Dementia Through Different Eyes: Explainable Modeling of Human and LLM Perceptions for Early Awareness

Lotem Peled-Cohen[*T] **Maya Zadok**[*T] **Nitay Calderon**[T]
Hila Gonen[U] **Roi Reichart**[T]
[T] Faculty of Data and Decision Sciences, Technion
[U] Paul G. Allen School of Computer Science & Engineering, University of Washington
{splotem,maya.zadok,nitay}@campus.technion.ac.il
hilagnn@gmail.com roiri@technion.ac.il

Abstract

Cognitive decline often surfaces in language years before diagnosis. It is frequently non-experts, such as those closest to the patient, who first sense a change and raise concern. As LLMs become integrated into daily communication and used over prolonged periods, it may even be an LLM that notices something is off. But what exactly do they notice–and *should be noticing*–when making that judgment? This paper investigates how dementia is perceived through language by non-experts. We presented transcribed picture descriptions to non-expert humans and LLMs, asking them to intuitively judge whether each text was produced by someone healthy or with dementia. We introduce an explainable method that uses LLMs to extract high-level, expert-guided features representing these picture descriptions, and use logistic regression to model human and LLM perceptions and compare with clinical diagnoses. Our analysis reveals that human perception of dementia is inconsistent and relies on a narrow, and sometimes misleading, set of cues. LLMs, by contrast, draw on a richer, more nuanced feature set that aligns more closely with clinical patterns. Still, both groups show a tendency toward false negatives, frequently overlooking dementia cases. Through our interpretable framework and the insights it provides, we hope to help non-experts better recognize the linguistic signs that matter.

1 Introduction

Dementia is a progressive neurodegenerative condition caused by various underlying pathologies, most commonly Alzheimer's disease (AD). Tens of millions are currently living with the disease worldwide, with this figure expected to double with each passing generation (Alzheimer's Disease International, 2025). Early diagnosis is critical for maximizing the effectiveness of both symptomatic and disease-modifying interventions, especially during mild cognitive impairment (MCI), a transitional stage between normal aging and dementia, where cognitive deficits are present but do not yet impair daily functioning (Prince et al., 2011; Rasmussen and Langerman, 2019; Sims et al., 2023).

Subtle language dysfunctions have long been recognized as early signs of cognitive decline, making linguistic signals valuable for early detection (Verma and Howard, 2012; Szatloczki et al., 2015b; Orimaye et al., 2017; Martínez-Nicolás et al., 2021; Cho et al., 2022). This has spurred hundreds of studies applying Natural Language Processing (NLP) methods to transcribed cognitive assessments (Peled-Cohen and Reichart, 2024), primarily to extract linguistic markers (disfluencies, rephrasing, part-of-speech ratios, etc; Soni et al., 2021; Adhikari et al., 2021; Farzana et al., 2022a; Williams et al., 2021) and to detect cognitive decline.

In reality, however, the initial "detection" of dementia symptoms rarely begins with a clinician or structured cognitive assessments. Instead, it is often the individuals themselves or their close environment, who first notice signs of cognitive decline and initiates a medical evaluation (van Harten et al., 2018; Scharre, 2019; Jessen et al., 2020). This highlights the importance of understanding how dementia is perceived by non-experts, i.e., those who are not yet patients, caregivers, or clinicians. Identifying which linguistic behaviors are perceived as related to dementia can reveal where public intuition aligns with clinical insight, and where it falls short and can benefit from further education.

Additionally, nowadays, it is not just humans who can track linguistic changes and raise concerns. Adults over 55, an age group at increased risk for cognitive decline (Bai et al., 2022), now regularly use large language models (LLMs) (Smith, 2025). Given these models' extensive world knowledge and continuous linguistic signal from the user, one can imagine LLMs flagging subtle shifts in language to indicate early signs of cognitive im-

* Equal contribution.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 13834–13860
November 4-9, 2025

pairment. It is therefore important to understand which cues drive LLMs' perceptions.

Our paper explores the concept of *dementia perception by non-experts*: which cues lead humans and LLMs to perceive someone as cognitively impaired based on their language. To study this, we collected perceptions from 27 humans and 3 LLMs–LLaMA 3.1 (Grattafiori et al., 2024), GPT-4o (OpenAI, 2023), and Gemini-1.5-Pro (Team et al., 2024)–who were presented with 514 transcriptions of a spontaneous speech task (Cookie Theft picture descriptions from the Pitt corpus; Becker et al., 1994; further detailed in Section 3.1). Humans and LLMs were asked to give their best intuitive judgment as to whether each text was produced by someone healthy or cognitively impaired. Throughout this work, we use "humans" to refer to non-expert humans, and "clinicians" or "clinical diagnosis" to refer to expert medical judgment.

To analyze perceptions, we propose a 4-step explainable method, inspired by studies such as Badian et al. (2023), Balek et al. (2024), and Lissak et al. (2024): (1) design intuitive, human-centered features in consultation with domain experts; (2) extract these features using an LLM as an annotator, with quality control; (3) train inherently simple and interpretable logistic regression models to predict perceptions and clinical diagnosis; and (4) analyze coefficients to identify the linguistic cues influencing how dementia is perceived. Sections 4 to 7 outline the full method and results.

Our pipeline is rooted in high-level features that capture nuanced aspects of picture descriptions. These 38 binary features (Table 1, Appendix A), were developed in accordance with established literature and in collaboration with domain experts. The features span five categories aligned with cognitive processes involved in describing a picture: visual processing (e.g., "I see a boy and a girl"); reasoning (e.g., "he it about to fall"); verbal expression (e.g., disfluencies); emotional reaction (e.g., "poor kids"); and personal interactions (e.g., "Is that what you meant?"). This categorization allows for an analysis beyond individual features, and conclusions that may generalize beyond picture descriptions to other texts produced by patients.

Traditionally, extracting such high-level features would require extensive feature engineering and a dedicated algorithmic logic for each feature, posing a significant scalability challenge. Manual annotation is also impractical, as our dataset includes 38 features for hundreds of transcriptions, amounting to over 19,000 annotations. To address this, we use LLMs as annotators and validate their output with statistical tests to ensure quality comparable to human annotations (Calderon et al., 2025b).

Our analysis reveals that human judgments are highly inconsistent and show a tendency toward false negatives, i.e., labeling clinically diagnoses cases as healthy. Humans appear to rely on a narrow set of simple, objective features, and sometimes interpret cues in ways that contradict clinical patterns. Notably, when asked to describe which linguistic cues shaped their perception, they often reported features that align with our predefined set, reinforcing its validity. However, these self-reports do not match the actual decision patterns, suggesting that people do not rely on what they think they do. LLMs, while also prone to false negatives, appear to rely on a much richer feature set, including sentiment-related cues. This highlights their surprisingly nuanced use of language in this task.

