semantically close ones. Similar observations are made by Nori et al. (2023a). However, it comes with the trade-off of greater fluctuations and less consistent Sensitivity and Specificity rates.

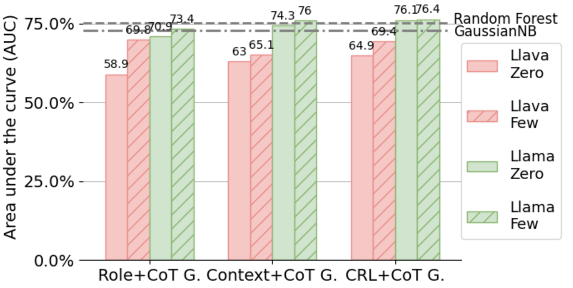


Figure 2: LLaVA and LLama performances on different Background (`Role`, `Context`, `Context+Role+Linguistic`) and Answer prompt (`G. CoT`) settings.

Vision Language Model *vs.* Language-only Model. While we were expecting VLMs to outperform pure LLMs due to their ability to process the picture, Figure 2 reveals that LLaVA significantly underperforms Llama on this task, with lower AUC scores up to 10% across various prompting, in both zero-shot and few-shot setups. Specifically, we observe some *extreme* predictions in zero-shot prompting, where LLaVA exclusively predicts either Patient or Control, a behavior never observed with Llama. While few-shot prompting brings some improvement, the model remains *biased* toward producing high Specificity or Sensitivity scores (see detailed scores in Appendix 5). We also note some anecdotal observations regarding LLaVA's reasoning. For instance, when asked to analyze step by step, instead of reasoning from a linguistic perspective, LLaVA simply lists the objects in the image, such as: "*1. The image shows a family. 2. The boy is standing. 3. Water is flowing.*" One plausible reason is that VLMs are not trained to capture subtle linguistic nuances as LLMs are,

		Random Few-shot			*k*NN Few-shot			Zero-shot		
B prompt	Q prompt	AUC	Sensitivity	Specificity	AUC	Sensitivity	Specificity	AUC	Sensitivity	Specificity
Role	Short	69.5 ± 7.5	74.4 ± 8.1	55.9 ± 14.8	64.5 ± 0.2	84.13 ± 0.0	38.8 ± 2.1	60.3 ± 1.1	96.4 ± 0.8	11.5 ± 0.8
	CoT	71.3 ± 3.4	75.4 ± 9.0	50.0 ± 17.9	70.7 ± 2.1	94.4 ± 1.1	23.1 ± 1.0	65.8 ± 0.5	91.13 ± 1.1	24.6 ± 2.5
	G. CoT	73.4 ± 7.0	74.9 ± 8.0	60.8 ± 12.1	72.9 ± 0.7	87.2 ± 2.4	33.6 ± 1.0	70.9 ± 0.4	84.7 ± 1.1	35.4 ± 2.1
Context	Short	69.2 ± 9.6	77.2 ± 12.8	50.0 ± 18.5	68.3 ± 0.9	61.1 ± 1.1	69.4 ± 1.0	69.4 ± 1.5	35.9 ± 2.0	93.5 ± 1.4
	CoT	65.0 ± 5.1	59.1 ± 12.4	65.3 ± 14.3	68.1 ± 2.7	61.1 ± 3.4	71.6 ± 0.0	68.9 ± 0.6	50.8 ± 1.1	73.9 ± 2.1
	G. CoT	76.0 ± 4.1	64.6 ± 4.8	75.9 ± 8.9	74.3 ± 2.1	80.2 ± 1.1	64.2 ± 0.0	74.3 ± 1.1	69.4 ± 2.2	69.3 ± 0.0
Context	Short	71.9 ± 3.7	58.1 ± 15.0	79.8 ± 10.5	71.1 ± 2.3	65.9 ± 1.1	71.6 ± 0.0	71.6 ± 0.5	72.6 ± 0.0	69.6 ± 1.4
+Role	CoT	71.9 ± 3.6	61.4 ± 11.0	78.4 ± 7.2	74.9 ± 0.1	77.0 ± 3.3	64.2 ± 0.0	72.9 ± 3.8	70.2 ± 3.4	70.8 ± 4.3
+Ling	G. CoT	76.4 ± 2.9	71.7 ± 7.6	75.8 ± 8.4	74.8 ± 1.3	83.2 ± 1.3	46.3 ± 0.0	76.1 ± 2.0	71.8 ± 3.4	73.9 ± 2.1

Table 3: Random, *k*NN few-shot, and zero-shot prompting results with Llama. Random few-shot results are averaged from three sampling. Best AUC in random, *k*NN, and zero-shot sections is highlighted ; second best is underlined.

as they are primarily pretrained on image-text pairs to recognize objects in images.

