The Eyes Don't Lie: Text Transcriptions Can Hide Dementia Presentation that Gaze Reveals

Anonymous ACL submission

Abstract

Current methods used to diagnose or monitor dementia-related cognitive decline predominantly rely on audio recordings. Such audio recordings can leak personally identifiable information and create new risks given deep fake technology. We introduce generative likelihood-based approaches to identify differences in healthy versus dementia-diagnosed participants via gaze tracking and text transcriptions during a standard diagnostic image description task without relying on sensitive audio information. Contrasting conventional wisdom, we find that text transcriptions alone are not a reliable measure of cognitive impairment in this task, finding gaze tracking to be more reliable, and suggesting existing results in language-based dementia detection rely primarily on audio signals.

1 Introduction

Continual monitoring of cognitive change can enable early detection of Alzheimer's and related dementias, facilitating earlier intervention and treatment (Rasmussen and Langerman, 2019). Existing dementia detection resources and methods largely focus on audio and text transcriptions, but we find that reliable detection from short interactions with participants is achievable through *gaze tracking* in tandem with text transcriptions.

Around 70% of Americans said they would want Alzheimer's disease identified if that knowledge led to earlier treatment (Alzheimer's Association, 2023), but available clinically validated measures of cognitive change for early detection of Alzheimer's take place at most every three (P et al., 2009; CB et al., 2023) to six (KV et al., 2024) For some, these important checks don't take place at all until after advanced symptoms are present. Developing computational models to detect the onset of dementia-related cognitive decline in time for medical intervention and evaluation is an under-

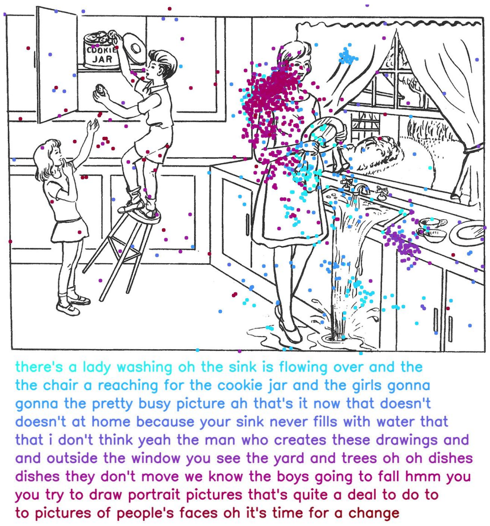


Figure 1: We investigate training-free dementia detection methods from gaze tracking (colored dots) and transcript text (colored words) of participants describing The Cookie Theft Picture.

explored problem, as most existing ML detection methods are based on data from a single assessment and modality of interaction, such as speech (Becker et al., 1994; Luz et al., 2020).

The speech data from those existing works often includes a verbal task where participants spend up to two minutes describing a line drawing scene (Figure 1). This task is a component of the The Boston Diagnostic Aphasia Examination (Goodglass et al., 2001) frequently used by clinicians for screening for dementia symptom presentation. Existing works that train machine learning models to detect the presence of dementia symptoms largely focus on audio signals or hand-crafted features summarizing aspects of text transcripts (Santander-Cruz et al., 2022; Kumar et al., 2022; Javeed et al., 2023; Shi et al., 2023). These approaches typically train simple classifiers such as SVM, Random Forest, and logistic regression (Diogo et al., 2022; Haider et al., 2020), or fine-tune existing pretrained models such as BERT (Balagopalan et al., 2020), RoBERTa (Matošević and Jović, 2022), and GPT-2 (Liu and Wang, 2023).

In this paper, we take a step towards non-invasive, privacy-preserving, in-home monitoring tools for detecting early signs and symptoms of dementia. We explore analysis methods on *raw* gaze and text data that require no hand-crafted features, federated learning across participants, or even back propagation gradient passes on existing models, all of which can inadvertently leak personally identifiable information. In short, we explore methods to detect dementia symptoms from the under explored spaces of gaze tracking and verbal text transcriptions. The contributions of the paper can be summarized as follows:

- We empirically demonstrate that the gaze of Control group focus on the areas-of-interest presented in The Cookie Theft Picture compared to that of participants with AD.
- We similarly demonstrate that the Control group's text transcript descriptions of The Cookie Theft Picture correspond more closely to the expectations of large, pretrained image captioning model when compared to participants with AD.
- Our analyses do not rely on hand-crafted features related to analyzing dementia presentation, and instead leverage pretrained models and statistical machine learning models to measure deviation from expected gaze patterns and sequences of descriptive words in terms of likelihood without any additional model training or fine-tuning.