To summarize, our contributions are as follows: (1) we propose an interpretable four-step method building on LLMs-as-annotators with statistical significance quality test for analyzing the linguistic features driving dementia perception; (2) we analyze the perceptions of non-experts (27 human annotators and three LLMs) and identify the linguistic behaviors they associate with dementia; (3) we examine the extent to which these behaviors overlap with features associated with clinical diagnosis and analyse human and LLM misperceptions.

We hope this study lays a foundation for future research on dementia perception from the perspective of all stakeholders. By shedding light on the linguistic cues that non-experts and LLMs rely on when assessing cognitive decline, we believe our findings can contribute to broader public awareness and support earlier detection. Finally, we aspire to encourage the interpretable and statistically grounded use of LLMs in sensitive domains, fostering interdisciplinary trust and real-world impact.

2 Related Work

NLP-Based Dementia Detection NLP is increasingly used in dementia research, typically applied to transcribed clinical assessments to detect cognitive decline or identify its linguistic markers (Peled-Cohen and Reichart, 2024). Commonly used datasets include the CCC corpus (Pope and Davis, 2011), the ADReSS and ADReSSo chal-

lenge sets (Luz et al., 2021a,b), and the popular Pitt corpus (Becker et al., 1994; detailed in Section 3.1). All four datasets provide transcribed speech from both healthy and cognitively impaired individuals.

These datasets have been used to extract linguistic markers such as syntactic complexity (Roark et al., 2007), idea density (Sirts et al., 2017), topic structure (Pompili et al., 2020), and meta-semantic terms, i.e., words expressing emotion or opinion (Choi et al., 2019). Disfluencies, pauses, and context shifts are other markers known to significantly influence model predictions (Kemper and Anagnopoulos, 1989; Adhikari et al., 2021; Farzana et al., 2022b). Markers are then used to train dementia detection algorithms, using traditional classifiers such as SVMs and Random Forests (Jarrold et al., 2014; Fraser et al., 2016b; Zhou et al., 2016) or transformers (Pappagari et al., 2020; Edwards et al., 2020; Balagopalan et al., 2021). Recently, LLMs have also been leveraged for feature extraction or embedding-based representations supporting dementia detection (Li et al., 2023; Liu et al., 2023; Bang et al., 2024; Botelho et al., 2024; BT and Chen, 2024; Koga et al., 2024; Latif and Kim, 2024; Runde et al., 2024). Despite these advances, automated detection has yet to be adopted in clinical settings, and early diagnosis still relies on self-reports from patients or their environment.

Explainable Dementia Detection In high-risk fields such as healthcare, explainability is crucial for establishing trust and encouraging the adoption of artificial intelligence systems (Adadi and Berrada, 2020). A wide range of interpretability paradigms is available to NLP practitioners (Stiglic et al., 2020; Calderon and Reichart, 2024; Viswan et al., 2024; Calderon et al., 2025a), and some dementia-related studies do leverage them. Karlekar et al. (2018b) analyzed model predictions using activation clustering and derivative saliency, while Ilias and Askounis (2022) examined text statistics using LIME (Ribeiro et al., 2016) to identify which words or phrases most influenced individual predictions. Others, such as Vimbi et al. (2024), use feature attribution methods that assign importance to individual features or tokens. While related in spirit, our approach goes further by attributing predictions to higher-level, cognitively grounded concepts, offering more interpretable and generalizable insights than assigning importance to raw inputs like tokens.

Societal Perception of Dementia Studies on how dementia is perceived by the general public primarily focus on the concept of stigma. Long recognized as a defining aspect of the dementia experience (Graham et al., 2003; Milne, 2010; Benbow and Jolley, 2012), stigma has been shown to negatively impact emotional well-being and contribute to delays in diagnosis and help-seeking (Swaffer, 2014; Gove et al., 2016; Nguyen and Li, 2020).

Few studies have used NLP to investigate this topic, and those that have focus almost exclusively on dementia portrayal in social media conversations (Oscar et al., 2017; Pilozzi and Huang, 2020; Tahami Monfared et al., 2022), revealing that patients are mocked or ridiculed. While social media is accessible and captures real-world language, it has key limitations: (1) scarce presence of dementia patients (Panzavolta et al., 2025), making their language hard to study; (2) content is often preplanned or edited; and (3) unlike the clinical data we use, social media data is not centered on cognitive abilities impacted by dementia, or designed to elicit narrative, reasoning, or spontaneous dialogue.

Our Novelty: To the best of our knowledge, this is the first work to examine how non-experts and LLMs perceive dementia through clinical cognitive assessments. We uniquely approach the task by introducing an interpretable method rooted in human-oriented, expert-guided high-level features, extracted using LLMs as annotators, and used to model perception. Unlike prior work, we do not focus on the prediction of clinical diagnosis but rather on understanding the non-experts: their reasoning, where their perceptions diverge from clinical judgment, and how intuition may be improved.

3 Cookie Theft Picture Descriptions

3.1 Background

In this section, we describe the data used to obtain and analyze perceptions. We rely on the Pitt corpus (Becker et al., 1994), a widely used dataset from the DementiaBank cohort (Lanzi et al., 2023), which includes longitudinal data from 98 healthy individuals and 196 dementia patients across varying stages and types, primarily AD. The corpus provides participant demographics, standardized cognitive scores and transcribed recordings of various linguistic tasks, such as the Cookie Theft picture description, which we focus on in our study.

Originally designed by Goodglass et al. (1983), the Cookie Theft picture description task is commonly used to elicit spontaneous speech in cogni-

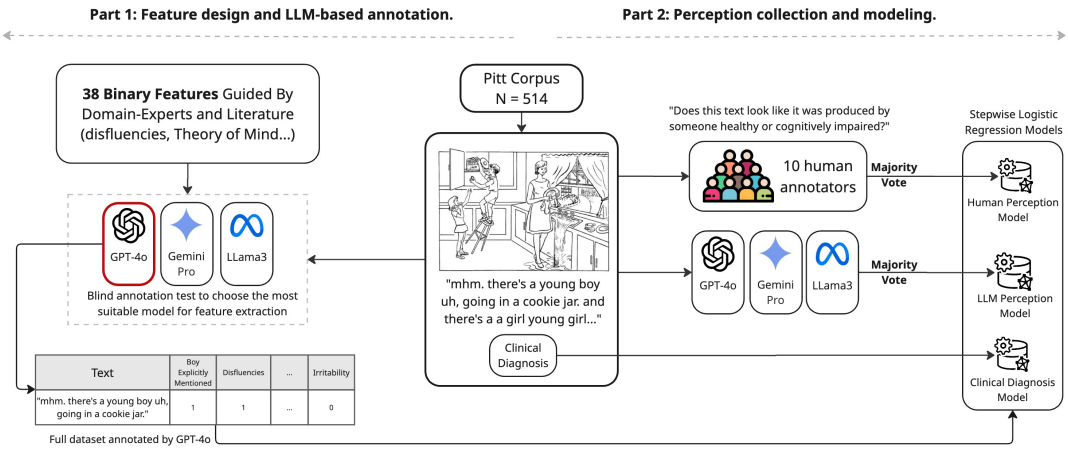

Figure 1: Illustration of the end-to-end methodological process.