Recent studies reveal that VLMs are prone to heavy *hallucinations* and can be easily misled by deceptive prompts (Qian et al., 2024; Zhang et al., 2024). To explore the "neuro-cognitive status" of LLaVA, we prompt the model to perform the picture description task, similar to the human participants. We then apply trained classifiers (GNB, LR) to analyze LLaVA's generated speech, with details in Appendix 5. Unsurprisingly, the supervised classifiers consistently predict LLaVA as a Patient with high probability: GNB$>90\%$, LR$>80\%$. Since LLaVA is unable to generate *normal* speech during the picture description task, it is not surprising that its predictions are not reliable.

5 Conclusion and Open Questions

In this paper, we explore the potential of foundation models for lightweight use in Alzheimer's Disease detection via Picture Description task. Using appropriate prompting strategies, we find that LLMs can be activated to exhibit specialist capabilities even in a no-data scenario, achieving performance comparable to supervised classifiers, while providing clear and insightful reasoning. VLMs, however, are not yet suited for complex language reasoning tasks, as they may themselves exhibit symptoms akin to "neurodegenerative" issues. In the near future, we plan to expand our methods from picture description narratives to conversational interactions such as semi-structured conversations (Goodkind et al., 2018), and other healthcare-related disease detection (Li et al., 2022).

Several open questions remain worth exploring, such as alternative methods for demonstration example selection and whether a best example-pair exists that could boost LLMs' performance. For multimodal models, a key challenge is how to effectively enhance their compositional capabilities, thus enabling them to process different information in a more intelligent and integrated way.

Limitations

Most LLMs do not consistently produce the same output due to the inherent randomness in their parameter initialization. To address this and ensure more robust results, we ran each setup at least three times on different servers. In most cases (zero-shot and *k*NN few-shot), the variations were minimal, reinforcing our conclusions about different prompting strategies. However, we found that the relative performance of prompt variations could vary significantly when using random few-shot prompting.

We evaluate two model variations, Llama and LLaVA, both of which are leading models of LLMs and VLMs, respectively. However, further research is needed to understand how different language models, architectures, and datasets may impact the sensitivity of prompt variations on this task.

Ethical Considerations

The dataset we use for this paper comes from the CANARY project at University of British Columbia (UBC), which was approved by the UBC clinical research ethics board (H17-02803-A036). During the experiments, we ensure that no private information—such as participants' health, clinical, or demographic data—is disclosed. This is a main reason for us exclusively testing with open-source language models.

Acknowledgments

The authors thank the anonymous reviewers, Ronald Liu for manually verifying the transcriptions, and Miini Teng along with the UBC Canary project members for their valuable comments and suggestions. The authors also extend their appreciation to all the trainees and research assistants who have contributed to the Canary project over the years by collecting and refining data. This work was supported by the Faculty of Medicine at the University of British Columbia. The computing resources are provided by the Digital Research Alliance of Canada (alliancecan.ca).

References

Sharon Ash and Murray Grossman. 2015. Why study connected speech production. *Cognitive neuroscience of natural language use*, pages 29–58.

Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. 2023. Whisperx: Time-accurate speech transcription of long-form audio. *INTERSPEECH 2023*.

Oswald Barral, Hyeju Jang, Sally Newton-Mason, Sheetal Shajan, Thomas Soroski, Giuseppe Carenini, Cristina Conati, and Thalia Field. 2020. Non-invasive classification of alzheimer's disease using eye tracking and language. In *Machine Learning for Healthcare Conference*, pages 813–841. PMLR.