2 Participant Gaze and Text Data

We analyze a dataset of participant tracked eye gaze and human-corrected transcripts of participant speech during the completion of The Cookie Theft Description Task. The study included 25 Control group participants with healthy cognitive function and 14 participants with an Alzheimer's Disease (AD) diagnosis. Participants were all patients at a local aging research center, at which they were also recruited for enrollment in the study.

During each participant session, we recorded eye gaze and audio while the participant viewed The Cookie Theft Picture on a Surface Laptop Studio equipped with an Intel Core i7 processor, 32 GB of RAM, a 1TB SSD, Microsoft OS, an NVIDIA GeForce RTX graphics card, and a Tobii Pro X3-120 eye tracker. Eye tracking was calibrated using Tobii Manager software, with gaze data gathered via the Tobii Pro SDK 3 [1]. Audio was processed to speech transcriptions standardized using the Automatic Speech Recognition (ASR) Vosk Model [2] followed by manual annotation by a person to correct any ASR errors. After removing gaze points from timesteps when none was tracked and cleaning up text transcriptions, we have an average of 8312.82$\pm$ 4993.53 gaze points and 167.87$\pm$58.47 transcribed words of description across the 39 total participants whose sessions lasted, on average, 94.89$\pm$21 seconds.

3 Hypotheses and Methods

The methods described in the paper do not involve training algorithms based on participant data or processing participant data for any sort of hand-crafted feature extraction. Instead, these methods utilize pre-trained, generative models to estimate likelihoods of observed data being generated by a background, "healthy" distribution. We hypothesize that:

H1 gaze points collected from participants in the AD group will exhibit lower likelihood of gaze being explained by annotated areas of interest in the stimulus image than will the Control group; and

H2 text transcribed from audio of participants in the AD group will exhibit lower likelihood due to syntactic fluency and topic consistency ($\mathbf{H2}_1$) as well as relevance to the stimulus image ($\mathbf{H2}_2$).

H3 gaze and text will reveal complementary participant cognitive function.

For the purposes of evaluating our hypothesis, we calculate the average *Negative Log-Likelihood* (NLL) of sequences of gaze points and transcription words for each participant. Note that a lower NLL corresponds to a lower likelihood, while a high NLL indicates a higher likelihood.

3.1 Gaze: Semantic GMM

We fit a Gaussian Mixture Models (GMM) to Areas of Interest (AOI) in The Cookie Theft Picture annotated by an experimenter. We learn a $k = 17$ component mixture of Gaussians, each defined by a mean (μ_k), covariance (σ_k), and mixing coefficient

[1] https://developer.tobiipro.com/python/python-oldmigrationsdk.html

[2] https://alphacephei.com/vosk/models/vosk-model-en-us-0.21.zip

Figure 2: Heatmap displaying the likelihood estimations by the GMM across The Cookie Theft Picture.

(π_k). Figure 2 visualizes the likelihood heatmap by pixel in the stimulus image of this fitted "Semantic GMM." We calculate the average NLL for a set of gaze points $\{x, y\}^N$ by:

$$\overline{\text{NLL}}_{\text{gaze}}(\{x, y\}^N) = -\frac{1}{N}\sum_{n=1}^{N} \ln\left(\sum_{k=1}^{K} \pi_k \mathcal{N}((x_n, y_n)|\mu_k, \Sigma_k)\right).$$

We calculate this average log-likelihood per participant, then analyze the differences in these gaze likelihood samples between the control and AD populations.

3.2 Text Transcripts: Pretrained LLMs

We utilize two pretrained large language models (LLMs) that decode autoregressively and can be run on-device to calculate the average likelihood of the sequence of transcribed tokens from participant descriptions. Given trained LMM parameters θ yielding a distribution $p_\theta(x_i|x_{1\ldots i-1})$ of next token probability, the average log-likelihood of a token sequence $\vec{x} :=$ LLM-Tokenizer($\vec{w}$) from participant transcript word sequence $\vec{w}$ is calculated as:

$$\overline{\text{NLL}}_{\text{text}}(\vec{x}) = -\frac{1}{N}\sum_{i=1}^{N} \ln p_\theta(x_i|x_{1\ldots i-1}).$$

GPT-2 (Radford et al., 2019) is a transformer-based model that was pre-trained on substantial English data using self-supervised learning techniques, primarily focusing on predicting the next word in sentences. We can consider the GPT-2 NLL values to represent the *prior* likelihood of text, where differences in NLL scores are likely to correspond to syntactic fluency and topic consistency (**H2**$_1$). We use the GPT-2 Large model which can be run on-device, and break transcripts into tokens using the GPT-2 Large tokenizer.

