NAACL 2025

Patient-Oriented Language Processing (CL4Health) Workshop

On Large Foundation Models and Alzheimer's Disease Detection

Chuyuan Li, Giuseppe Carenini, Thalia S. Field

The University of British Columbia

{chuyuan.li, thalia.field}@ubc.ca, carenini@cs.ubc.ca

Background

- **Emergent Capabilities**: Large proprietary Language Models (LLMs) such as GPT-4 have shown impressive performance on professional benchmarks in the health domain.
- **Interpretable Explanations**: LLMs can generate interpretable explanations to their predictions, providing clinical doctors with valuable insights into their reasoning.
- **Considerations in Healthcare Domain**: Third-party commercial LLMs is not always feasible due to concerns about traceability, privacy, and security.

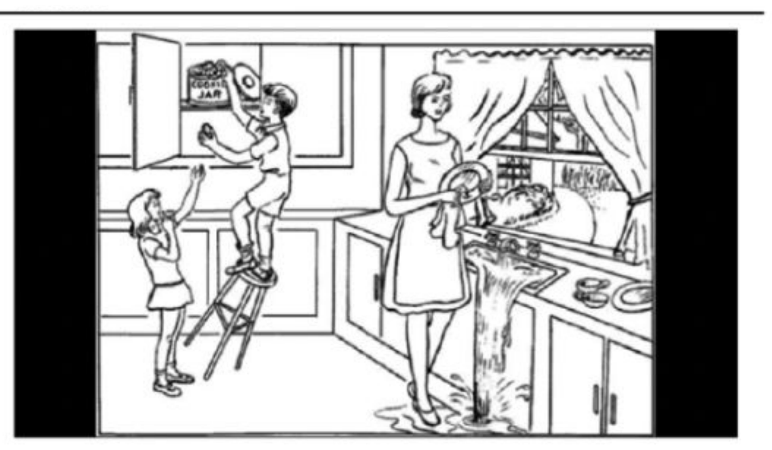
In this paper, we explore using **small (e.g., less than 10B), cost-effective open-source** Foundation Models such as **Llama-3.1-8B** (language-only) and **Llama3-LLaVA-NeXT** (vision language model) for AD detection.

Task and Dataset

- Task: *Cookie Theft* Picture Description.
- Canary dataset [1]: it contains 130 participants where 67 are healthy controls and 63 are AD patients. Patients are diagnosed or exhibiting initial symptoms potentially progressing to AD.

Picture description task

"You will be shown a picture on the screen. Describe everything you see going on in this picture. Try not to look away from the screen while describing the picture."

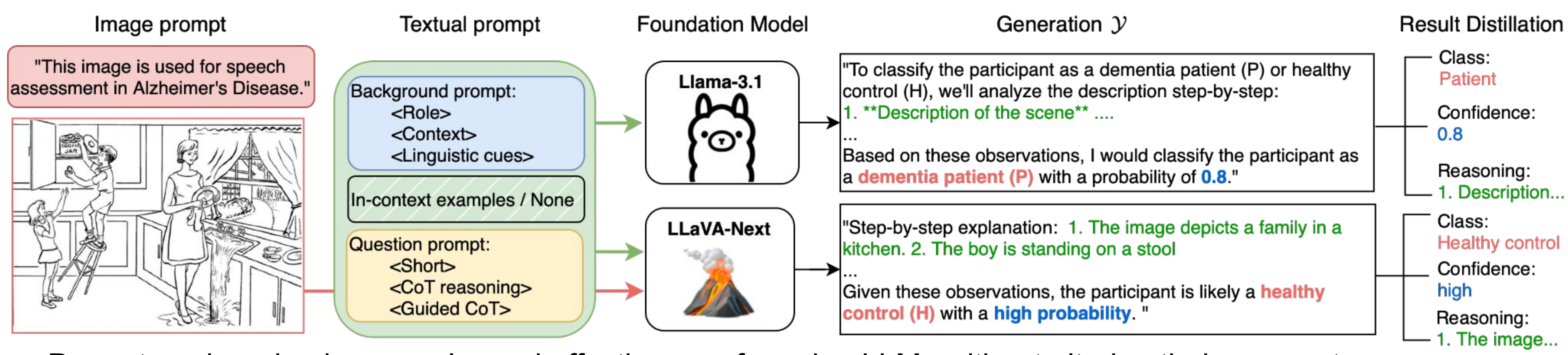


Group	#	Age	Gender	MoCA
Patient	63	72 ± 9	31M / 34F	18 ± 7
Control	67	62 ± 15	22M / 45F	27 ± 3

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

[1] *Hyeju Jang et al. 2021. Classification of alzheimer's disease leveraging multi-task machine learning analysis of speech and eye-movement data. Frontiers in Human Neuroscience.*

Power of Prompting



- Prompt engineering is a popular and effective way for using LLMs without altering their parameters.
- We design our prompts in a systematic way to unleash the inner specialist capabilities of LLMs, including:
  - Background prompt: with cue phrases "Role", "Context", and "Linguistic cues"
  - Example prompt: *In-Context Learning* pairs, we employ fixed (random) and dynamic (kNN) selection with one positive and one negative demonstration.
  - Question prompt: compare Short answer, Chain-of-thought (CoT), and Guided CoT for LLM output.
  - Each prompt setting was run 6 times with a lower temperature (0.1) to mitigate model instability.

Results and Take-aways

- Comprehensive background prompt and CoT reasoning gives optimal performance, even surpass supervised classifiers.

Background	Question	AUC	Sensitivity	Specificity
Role	Short	60.3 ± 1.1	$\mathbf{96.4} \pm 0.8$	11.5 ± 0.8
	CoT	65.8 ± 0.5	91.13 ± 1.1	24.6 ± 2.5
	G. CoT	$\mathbf{70.9} \pm 0.4$	84.7 ± 1.1	$\mathbf{35.4} \pm 2.1$
Context	Short	69.4 ± 1.5	35.9 ± 2.0	$\mathbf{93.5} \pm 1.4$
	CoT	68.9 ± 0.6	50.8 ± 1.1	73.9 ± 2.1
	G. CoT	$\mathbf{74.3} \pm 1.1$	$\mathbf{69.4} \pm 2.2$	69.3 ± 0.0
Context +Role +Ling	Short	71.6 ± 0.5	$\mathbf{72.6} \pm 0.0$	69.6 ± 1.4
	CoT	72.9 ± 3.8	70.2 ± 3.4	70.8 ± 4.3
	G. CoT	$\mathbf{76.1} \pm 2.0$	71.8 ± 3.4	$\mathbf{73.9} \pm 2.1$
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

- Vision-language model (VLMs) like LLaVA, despite its additional vision (image), **underperforms** language-only LLMs like Llama, both in zero-shot and few-shot settings.
- Sanity check on LLaVA reveals that it's unable to generate normal speech during the picture description task, raising open questions on VLMs compositional capabilities.