Anastasiya Belyaeva, Justin Cosentino, Farhad Hormozdiari, Krish Eswaran, Cory Shravya Shetty, Y McLean, Greg Corrado, and Nicholas A Furlotte1 AndrewB Carroll. 2023. Multimodal llms for health grounded in individual-specific data. In *Machine Learning for Multimodal Healthcare Data: First International Workshop, ML4MHD 2023, Honolulu, Hawaii, USA, July 29, 2023, Proceedings*, volume 14315, page 86. Springer Nature.

Shauna Berube, Jodi Nonnemacher, Cornelia Demsky, Shenly Glenn, Sadhvi Saxena, Amy Wright, Donna C Tippett, and Argye E Hillis. 2019. Stealing cookies in the twenty-first century: Measures of spoken narrative in healthy versus speakers with aphasia. *American journal of speech-language pathology*, 28(1S):321–329.

Helen Bird, Matthew A Lambon Ralph, Karalyn Patterson, and John R Hodges. 2000. The rise and fall of frequency and imageability: Noun and verb production in semantic dementia. *Brain and language*, 73(1):17–49.

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

Tom B Brown. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Shih-Han Chou, Miini Teng, Harshinee Sriram, Chuyuan Li, Giuseppe Carenini, Cristina Conati, Thalia S Field, Hyeju Jang, and Gabriel Murray. 2025. Multimodal classification of alzheimer's disease by combining facial and eye-tracking data. In *Machine Learning for Health (ML4H)*, pages 219–232. PMLR.

Hejie Cui, Zhuocheng Shen, Jieyu Zhang, Hui Shao, Lianhui Qin, Joyce C Ho, and Carl Yang. 2024. Llms-based few-shot disease predictions using ehr: A novel approach combining predictive agent reasoning and critical agent instruction. *arXiv preprint arXiv:2403.15464*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Anna Favaro, Najim Dehak, Thomas Thebaud, Jesús Villalba, Esther Oh, and Laureano Moro-Velázquez. 2024. Discovering invariant patterns of cognitive decline via an automated analysis of the cookie thief picture description task. In *The Speaker and Language Recognition Workshop (Odyssey 2024)*, pages 201–208.

Shangbin Feng, Weijia Shi, Yike Wang, Wenxuan Ding, Vidhisha Balachandran, and Yulia Tsvetkov. 2024. Don't hallucinate, abstain: Identifying llm knowledge gaps via multi-llm collaboration. *arXiv preprint arXiv:2402.00367*.

Katrina E Forbes-McKay and Annalena Venneri. 2005. Detecting subtle spontaneous language decline in early alzheimer's disease with a picture description task. *Neurological sciences*, 26:243–254.

Kathleen C Fraser, Kristina Lundholm Fors, Marie Eckerström, Fredrik Öhman, and Dimitrios Kokkinakis. 2019. Predicting mci status from multimodal language data using cascaded classifiers. *Frontiers in aging neuroscience*, 11:205.

Kathleen C Fraser, Jed A Meltzer, and Frank Rudzicz. 2016. Linguistic features identify alzheimer's disease in narrative speech. *Journal of Alzheimer's Disease*, 49(2):407–422.

Elaine Giles, Karalyn Patterson, and John R Hodges. 1996. Performance on the boston cookie theft picture description task in patients with early dementia of the alzheimer's type: missing information. *Aphasiology*, 10(4):395–408.

Adam Goodkind, Michelle Lee, Gary E Martin, Molly Losh, and Klinton Bicknell. 2018. Detecting language impairments in autism: A computational analysis of semi-structured conversations with vector semantics. In *Proceedings of the Society for Computation in Linguistics (SCiL) 2018*, pages 12–22.

Loukas Ilias and Dimitris Askounis. 2022. Multimodal deep learning models for detecting dementia from speech and transcripts. *Frontiers in Aging Neuroscience*, 14:830943.

Hyeju Jang, Thomas Soroski, Matteo Rizzo, Oswald Barral, Anuj Harisinghani, Sally Newton-Mason, Saffrin Granby, Thiago Monnerat Stutz da Cunha Vasco, Caitlin Lewis, Pavan Tutt, et al. 2021. Classification of alzheimer's disease leveraging multi-task machine learning analysis of speech and eye-movement data. *Frontiers in Human Neuroscience*.

Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. 2021. How can we know when language models know? on the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9:962–977.