tive assessments (Goren et al., 1992; Fraser et al., 2016a; Berube et al., 2022; Butala et al., 2022). In this task, participants are shown an image of a domestic scene (Figure 4) and asked to describe what they see, with their responses typically recorded, transcribed, analyzed, and scored. However, this task provides clinicians with more than just standardized scores, as they commonly use it to form *a general impression* of an individual's information processing, linguistic performance, motor speech function, and communicative ability. The fact that clinicians routinely rely on this task to make such holistic judgments highlights its practical value for our study of perception and intuition.

3.2 Preprocessing

We extract 514 Cookie Theft Picture descriptions and their corresponding clinical diagnosis ('Healthy Control', 'MCI', 'Possible AD', 'Probable AD', 'AD', and 'Other'). We then binarize these into two classes: 'Healthy' and 'Dementia', with the latter including all samples not labeled as 'Healthy Control'. This binary framing simplifies the task for non-expert annotators, since differentiating between the types and stages of cognitive decline is challenging even for trained clinicians.

As mentioned in Section 2, this corpus has been extensively used in NLP-focused dementia research. Building on these studies, we apply standard preprocessing to the Cookie Theft transcripts, separating the participant's speech from the interviewer's and removing extraneous characters and interview-specific annotations. For the remainder of the study, the input provided to both humans and LLMs consists solely of raw transcriptions, with no additional demographic information or clinical labels. This helps mitigate potential biases, such as assumptions based on the speaker's age, while also complying with the dataset's safety regulations.

4 Perception Annotations

4.1 Human Perception

We recruited 27 non-expert annotators to read the preprocessed picture descriptions and intuitively judge each as "Healthy" or "Dementia". Annotators received no prior instructions regarding which cues to consider when making their decisions (see Appendix B.1.1 for annotation guidelines). Generally, we focused on young non-expert adults, none have prior experience as dementia caregivers. Full demographic details about our annotators are presented in Table 2. We deliberately chose not to include other populations, such as caregivers or clinicians, as (a) expert and non-expert groups are likely to produce significantly different perception signals, prompting dedicated studies; and (b) young adults are a particularly relevant group for studying dementia perception, as they are increasingly likely to observe early signs of MCI in their close circles. Understanding what they perceive, and eventually helping them recognize those signs, is crucial.

Each description was labeled by 10 annotators, and the majority vote was used as the final perception label. No ties occurred in any of the samples. Annotations show a relatively low inter-annotator agreement (Fleiss' $\kappa = 0.28$), which is unsurprising given the inherent subjectivity of the task (Rottger et al., 2022). After completing the task, annotators were asked to describe any cues they noticed that may have influenced their decision.

Category	Definition	Example feature
Objective Interpretation features	Whether the speaker refers to elements that are indisputably visible in the picture	Does the speaker **mention the boy?** Y/N
Subjective Interpretation features	Whether the speaker presents higher-level interpretations that may or may not be accurate	Does the speaker **assume any sound was made**? Y/N
Linguistic features	How the speaker uses language	Does the speaker **use a rich vocabulary?** Y/N
Human Experience features	Whether the speaker expresses their own state or emotions	Does the speaker **express irritability**? Y/N
Interview Context features	Whether the text includes references to the interview situation itself	Does the speaker **ask the interviewer for a clarification**? Y/N

Figure 2: Feature categories, definitions, and examples.

4.2 LLM Perception

GPT-4o, LLaMA 3.1, and Gemini-1.5-Pro were provided with the same transcripts and also asked to provide their best judgment (see Appendix B.2 for the full prompt). Each model labeled the entire dataset, and we used their majority vote in our analysis. LLMs show stronger inter-annotator agreement than humans (Fleiss' $\kappa = 0.465$), perhaps due to their shared training data and structure.

5 Feature Extraction Methodology

5.1 Feature Design

Our core intuition is that when a text is perceived as produced by someone with dementia, judgment is not based on computational scores such as noun-to-verb ratio (Williams et al., 2023). Instead, it often stems from a gut feeling–whether the text feels informative or empty, comprised of rich or repetitive vocabulary, etc. We therefore aim to define features in a more intuitive manner. For example, to represent the noun-to-verb ratio, one can ask: "Did the speaker focus on actions over objects?". This framing makes the feature easier to interpret and potentially adopt as a guideline.

In collaboration with a neurologist and a neuropsychologist, we defined **38 binary features** that capture informative aspects of the picture descriptions in an intuitive manner, and are anchored in dementia research (see Table 1, Appendix A for all features, sources and examples). The features span five categories (Figure 2) aligned with cognitive processes involved in picture description: what is directly observed (**Objective Interpretation**); what is inferred or assumed beyond what is seen (**Subjective Interpretation**); how these observations are linguistically expressed (**Linguistic**); and the emotional or experiential states expressed throughout (**Human Experience** and **Interview Context**). Grouping features into broader categories enables higher-level analysis and more generalizable insights; Cummings (2019b), for example, found that related feature groups may be more reliable diagnostic markers than individual cues.

5.2 Feature Annotation

Building on studies such as He et al. (2024); Tan et al. (2024) and Badian et al. (2023), we use LLMs as annotators (Nahum et al., 2024; Calderon et al., 2025b). We experiment with LLaMA 3.1, GPT-4o, and Gemini-1.5-Pro. Each model evaluated one picture description at a time, and provided a binary Yes/No response for each feature. See the complete list of features and prompts in Appendix A.