BLIP, Bootstrapping Language-Image Pre-training (Li et al., 2022), is a model pretrained on large-scale image-text datasets using self-supervised learning techniques to autoregressivelyi predict textual descriptions of input images. The BLIP NLL values represent *posterior* likelihoods of text descriptions conditioned on The Cookie Theft Picture stimulus, and may expose more nuanced differences in semantic relevance between control and AD participants (**H2**$_2$). We use the BLIP-image-captioning-base [3], a BLIP processor which wraps a BERT tokenizer [4] and BLIP image processor into a single processor. For a fair comparison against GPT-2, we additionally test BLIP with a blank image input, treating it as another *prior* likelihood measure in that case.

4 Experiments and Results

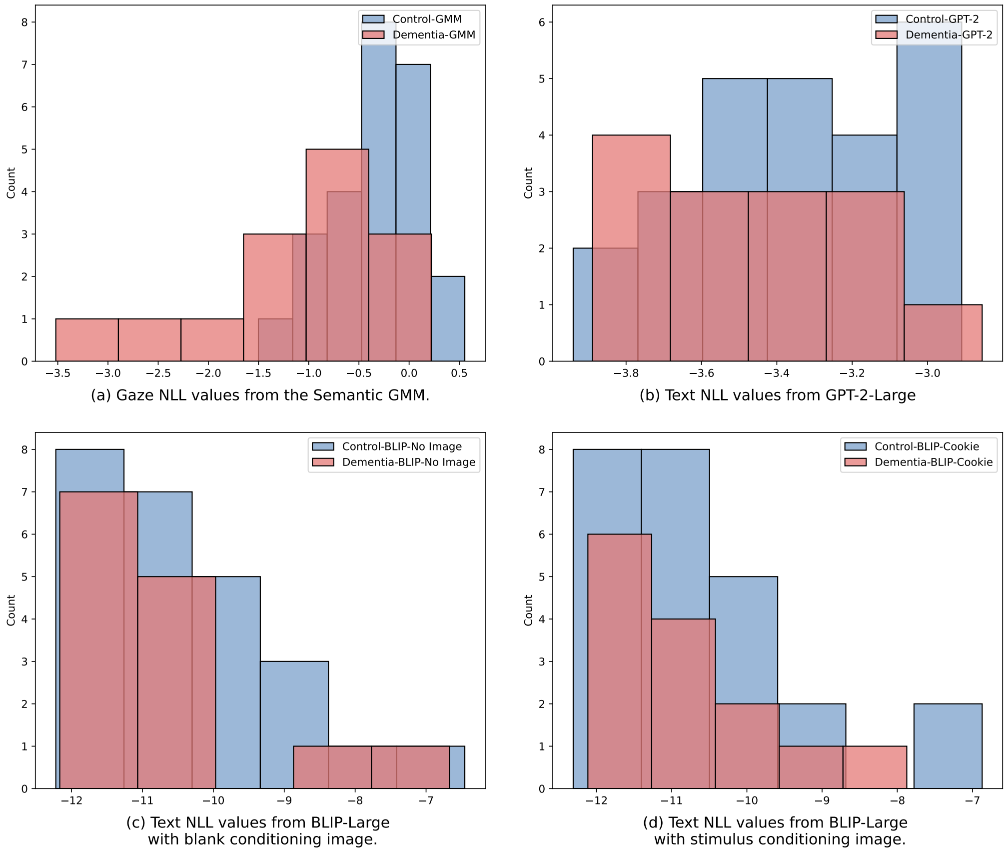
The experimental results reveal that significant differences exist between the Control and AD groups when analyzing eye gaze data. However, the differences between text transcripts are less consistent. In multimodal analyses combining eye gaze and text transcripts, the Hotelling T-square indicate significant differences between the two groups when using GPT-2.

Gaze reveals AD symptoms. To evaluate hypothesis **H1**, we compared the population of $\overline{\text{NLL}}_{\text{gaze}}$ values of the 25 control patients to those of the 14 patients with an AD diagnosis using a one-sided, Welch's unequal variances t-tests. Figure 3(a) shows histograms of $\overline{\text{NLL}}_{\text{gaze}}$ values between the populations. The average $\overline{\text{NLL}}_{\text{gaze}}$ of the control group was found to be statistically significantly higher than that of the AD group, with p-value .0158, providing supporting evidence for **H1**.