Mingyu Jin, Qinkai Yu, Dong Shu, Chong Zhang, Lizhou Fan, Wenyue Hua, Suiyuan Zhu, Yanda Meng, Zhenting Wang, Mengnan Du, et al. 2024. Health-llm: Personalized retrieval-augmented disease prediction system. *arXiv preprint arXiv:2402.00746*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Chuyuan Li, Chloé Braud, and Maxime Amblard. 2022. Multi-task learning for depression detection in dialogs. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 68–75.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2023. Holistic evaluation of language models. *Transactions on Machine Learning Research*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024b. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023a. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*.

Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, et al. 2023b. Can generalist foundation models outcompete special-purpose tuning? case study in medicine. *Medicine*, 84(88.3):77–3.

Shuyin Ouyang, Jie M Zhang, Mark Harman, and Meng Wang. 2023. Llm is like a box of chocolates: the non-determinism of chatgpt in code generation. *arXiv preprint arXiv:2308.02828*.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Roy H Perlis. 2023. Application of gpt-4 to select next-step antidepressant treatment in major depression. *MedRxiv*.

Yusu Qian, Haotian Zhang, Yinfei Yang, and Zhe Gan. 2024. How easy is it to fool your multimodal llms? an empirical analysis on deceptive prompts. *arXiv preprint arXiv:2402.13220*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Abel Salinas and Fred Morstatter. 2024. The butterfly effect of altering prompts: How small changes and jailbreaks affect large language model performance. *arXiv preprint arXiv:2401.03729*.

Saman Sarraf, Arman Sarraf, Danielle D DeSouza, John AE Anderson, Milton Kabia, and Alzheimer's Disease Neuroimaging Initiative. 2023. Ovitad: Optimized vision transformer to predict various stages of alzheimer's disease using resting-state fmri and structural mri data. *Brain Sciences*, 13(2):260.

Douglas W Scharre. 2019. Preclinical, prodromal, and dementia stages of alzheimer's disease. *Pract Neurol*, 15:36–47.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2024. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. In *The Twelfth International Conference on Learning Representations (ICLR)*.

Zhengyan Sheng, Zhiqiang Guo, Xin Li, Yunxia Li, and Zhenhua Ling. 2022. Dementia detection by fusing speech and eye-tracking representation. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6457–6461. IEEE.

Sonish Sivarajkumar, Mark Kelley, Alyssa Samolyk-Mazzanti, Shyam Visweswaran, and Yanshan Wang. 2024. An empirical evaluation of prompting strategies for large language models in zero-shot clinical natural language processing: algorithm development and validation study. *JMIR Medical Informatics*, 12:e55318.

Yifan Song, Guoyin Wang, Sujian Li, and Bill Yuchen Lin. 2024. The good, the bad, and the greedy: Evaluation of llms should not ignore non-determinism. *arXiv preprint arXiv:2407.10457*.

Harshinee Sriram, Cristina Conati, and Thalia Field. 2023. Classification of alzheimer's disease with deep learning on eye-tracking data. In *Proceedings of the 25th International Conference on Multimodal Interaction*, pages 104–113.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D Manning. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5433–5442.

Zhuo Wang, Rongzhen Li, Bowen Dong, Jie Wang, Xiuxing Li, Ning Liu, Chenhui Mao, Wei Zhang, Liling Dong, Jing Gao, et al. 2023. Can llms like gpt-4 outperform traditional ai tools in dementia diagnosis? maybe, but not today. *arXiv preprint arXiv:2306.01499*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

Kathryn M Yorkston and David R Beukelman. 1980. An analysis of connected speech samples of aphasic and normal speakers. *Journal of speech and hearing disorders*, 45(1):27–36.

Letian Zhang, Xiaotong Zhai, Zhongkai Zhao, Yongshuo Zong, Xin Wen, and Bingchen Zhao. 2024. What if the tv was off? examining counterfactual reasoning abilities of multi-modal language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21853–21862.

Appendix A Prompt Templates

We provide prompting templates used in Llama and LLaVA in Table 5 and Table 6, respectively.

Appendix B Supervised Classifiers

Following Jang et al. (2021), we reduplicate the supervised learning results using Logistic Regression (LR), Random Forest (RF), and Gaussian Naive Bayes (GNB), all implemented with Scikit-learn library (Pedregosa et al., 2011). We split the dataset into 10 folds with 10 different seeds for cross-validation. The micro-averaged scores are given in Table 4, in comparison with our best zero-shot and few-shot prompting strategies.