To test whether different LLMs can reliably label the 38 binary features, we ran a blind annotation study. Three human annotators independently labeled a subset of the corpus (10 descriptions $\times$ 38 features, totaling $N = 380$ values per annotator). Inter-annotator agreement was solid, with Fleiss' $\kappa = 0.557$. We then applied the Alternative Annotator Test (Calderon et al., 2025b), a statistical method for evaluating whether LLMs can replace human annotators and for comparing different LLMs. GPT-4o performed best, with a 90% chance that its answers were as good as or better than those of humans. It outperformed Gemini-1.5-Pro (83%) and Llama-3.1 (70%). Notably, GPT-4o passed the test with a conservative threshold ($\varepsilon = 0.1$), limiting the acceptable disagreement between LLM and human annotations, as recommended by Calderon et al. (2025b). Given this statistical justification, we used GPT-4o to label all 38 binary features across our 514 descriptions, a total of 19,532 annotated values. Appendix C details the feature value distribution and internal correlations.

6 Perception Modeling

We use stepwise logistic regression, an inherently interpretable model suited for our binary outcomes (Dementia vs. Healthy). The stepwise approach simplifies the model by eliminating weak predic-

Category	Feature Name	Sources	Example
Linguistic	Circumlocution (Wordiness)	Nicholas and Brookshire (1993); Kavé and Dassa (2018); Cho et al. (2021)	"a stool which is about and he he is getting a cookie"
	Grammatical Inaccuracies	Croisile et al. (1996); Fraser et al. (2016a)	"they be stealing"
	Introduction	Ortiz-Perez et al. (2023)	"This is a family scene. A boy..."
	Naming Characters	Kempler and Zelinski (1994); Kempler and Goral (2008)	"Johnny here is..."
	Non-specific language	Nicholas and Brookshire (1993); Kavé and Dassa (2018); Cummings (2019a); Cho et al. (2021)	"The thing there.."
	Rich vocabulary	Fraser et al. (2016a); Kavé and Dassa (2018); Cho et al. (2021); Williams et al. (2023)	"She is probably daydreaming. The faucet..."
	Short sentences	Mueller et al. (2016); Fraser et al. (2016a), Kavé and Levy (2003); Forbes-McKay and Venneri (2005)	"Stealing cookies. Falling stool."
	Starts with interjection	Karlekar et al. (2018a); Ortiz-Perez et al. (2023)	"Well. Mother is.."
	Disfluencies	Nicholas and Brookshire (1993); Szatloczki et al. (2015a); Mueller et al. (2018); Kumar et al. (2022)	"She is uh, uh, umm.."
	Self corrections	Rudzicz et al. (2014); Mueller et al. (2018)	"Mother's washing, uh, drying dishes"
	Actions over objects	Williams et al. (2023); Kavé and Dassa (2018)	"She's Washing, they're stealing" (vs. "I see curtains, shoes")
Objective Interpretation	Boy explicitly mentioned	Croisile et al. (1996); Nicholas and Brookshire (1993); Cummings (2019a)	"A young lad here"
	Girl explicitly mentioned	Croisile et al. (1996); Nicholas and Brookshire (1993); Cummings (2019a)	"Sister is laughing"
	Kitchenware attention	Croisile et al. (1996); Nicholas and Brookshire (1993); Cummings (2019a)	"Two cups and a plate"
	Mother explicitly mentioned	Croisile et al. (1996); Nicholas and Brookshire (1993); Cummings (2019a)	"Mother is..."
	Outside mentioned	Croisile et al. (1996); Nicholas and Brookshire (1993); Cummings (2019a)	"The garage is..."
Subjective Interpretation	Assumes sound	Yorkston and Beukelman (1980)	"Sister is saying..."
	Cause and effect	Croisile et al. (1996); Cummings (2019a)	"Boy is about to fall"
	Criticise characters	Yorkston and Beukelman (1980)	"Mother is doing a bad job"
	Empathy	Chow et al. (2023); Demichelis et al. (2020)	"Poor boy is about to fall"
	Improbable interpretation	Nicholas and Brookshire (1993)	"Pineapple jar"
	Mentioning many objects	Nicholas and Brookshire (1993); Williams et al. (2023)	"Shoes, dress, cupboard handles..."
	Other characters mentioned	Nicholas and Brookshire (1993)	"Baby is crying in the other room"
	Probable speculation	Nicholas and Brookshire (1993)	"Sink is probably clogged"
	Theory of Mind- Boy	Cummings (2019a); Demichelis et al. (2020); Zegarra-Valdivia et al. (2023)	"Boy is handing a cookie to his sister"
	Theory of Mind- Girl	Cummings (2019a); Demichelis et al. (2020); Zegarra-Valdivia et al. (2023)	"Sister is laughing..."
	Theory of Mind- Mother	Cummings (2019a); Demichelis et al. (2020); Zegarra-Valdivia et al. (2023)	"Mother is thinking about something..."
	Weather conditions mentioned	Nicholas and Brookshire (1993)	"It's a sunny day"
Human Experience	Checking Previously Said	Cipriani et al. (2013)	"Did I already say that?"
	Continuing after saying done	Baylis et al. (2004)	"That's all I can see. They boy is..."
	Hesitation	Rudzicz et al. (2014); Ortiz-Perez et al. (2023)	"I'm not sure, maybe stealing cookies"
	Irritability	Cummings (2019a)	"Can I go now?"
	Lightheartedness	Cummings (2019a); Granitsas (2020)	"Wowie! This is a mess [laughs]"
	Reiterating ideas	Croisile et al. (1996); Stegmann et al. (2021); Kumar et al. (2022)	"Boy is stealing. Mother is washing dishes. Boy is taking cookies"
	Sad depressed despaired	Cummings (2019a)	"Please stop, I can't take this anymore"
	Self limitations	Rudzicz et al. (2014); Nicholas and Brookshire (1993)	"I don't see anything else."
	Vision difficulties	Lawrence et al. (2009)	"Let me put on my glasses"
Interview Setting	Clarification required	Rudzicz et al. (2014); Karlekar et al. (2018a)	"Did you want only actions?"

Table 1: Full feature list, literature sources and examples.

tors, and is well-suited for small datasets. We selected logistic regression since, under standard assumptions, particularly when relevant confounders are included, coefficients can roughly approximate causal effects (Cinelli et al., 2024). In the absence of a formal causal graph, randomized control trials, etc., regression offers a close approximation to causal insights, albeit with caution.

Since our goal is to extract interpretable insights, we fit the model to the full dataset. We evaluate models using McFadden's R^2 (McFadden, 1972), a standard goodness-of-fit metric for logistic regression that quantifies how well the predictors explain the outcome relative to a null model, based on log-likelihood ratio (Shtatland et al., 2002; Smith and McKenna, 2013). Values between 0.2-0.4 are considered strong and comparable to R^2 values of 0.6-0.8 in linear regression (Domencich and McFadden, 1975; Louviere et al., 2000). To comply with standard NLP practices, we also conduct a complementary analysis of predictive performance and report standard metrics in Appendix D.3.