Transcription text is not enough. To evaluate **H2**, we compared populations of $\overline{\text{NLL}}_{\text{text}}$ values between 24 control and 14 AD patient groups using autoregressive text-only and image-conditioned LLMs using one-sided, Welch's unequal variances t-tests. Figures 3(b), 3(c), and 3(d) show the distribution of $\overline{\text{NLL}}_{\text{text}}$ values between each population as estimated by GPT-2, BLIP with a blank conditioning image, and BLIP conditioned on The

[3] https://huggingface.co/Salesforce/blip-image-captioning-base

[4] https://huggingface.co/docs/transformers/v4.41.3/en/model_doc/bert#transformers.BertTokenizerFast



(a) Gaze NLL values from the Semantic GMM.

(b) Text NLL values from GPT-2-Large

(c) Text NLL values from BLIP-Large with blank conditioning image.

(d) Text NLL values from BLIP-Large with stimulus conditioning image.

Figure 3: Average NLL values from control group gaze and text transcripts estimated via the Semantic GMM (a), GPT-2-Large (b), BLIP-Large conditioned on a blank image (c), and BLIP-Large conditioned on the stimulus image.

Cookie Theft Picture stimulus image. The corresponding p-values are .0795, .317, and .355, respectively. These results suggest that there may be support for **H2**$_1$, that there are measurable likelihood-based differences in control versus AD patient transcripts with respect to syntactic fluency and topical consistency (as measured by GPT-2; $p = 0.0795$). However, the image-conditioned BLIP model, with both a blank image and the actual stimulus image, show no substantial differentiation in likelihood estimates of transcription tokens between the groups; we suspect this result may arise from the misalignment between BLIP's image caption language pretraining data and the long form text transcription descriptions of images.

Transcription May Not Complement Gaze. We used a Multivariate Hotelling's T-square test to compare participant $\overline{\text{NLL}}_{\text{gaze}}$ and $\overline{\text{NLL}}_{\text{text}}$ data simultaneously. This multivariate population difference was found statistically significant, but we repeated the test with identical 0 values substituted for $\overline{\text{NLL}}_{\text{text}}$ for all participants also found significance. Our findings do not support **H3**. Participant $\overline{\text{NLL}}_{\text{text}}$ contributed no significant information about the presenc or absence of dementia symptoms compared to $\overline{\text{NLL}}_{\text{gaze}}$ alone.

5 Future Work

While our areas of interest for gaze analysis are hand-annotated, we note that pretrained segmentation models such as Meta AI's Segment Anything (Kirillov et al., 2023) may handle line drawings like The Cookie Theft Picture. Additionally, methods like MDETR (Kamath et al., 2021) can identify image regions corresponding to input language, opening another way to measure alignment of participant transcripts. Similarly, while our transcriptions are hand-corrected, we note that the ASR system produced an estimated WER rate of only 5.43, and that future work may be able to incorporate visual priors from the image itself to improve automatic transcription (Chang et al., 2023).

Limitations

We acknowledge that our study is based on a small sample of 39 participants, and the demographics are not balanced. Specifically, 71% of the participants are white Caucasians, and there is a 1 to 2 ratio of Alzheimer's Disease (AD) patients to healthy controls. This demographic imbalance may limit the generalizability of our findings to the broader population. However, we believe that our analysis highlights the value of methods like estimation log-likelihood for small datasets in both unimodal and multimodal approaches to dementia assessment. Our findings demonstrate the potential of using limited data effectively, offering evaluation metrics that can be applied to other multimodal tasks where access to large datasets is restricted.

Ethical Impact

This study recognizes the ethical concerns regarding privacy and potential information leakage in the collection and analysis of eye gaze data and text transcripts. To address these issues, we have implemented stringent data protection protocols, including anonymization, secure storage, and strict access controls. Informed consent was obtained from all participants, ensuring they understand how their data will be used and protected. Our research team is dedicated to continuously improving our practices to uphold the highest ethical standards, ensuring that the benefits of our research are achieved without compromising participant privacy and trust.

References

Alzheimer's Association. 2023. 2023 Alzheimer's disease facts and figures. *Alzheimer's & Dementia*, 19(4):1598–1695.