Note that noting that our results differ slightly from those reported in Jang et al. (2021), as we do not use the exact same training samples (79 Patients and 83 Controls *vs.* our dataset with 63 Patients and 67 Controls). Additionally, we employ different speech-to-text methods, which may have led to variations in the transcripts.

Training	Model	AUC	Sensitivity	Specificity
Supervised	GNB	72.8 ± 2.2	64.1 ± 2.2	66.5 ± 3.5
	LR	73.2 ± 1.7	68.5 ± 3.8	70.2 ± 1.6
	RF	75.2 ± 3.1	67.7 ± 4.6	73.1 ± 3.6
Ours (zero-shot)	Llama-3.1	76.1 ± 2.0	**71.8** ± 3.4	73.9 ± 2.1
Ours (few-shot)	Llama-3.1	**76.4** ± 2.9	71.7 ± 7.6	**75.8** ± 8.4

Table 4: Comparison of fully supervised classifiers (top) and our methods using LLMs (bottom). RF: random forest, GNB: Gaussian Naive Bayes, LR: logistic regression. Supervised results are averaged over 10-seed 10-fold cross-validation. Prompting results are averaged over 2-trial 3-run per setup.

Appendix C LLaVA Prompting Results

We report zero-shot and few-shot prompting results with LLaVA-NeXT-8B in Table 7. Textual prompts and hyper-parameters are the same as with Llama. Different from Llama, we do not observe a consistent improvement with more complete prompts. Plausibly, LLaVA is not capable of processing longer and more complex textual information. We also observe some *extreme* predictions where the model only predicts Patient or Control, as highlighted in pink in Table 7.

Appendix D LLaVA's Speech on Picture Description Task

We prompt LLaVA twenty times on picture description task. The instruction is provided in the same way as for human participants. We then use supervised classifiers to categorize its responses. Across all of LLaVA's outputs, the three classifiers consistently classify them as "Patient" with high probabilities: RF $> 60\%$, GNB $> 90\%$, and LR $> 80\%$. Two examples are given in Table 8.

Llama-3.1	
Strategy	Template
Background Prompt	`Role:` You are a medical expert in Alzheimer’s disease. You analyze linguistic features in the patient’s speech, such as lexical richness, syntactic complexity, grammatical correctness, information content, and semantic coherence. Based on the participant’s speech, provide an initial diagnosis of dementia patient (P) and healthy control (H). `Context:` The Boston Cookie Theft picture description task is a well established speech assessment in Alzheimer’s disease. During the task, participants are shown the picture and are asked to describe everything they see in the scene using as much time as they would like. Based on the participant’s description, make a classification of dementia patient (P) versus healthy control (H). `Context+Role+Ling:` The Boston Cookie Theft picture description task is a well established speech assessment in Alzheimer’s disease. During the task, participants are shown the picture and are asked to describe everything they see in the scene using as much time as they would like. The objects (also known as information units) in this picture includes: “cookie”, “girl”, “boy”, “woman”, “jar”, “stool”, “plate”, “dishcloth”, “water”, “window”, “cupboard”, “curtain”, “dishes”, “sink”. You are a medical expert in Alzheimer’s disease. You analyze linguistic features in the patient’s speech, such as lexical richness, syntactic complexity, grammatical correctness, information units, and semantic coherence. Based on the participant’s description of the picture, provide an initial diagnosis of dementia patient (P) and healthy control (H).
Example Prompt	`Zero-shot: None` `Few-shot:` Example: ## Text: <text> ## Answer: healthy control (H). ## Text: <text> ## Answer: dementia patient (P).
Question Prompt	`Short:` Given the text below, classify the participant as a dementia patient (P) or healthy control (H). Please give an answer and a probability without explanation. `CoT:` Given the text below, classify the participant as a dementia patient (P) or healthy control (H). First explain step-by-step and then give a prediction with a probability. `G. CoT:` Given the text below, classify the participant as a dementia patient (P) or healthy control (H). Please first reason from the following perspectives: (1) Vocabulary richness: such as the usage of different words; (2) Syntactic complexity: such as the length of the sentence and the number of subordinate clauses; (3) Information content: whether the participant describe most of the information units in the picture; (4) Semantic coherence: such as the usage of connectives and the change in description from one information unit to another; (5) Fluency and repetitiveness: whether the text is fluent with less repetitive sentences. Based on your reasoning, please give a prediction and the corresponding probability.