After training we analyze coefficients of features with significant effects (Hosmer Jr et al., 2013). As an example, the feature *disfluencies* has a coefficient of $\beta = 2.16$ in the LLM perception model. In logistic regression, this means the log-odds of the model assigning a "Dementia" label increase by 2.16 when *disfluencies* are present, holding all else constant. Converting this to an odds ratio, the model is approximately 9 times more likely ($e^{2.16} \approx 8.67$) to assign a "Dementia" label than a "Healthy" one. In contrast, *rich vocabulary* has $\beta = -1.46$ in the LLM perception model, corresponding to an odds ratio of about $e^{-1.46} \approx 0.23$, meaning the model is *less likely* to assign a "Dementia" label when this feature is present. In general, positive coefficients indicate a shift toward the "Dementia" label (1) and negative toward "Healthy" (0), with larger values reflecting greater impact.

7 Results

7.1 Modeling Perceptions and Diagnosis

Figure 3 illustrates the coefficient values of statistically significant features associated with human perception, LLM judgments, and clinical diagnosis. For all coefficients and corresponding significance values, see Table 3, Appendix A.

McFadden's R^2 values indicate a good model fit for clinical diagnosis (0.209) and a very strong fit for LLM dementia perception (0.527), suggesting that our model and features captures a reliable underlying signal. Human perception, however, was harder to model, with McFadden's R^2 = 0.058.

Only a small set of straightforward features are significantly associated with human perception: *non-specific language, short sentences, girl explicitly mentioned*, and *mother explicitly mentioned*, all significantly associated with clinical diagnosis as well. Additional features significantly linked to clinical diagnosis include *actions over objects, other characters mentioned*, and *weather conditions mentioned*. Interestingly, while clinicians associate *short sentences* with dementia, non-experts interpret them as a sign of cognitive health.

Coefficient analysis reveals that all three judgment types are significantly associated with linguistic features such as *non-specific language*, as well as objective interpretations (e.g., whether the boy, girl, or mother is explicitly mentioned). LLMs rely on a broader range of features and categories than clinical diagnoses, showing greater sensitivity to subjective interpretation cues, such as using Theory of Mind (describing others' emotions, intentions, or thoughts) or referring to characters not present in the picture. LLMs also place greater emphasis on emotional expression (*lightheartedness, self-limitations, sad-depressed-despaired*).

7.2 Diving Into Misperceptions

We begin with a data analysis, examining how human and LLM judgments align with clinical diagnoses. Figure 3 presents the corresponding confusion matrices. It also includes a Venn diagram showing the overlap between clinically diagnosed cases and those perceived as dementia by humans and LLMs. Among the 283 clinically diagnosed dementia cases, humans correctly identified 57%. LLMs, though more conservative in assigning the dementia label, matched 60%. Some misalignment is expected, as our non-experts rely solely on picture descriptions, while clinical diagnoses draw on a broader range of signals. Errors by both humans and LLMs followed a similar pattern: for LLMs, 70% of errors were false negatives (i.e., missing clinically diagnosed cases), while 30% were false positives. Humans showed a similar trend, with a 65-35 ratio, suggesting that both groups have blind spots and room for improvement.

Next, we examine two key subsets: cases where humans disagreed with both LLMs and clinicians ($n = 121$), e.g., perceiving dementia when both

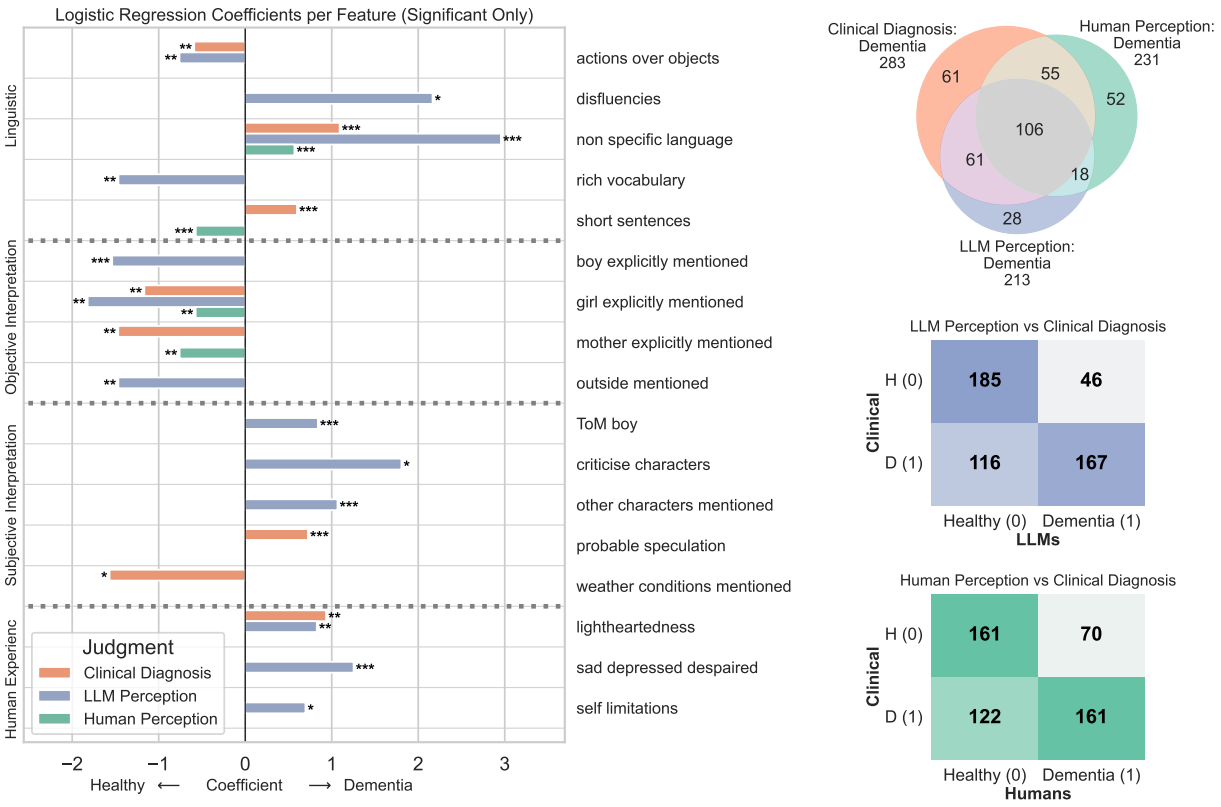

Figure 3: Main results from the logistic regression coefficient analysis, and perception disagreements. **Left:** Statistically significant features associated with clinical diagnosis, LLM perception, and human perception. Colors indicate the source of judgment; bar direction reflects the sign of the logistic regression coefficient (right = dementia, left = healthy). Dotted lines separate feature categories (Linguistic, Objective Interpretation, etc.). Significance levels: * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$. **Top Right:** Overlap between clinically diagnosed cases and those perceived as dementia by humans and/or LLMs. Of 283 diagnosed, 98 (green, teal, and purple) were missed by humans, LLMs, or both. **Bottom Right:** Confusion matrices showing alignment between perceptions and diagnosis.