Aparna Balagopalan, Benjamin Eyre, Frank Rudzicz, and Jekaterina Novikova. 2020. To bert or not to bert: comparing speech and language-based approaches for alzheimer's disease detection. *arXiv preprint arXiv:2008.01551*.

J. T. Becker, F. Boller, O. L. Lopez, J. Saxton, and K. L. McGonigle. 1994. The natural history of alzheimer's disease: description of study cohort and accuracy of diagnosis. *Archives of Neurology*, 51(6):585–594. Grant Support: NIA AG03705 and AG05133.

Young CB, Mormino EC, Poston KL, Johnson KA, Rentz DM, Sperling RA, and Papp KV. 2023. Computerized cognitive practice effects in relation to amyloid and tau in preclinical Alzheimer's disease: Results from a multi-site cohort. *Alzheimers Dement (Amst).*, 15(1).

Allen Chang, Xiaoyuan Zhu, Aarav Monga, Seoho Ahn, Tejas Srinivasan, and Jesse Thomason. 2023. Multimodal speech recognition for language-guided embodied agents. In *Annual Conference of the International Speech Communication Association (INTERSPEECH)*.

Vasco Sá Diogo, Hugo Alexandre Ferreira, Diana Prata, and Alzheimer's Disease Neuroimaging Initiative. 2022. Early diagnosis of alzheimer's disease using machine learning: a multi-diagnostic, generalizable approach. *Alzheimer's Research & Therapy*, 14(1):107.

Harold Goodglass, Edith Kaplan, and Sandra Weintraub. 2001. *BDAE: The Boston Diagnostic Aphasia Examination*. Lippincott Williams & Wilkins Philadelphia, PA, n/a.

Fasih Haider, Sofia De La Fuente Garcia, Pierre Albert, and Saturnino Luz. 2020. Affective speech for alzheimer's dementia recognition. *LREC: Resources and ProcessIng of linguistic, para-linguistic and extra-linguistic Data from people with various forms of cognitive/psychiatric/developmental impairments (RaPID)*, pages 67–73.

Ashir Javeed, Ana Luiza Dallora, Johan Sanmartin Berglund, Arif Ali, Liaqata Ali, and Peter Anderberg. 2023. Machine learning for dementia prediction: a systematic review and future research directions. *Journal of medical systems*, 47(1):17.

Aishwarya Kamath, Mannat Singh, Yann LeCun, Ishan Misra, Gabriel Synnaeve, and Nicolas Carion. 2021. MDETR–modulated detection for end-to-end multi-modal understanding. *arXiv preprint arXiv:2104.12763*.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. 2023. Segment anything. *arXiv:2304.02643*.

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

Papp KV, Jutten RJ, Soberanes D, Weizenbaum E, Hsieh S, Molinare C, Buckley R, Betensky RA, Marshall GA, Johnson KA, Rentz DM, Sperling R, and Amariglio RE. 2024. Early detection of amyloid-related changes in memory among cognitively unimpaired older adults with daily digital testing. *Ann Neurol.*, 95(3):507–517.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR.

Ning Liu and Lingxing Wang. 2023. An approach for assisting diagnosis of alzheimer's disease based on natural language processing. *Frontiers in Aging Neuroscience*, 15.

Saturnino Luz, Fasih Haider, Sofia de la Fuente, Davida Fromm, and Brian MacWhinney. 2020. Alzheimer's dementia recognition through spontaneous speech: The ADReSS challenge. In *Annual Conference of the International Speech Communication Association (INTERSPEECH)*.

Lovro Matošević and Alan Jović. 2022. Accurate detection of dementia from speech transcripts using roberta model. In *2022 45th Jubilee International Convention on Information, Communication and Electronic Technology (MIPRO)*, pages 1478–1484. IEEE.

Maruff P, Thomas E, Cysique L, et al. 2009. Validity of the CogState brief battery: relationship to standardized tests and sensitivity to cognitive impairment in mild traumatic brain injury, schizophrenia, and AIDS dementia complex. *Arch Clin Neuropsychol.*, 24(2):165–178.

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.

Jill Rasmussen and Haya Langerman. 2019. Alzheimer's disease — why we need early diagnosis. *Degenerative neurological and neuromuscular disease*, 9:123–130.

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

Mengke Shi, Gary Cheung, and Seyed Reza Shahamiri. 2023. Speech and language processing with deep learning for dementia diagnosis: A systematic review. *Psychiatry Research*, page 115538.