Table 5: Prompting template used in Llama.

LLaVA-NeXT

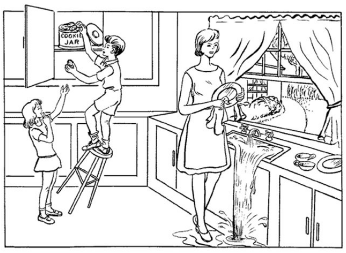
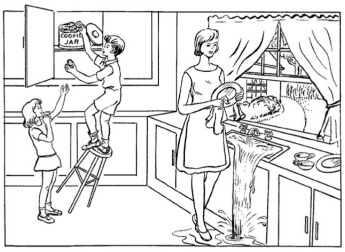
Strategy	Template
Background Prompt	 `Role:` This image is used for speech assessment in Alzheimer's disease. You are a medical expert in Alzheimer's disease. You analyze linguistic features in the patient's speech, such as lexical richness, syntactic complexity, grammatical correctness, information content, and semantic coherence. Based on the participant's speech, provide an initial diagnosis of dementia patient (P) and healthy control (H). `Context:` This image is used in Boston Cookie Theft picture description task, which is a well established speech assessment in Alzheimer's disease. During the task, participants are shown the picture and are asked to describe everything they see in the scene using as much time as they would like. Based on the participant's description, make a classification of dementia patient (P) versus healthy control (H). `Context+Role+Ling:` This image is used in Boston Cookie Theft picture description task, which is a well established speech assessment in Alzheimer's disease. During the task, participants are shown the picture and are asked to describe everything they see in the scene using as much time as they would like. The objects (also known as information units) in this picture includes: "cookie", "girl", "boy", "woman", "jar", "stool", "plate", "dishcloth", "water", "window", "cupboard", "curtain", "dishes", "sink". You are a medical expert in Alzheimer's disease. You analyze linguistic features in the patient's speech, such as lexical richness, syntactic complexity, grammatical correctness, information units, and semantic coherence. Based on the participant's description of the picture, provide an initial diagnosis of dementia patient (P) and healthy control (H).
Example Prompt	`Zero-shot: None` `Few-shot:` Example: ## Text: <text> ## Answer: healthy control (H). ## Text: <text> ## Answer: dementia patient (P).
Question Prompt	`Short:` Given the text below, classify the participant as a dementia patient (P) or healthy control (H). Please give an answer and a probability without explanation. `CoT:` Given the text below, classify the participant as a dementia patient (P) or healthy control (H). First explain step-by-step and then give a prediction with a probability. `G. CoT:` Given the text below, classify the participant as a dementia patient (P) or healthy control (H). Please first reason from the following perspectives: (1) Vocabulary richness: such as the usage of different words; (2) Syntactic complexity: such as the length of the sentence and the number of subordinate clauses; (3) Information content: whether the participant describe most of the information units in the picture; (4) Semantic coherence: such as the usage of connectives and the change in description from one information unit to another; (5) Fluency and repetitiveness: whether the text is fluent with less repetitive sentences. Based on your reasoning, please give a prediction and the corresponding probability.

Table 6: Prompting template used in LLaVA.