others judged healthy, and cases where LLMs disagreed with both humans and clinicians ($n = 88$). In this analysis, we go beyond comparing perceptions and diagnoses. Instead, we ask: what cues are so prominent that both clinicians and LLMs capture them, but humans miss? And which cues do clinicians and humans detect, but LLMs overlook? We then train a stepwise logistic regression model on each subset to predict human and LLM misperceptions. As before, positive coefficients indicate a shift toward the "Dementia" label, and negative toward "Healthy"– but here, they reflect *false dementia* and *false healthy*, respectively.

The human misperception model shows a strong fit (McFadden's $R^2 = 0.6$) and reveals two systematic patterns in human misjudgment: **(1) Features used differently in misperceived texts:** *non-specific language* received a positive coefficient in the full dataset of 514 samples, indicating that, in general, humans associate non-specific language with dementia. However, in the smaller misperception subset, this feature received a *negative coefficient*, indicating a shift toward the "Healthy" label (a *false healthy* judgment, in this context). This suggests that human reliance on this cue is inconsistent-typically treated as a sign of dementia but sometimes overlooked or misinterpreted.

(2) Features emerging only in misperceptions: Some features are not significant in the overall human perception model but become significant in misperception cases. Namely, *rich vocabulary, actions over objects, outside mentioned*, and *boy mentioned* all show a shift toward false healthy judgments, i.e., mistakenly labeling clinically diagnosed dementia cases as healthy. *Lightheartedness*, on the other hand, becomes significant in the misperception model with a shift toward false dementia, suggesting that humans may sometimes associate this cue with cognitive decline. This indicates that certain features, while not consistently relied upon in the general model, may exert misleading influence in specific cases of misjudgment.

The LLM disagreement analysis revealed a quasi-perfect separation: whenever features such as *hesitation*, *reiterating idea*, *grammatical inaccuracies*, or *disfluencies* received a value of '0' (i.e., were not present in the text) the LLM majority vote always labeled the text as "Healthy". Other features, namely *sadness, many objects mentioned, mother mentioned, other characters mentioned* were sparse in the disagreement set– only seven texts expressed sadness, five mentioned other characters, etc. This suggests LLM errors may also stem from data sparsity.

7.3 Self-reported Human Rationale

At the end of the task, without having seen our predefined features, annotators answered an open-ended question: Did you notice any patterns in the text that helped with your judgment? These retrospective reports, provided by 18 of 27 participants, may not fully capture real-time reasoning, yet, they offer valuable insights into the behaviors that annotators noticed and said they relied on.

Upon manual review, we found that 65% of the self-reported cues closely align with our predefined feature set, an encouraging result that suggests our features are naturally noticed by humans. The most frequently cited were *reiterating ideas*, *disfluencies*, and *improbable interpretation*, mentioned by 10, 7, and 5 annotators, respectively. Notably, all four cues revealed as significant for human perception in our coefficient analysis (Figure 3) were also self-reported by annotators. A full analysis, including details of newly emergent features from participant responses, is provided in Appendix E.

8 Discussion and Insights

8.1 Human Perception of Dementia

Modeling human perception proved particularly challenging. The low McFadden's R^2 observed for the human perception model likely reflects not only model limitations but also the inherently noisy and inconsistent nature of human judgment. This is evidenced by low inter-annotator agreement and conflicting annotator rationales, both in the features they cited and the direction of their presumed influence. Notably, these self-reported rationales did not align with the statistically significant features identified by the model, suggesting that people may not rely on what they think they do.

Our model shows that humans often rely on a narrow set of simplistic cues, whereas clinical diagnoses draw on a broader range of signals, including sentiment-related cues, which even LLMs managed to capture. Moreover, humans' misinterpretation of cues– associating *short sentences* with descriptions from healthy participants, unlike clinical diagnosis– underscores the importance of education about a broader and more nuanced set of linguistic indicators of dementia.

8.2 LLMs' Perception

Unlike humans, LLMs rely on a surprisingly wide set of features, spanning four of our five categories. This use of varied signals may reflect the extensive prior knowledge embedded in their training data, for instance, "learning" that Theory of Mind is linked to certain types of dementia (Bora et al., 2015). This reinforces the importance of rich background knowledge and training to help non-experts and clinicians attend to a broader range of cues.

Both humans and LLMs show a tendency toward false negatives, misjudging dementia cases as healthy. However, in the case of LLMs, a clear pattern emerges: when no linguistic difficulties are present, the model is extremely prone to assigning a "healthy" label. Thus, LLMs may be quick to judge a speaker as cognitively healthy if no linguistic dysfunction is apparent, potentially leading to missed signs of early cognitive impairment–an important limitation to recognize and study.

9 Conclusions

This study explores how non-expert humans and LLMs perceive dementia in transcribed picture descriptions, and how their perceptions align with clinical diagnoses. We present an inherently interpretable method using high-level, expert-guided features annotated by GPT-4o, followed by logistic regression and coefficient analysis. Ultimately, our work highlights: (1) the inherent difficulty of this task, both in perceiving dementia through text alone and in modeling such a complex phenomenon; (2) the importance of educating both humans and LLMs to recognize a broader range of linguistic signals to improve early detection; and (3) the value of interpretable NLP for advancing research and care across all dementia stakeholders.

10 Limitations

Our study makes deliberate modeling and methodological choices to enable a focused analysis of how dementia is perceived. We use the Pitt corpus,

which contains several hundred samples-small by NLP standards but relatively large in clinical research. It is the standard dataset in studies on NLP and dementia (Peled-Cohen and Reichart, 2024), making our choice consistent with established practice. Future work should apply our methodology to datasets of varying sizes and sources, given its promising results on the Pitt corpus.

Perceptions were collected from 27 individuals of a specific demographic, which limits representativeness. We focused on non-experts, as early signs of decline are often noticed by family members without formal training, to capture baseline intuitions and their limitations. Annotators were proficient but not native English speakers, which may have influenced judgments. Future studies should involve broader populations, including caregivers, clinicians, and participants with more diverse age, education, and language backgrounds.

We emphasize binary features for interpretability. While quantitative features (e.g., verb counts) are widely studied, we aimed for a more human-oriented framework translatable into actionable insights. Ordinal scales (e.g., rating disfluency from 1–5) proved less reliable and harder to align between human and LLM annotations. Future work could extend this approach with quantitative metrics such as idea density or lexical diversity.

Extracting features with LLMs introduces known challenges, such as prompt sensitivity and inconsistency, especially for features reflecting human intuition. We mitigated this by conducting human evaluation and applying a statistical test (Calderon et al., 2025b), but future work should expand this approach to larger annotator groups and a broader range of models.

We chose logistic regression for its interpretability and clear coefficient structure, which provide directionally meaningful insights into how individual features relate to perception labels. However, logistic regression assumes linearity and may struggle with complex interactions, and its predictive power may be limited compared to larger models. Future work should explore how to balance interpretability with more expressive modeling.

Another dimension can shape human perception of dementia: long-term personal relationships. In our study, judgments were based on a single written description by an unfamiliar person. In real-life settings, however, impressions are likely to be formed through repeated personal interactions. We assume that participants' decisions were guided by internalized notions of what a "typical" healthy individual sounds like, shaped by general assumptions, societal stereotypes, or portrayals in popular media. This would likely differ if participants were evaluating someone they knew well, drawing on a sense of the individual's baseline behavior and personality changes over time.

Finally, we use textual picture descriptions only, which do not encompass other modalities that are known to play an important role in discourse (such as motor abilities). While some features (e.g., disfluencies) may indirectly reflect physical difficulties in speech production, these are primarily expressed through vocal modalities that are absent from text-based data. In addition, non-verbal gestures (e.g., facial expressions, gaze direction) and prosodic features (e.g., intonation, pitch, speech rate) hold valuable information about cognitive and emotional functioning. These signals are integral to how people perceive and interpret communication in everyday interactions. Future work could address this limitation by incorporating multi-modal inputs such as audio and video.

11 Ethical Considerations

Privacy. Data confidentiality is a major concern when dealing with clinical data. The dataset used in this study, the Pitt corpus from DementiaBank, is available for research purposes and has already been annonymized. We used only the de-identified transcriptions of the Cookie Theft task, without any accompanying metadata to elicit perceptions. As a result, all data processed in this work was entirely free of personally identifiable information (PII) and handled in accordance with ethical standards.

LLM bias. LLMs are increasingly used in healthcare research, yet are known to inherit biases from their training data. These biases often reflect societal, linguistic, and cultural norms rather than clinically grounded principles (Busch et al., 2025). They may stem from data imbalance or reliance on proxy variables (Chen et al., 2018; Obermeyer et al., 2019). In our work, we explicitly frame model outputs as perceptions to underscore this limitation: LLM outputs do not reflect clinical truth but rather represent patterns the models are able to extract from language data.

Lack of professional support when relying on observations by non-experts. When relying on

non-experts, particularly close family members or friends, ethical concerns arise regarding their role and the potential impact on relationships. While increasing public awareness of early indicators is valuable, non-experts are not equipped (nor should they be expected) to deliver a diagnosis or communicate life-changing information. Suspicions raised by loved ones can unintentionally undermine trust, introduce stigma, or lead to unjustified restrictions on the autonomy of the person in question. Moreover, such situations can place a significant emotional burden on both parties, as the mere suspicion of a serious and threatening condition like dementia may evoke fear, helplessness, and anxiety. For the person suspected of cognitive decline, being perceived as impaired by someone close can be deeply distressing and lead to feelings of isolation. On the other hand, cases of false negatives, i.e., where early signs are overlooked and diagnosis is delayed, may result in guilt or self-blame for not having acted sooner. These emotional and relational consequences highlight the need for professional involvement.

The responsibility for diagnosing dementia and delivering difficult news should lie with trained clinicians, who are equipped to do so with sensitivity and to offer appropriate support and care planning. While non-experts may play a valuable role in prompting medical evaluation when they observe concerning changes, they should not carry the burden of diagnostic responsibility. Crucially, failing to notice subtle early symptoms should not be regarded as a personal failure. Dementia is a complex and gradually evolving condition, and even experienced professionals can find its early detection challenging.

References

Amina Adadi and Mohammed Berrada. 2020. Explainable ai for healthcare: From black box to interpretable models. In *Embedded Systems and Artificial Intelligence*, volume 1045 of *Advances in Intelligent Systems and Computing*, pages 327–337. Springer Singapore.

Surabhi Adhikari, Surendrabikram Thapa, Priyanka Singh, Huan Huo, Gnana Bharathy, and Mukesh Prasad. 2021. A comparative study of machine learning and NLP techniques for uses of stop words by patients in diagnosis of alzheimer's disease. In *International Joint Conference on Neural Networks, IJCNN 2021, Shenzhen, China, July 18-22, 2021*, pages 1–8. IEEE.

Galit Agmon, Sameer Pradhan, Sharon Ash, Naomi Nevler, Mark Liberman, Murray Grossman, and Sunghye Cho. 2024. Automated Measures of Syntactic Complexity in Natural Speech Production: Older and Younger Adults as a Case Study. *Journal of Speech, Language, and Hearing Research*, 67(2):545–561. Publisher: American Speech-Language-Hearing Association.

Alzheimer's Disease International. 2025. Dementia statistics. https://www.alzint.org/about/dementia-factsfigures/dementia-statistics/. [Accessed 2025-05-08].

Yael Badian, Yaakov Ophir, Refael Tikochinski, Nitay Calderon, Anat Brunstein Klomek, Eyal Fruchter, and Roi Reichart. 2023. Social media images can predict suicide risk using interpretable large language-vision models. *The Journal of Clinical Psychiatry*, 85(1):50516.

Wei Bai, Pan Chen, Hong Cai, Qinge Zhang, Zhaohui Su, Teris Cheung, Todd Jackson, Sha Sha, and Yu-Tao Xiang. 2022. Worldwide prevalence of mild cognitive impairment among community dwellers aged 50 years and older: a meta-analysis and systematic review of epidemiology studies. *Age and ageing*, 51(8):afac173.

Aparna Balagopalan, Benjamin Eyre, Jessica Robin, Frank Rudzicz, and Jekaterina Novikova. 2021. Comparing pre-trained and feature-based models for prediction of alzheimer's disease based on speech. *Frontiers in aging neuroscience*, 13:635945.