		Random Few-shot			kNN Few-shot			Zero-shot		
B prompt	Q prompt	AUC	Sensitivity	Specificity	AUC	Sensitivity	Specificity	AUC	Sensitivity	Specificity
Role	Short	67.1 ± 7.0	45.1 ± 9.7	85.8 ± 10.1	58.1 ± 0.0	76.2 ± 0.0	40.3 ± 0.0	57.4 ± 0.0	19.4 ± 0.0	92.3 ± 0.0
	CoT	62.4 ± 9.3	27.2 ± 16.2	90.8 ± 5.6	70.6 ± 1.0	81.0 ± 2.2	41.0 ± 1.1	49.2 ± 0.0	19.4 ± 0.0	92.3 ± 0.0
	G. CoT	54.9 ± 5.2	3.0 ± 7.1	100.0 ± 0.0	69.8 ± 1.3	81.8 ± 1.1	39.6 ± 1.1	58.9 ± 0.0	37.1 ± 0.0	72.3 ± 0.0
Context	Short	67.1 ± 7.0	54.3 ± 14.6	73.6 ± 12.5	67.3 ± 0.4	73.0 ± 0.0	57.5 ± 1.1	50.0 ± 0.0	100.0 ± 0.0	0.0 ± 0.0
	CoT	61.4 ± 7.3	32.4 ± 7.4	89.2 ± 6.1	64.9 ± 2.5	77.0 ± 3.4	47.1 ± 3.2	59.0 ± 0.0	14.5 ± 0.0	100.0 ± 0.0
	G. CoT	52.2 ± 1.8	3.2 ± 2.4	99.3 ± 1.6	65.1 ± 3.6	84.9 ± 1.1	29.9 ± 2.1	63.0 ± 0.0	66.1 ± 0.0	63.1 ± 0.0
Context	Short	63.4 ± 4.3	30.7 ± 13.9	89.6 ± 11.2	65.8 ± 0.0	74.6 ± 0.0	56.7 ± 2.1	60.9 ± 0.0	82.3 ± 0.0	38.5 ± 0.0
+Role	CoT	51.6 ± 11.1	19.3 ± 10.2	89.8 ± 8.8	65.1 ± 1.5	81.8 ± 1.1	35.1 ± 3.1	55.9 ± 0.0	33.9 ± 0.0	80.0 ± 0.0
+Ling	G. CoT	50.1 ± 4.7	3.2 ± 3.6	99.3 ± 1.6	69.4 ± 0.4	84.1 ± 2.2	35.8 ± 4.2	64.9 ± 0.0	30.7 ± 0.0	95.4 ± 0.0

Table 7: Random, *k*NN few-shot, and zero-shot prompting results with LLaVA. Random few-shot results are averaged from three sampling. Best AUC in random, *k*NN, and zero-shot sections is highlighted; second best is underlined. *Extreme* prediction such as very high Sensitivity and very low Specificity (or reversely) is highlighted.

Picture Description Task on LLaVA-NeXT



`Prompt:` You will do a picture description task. You'll be shown a picture. Please describe everything you see going on in the picture.

`LLaVA:` The image appears to be a black and white line drawing depicting a domestic scene. There are four figures in the image, each outlined and shaded to suggest depth and form. The drawing style is simple and cartoonish, with no color or shading other than the outlines and the shading within the outlines. In the foreground, there is a woman standing in the kitchen, holding a pan. She seems to be engaged in cooking or preparing food. To her right, there is a child standing on a stool, reaching for something on a shelf. The child is wearing a short-sleeved shirt and shorts, and appears to be in the process of retrieving an item. In the background, there is another child standing on the floor, looking up at the woman. This child is also wearing a short-sleeved shirt and shorts, and seems to be observing the woman's actions. On the far right, there is a baby lying on a surface, possibly a bed or a couch, with a blanket or sheet covering them. The baby appears to be asleep or resting. The kitchen area includes a sink and a countertop with various items such as a bowl and a bottle. There is also a window with curtains, and a door that is partially open. The room has a simple, clean layout with minimal decoration. The overall scene suggests a domestic setting, possibly a family home, with the woman and children engaged in everyday activities.

`LLaVA:` The image appears to be a black and white line drawing depicting a domestic scene. There are four characters in the picture: 1. A woman standing in the center, holding a plate. She is wearing a dress and seems to be engaged in an activity related to food preparation or serving. 2. A child standing on a stool to the left of the woman, reaching for something on a shelf. The child is wearing a short-sleeved top and shorts. 3. Another child standing to the right of the woman, also reaching for something on a shelf. This child is wearing a long-sleeved top and shorts. 4. A baby lying on a surface to the right of the woman, seemingly asleep or resting. The setting appears to be a kitchen or a room with a counter and a sink. There is a window with curtains to the right, and a door to the left. On the counter, there are various items including what looks like a bowl and a jar. The overall scene suggests a family moment, possibly meal preparation or cleanup. The characters are drawn in a simple, cartoon-like style, which is common in educational materials or children's books.

Table 8: Two examples of LLaVA's generations on Cookie Theft task. The *hallucinated* context are marked in red.