Vojtěch Balek, Lukáš Sỳkora, Vilém Sklenák, and Tomáš Kliegr. 2024. Llm-based feature generation from text for interpretable machine learning. *arXiv preprint arXiv:2409.07132*.

Jeong-Uk Bang, Seung-Hoon Han, and Byung-Ok Kang. 2024. Alzheimer's disease recognition from spontaneous speech using large language models. *ETRI Journal*, 46(1):96–105.

Gordon C Baylis, Leslie L Baylis, and Christopher L Gore. 2004. Visual neglect can be object-based or scene-based depending on task representation. *Cortex*, 40(2):237–246.

James T Becker, François Boiler, Oscar L Lopez, Judith Saxton, and Karen 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.

Susan Mary Benbow and David Jolley. 2012. Dementia: stigma and its effects. *Neurodegenerative Disease Management*, 2(2):165–172.

Shauna K. Berube, Emily Goldberg, Shannon M. Sheppard, Alexandra Zezinka Durfee, Delaney Ubellacker, Alexandra Walker, Colin M. Stein, and Argye E. Hillis. 2022. An Analysis of Right Hemisphere Stroke Discourse in the Modern Cookie Theft Picture.

American Journal of Speech-Language Pathology, 31(5S):2301–2312. Publisher: American Speech-Language-Hearing Association.

Gerhard Blanken, Jürgen Dittmann, J. Christian Haas, and Claus-W. Wallesch. 1987. Spontaneous speech in senile dementia and aphasia: Implications for a neurolinguistic model of language production. *Cognition*, 27(3):247–274.

Emre Bora, Mark Walterfang, and Dennis Velakoulis. 2015. Theory of mind in behavioural-variant frontotemporal dementia and alzheimer's disease: a meta-analysis. *Journal of Neurology, Neurosurgery & Psychiatry*, 86(7):714–719.

Catarina Botelho, John Mendonça, Anna Pompili, Tanja Schultz, Alberto Abad, and Isabel Trancoso. 2024. Macro-descriptors for alzheimer's disease detection using large language models. In *Proc. Interspeech*, pages 1975–1979.

Balamurali BT and Jer-Ming Chen. 2024. Performance assessment of chatgpt versus bard in detecting alzheimer's dementia. *Diagnostics*, 14(8):817.

Felix Busch, Lena Hoffmann, Christopher Rueger, Elon HC van Dijk, Rawen Kader, Esteban Ortiz-Prado, Marcus R Makowski, Luca Saba, Martin Hadamitzky, Jakob Nikolas Kather, et al. 2025. Current applications and challenges in large language models for patient care: a systematic review. *Communications Medicine*, 5(1):26.

Ankur Butala, Kevin Li, Aathman Swaminathan, Susan Dunlop, Yekaterina Salnikova, Bronte Ficek, Brandon Portnoff, Michael Harper, Bailey Vernon, Bela Turk, Zoltan Mari, and Alexander Pantelyat. 2022. Parkinsonics: A Randomized, Blinded, Cross-Over Trial of Group Singing for Motor and Nonmotor Symptoms in Idiopathic Parkinson Disease. *Parkinson's Disease*, 2022:4233203.

Nitay Calderon, Liat Ein-Dor, and Roi Reichart. 2025a. Multi-domain explainability of preferences. *arXiv preprint arXiv:2505.20088*.

Nitay Calderon and Roi Reichart. 2024. On behalf of the stakeholders: Trends in nlp model interpretability in the era of llms. *arXiv preprint arXiv:2407.19200*.

Nitay Calderon, Roi Reichart, and Rotem Dror. 2025b. The alternative annotator test for llm-as-a-judge: How to statistically justify replacing human annotators with llms. *CoRR*, abs/2501.10970.

Irene Chen, Fredrik D Johansson, and David Sontag. 2018. Why is my classifier discriminatory? *Advances in neural information processing systems*, 31.

Sunghye Cho, Katheryn Alexandra Quilico Cousins, Sanjana Shellikeri, Sharon Ash, David John Irwin, Mark Yoffe Liberman, Murray Grossman, and Naomi Nevler. 2022. Lexical and acoustic speech features relating to alzheimer disease pathology. *Neurology*, 99(4):e313–e322.

Sunghye Cho, Naomi Nevler, Sharon Ash, Sanjana Shellikeri, David J. Irwin, Lauren Massimo, Katya Rascovsky, Christopher Olm, Murray Grossman, and Mark Liberman. 2021. Automated analysis of lexical features in frontotemporal degeneration. *Cortex*, 137:215–231.

Jinho D Choi, Mengmei Li, Felicia Goldstein, and Ihab Hajjar. 2019. Meta-semantic representation for early detection of alzheimer's disease. In *Proceedings of the First International Workshop on Designing Meaning Representations*, pages 82–91.

Tiffany E Chow, Christina R Veziris, Nidhi Mundada, Alexis I Martinez-Arroyo, Joel H Kramer, Bruce L Miller, Howard J Rosen, Maria Luisa Gorno-Tempini, Katherine P Rankin, William W Seeley, et al. 2023. Medial temporal lobe tau aggregation relates to divergent cognitive and emotional empathy abilities in alzheimer's disease. *Journal of Alzheimer's disease*, 96(1):313–328.

Carlos Cinelli, Andrew Forney, and Judea Pearl. 2024. A crash course in good and bad controls. *Sociological Methods & Research*, 53(3):1071–1104.

Gabriele Cipriani, Marcella Vedovello, Martina Ulivi, Angelo Nuti, and Claudio Lucetti. 2013. Repetitive and stereotypic phenomena and dementia. *American Journal of Alzheimer's Disease & Other Dementias®*, 28(3):223–227.

Bernard Croisile, Bernadette Ska, Marie-Josee Brabant, Annick Duchene, Yves Lepage, Gilbert Aimard, and Marc Trillet. 1996. Comparative Study of Oral and Written Picture Description in Patients with Alzheimer's Disease. Technical report.

Louise Cummings. 2019a. Describing the Cookie Theft picture. *Pragmatics and Society*, 10(2):153–176.

Louise Cummings. 2019b. Describing the cookie theft picture: sources of breakdown in alzheimer's dementia. *Pragmatics and Society*, 10(2):153–176.

Olivia P. Demichelis, Sarah P. Coundouris, Sarah A. Grainger, and Julie D. Henry. 2020. Empathy and Theory of Mind in Alzheimer's Disease: A Meta-analysis. *Journal of the International Neuropsychological Society*, 26(10):963–977.

Thomas A Domencich and Daniel McFadden. 1975. Urban travel demand-a behavioral analysis. Technical report.

Erik Edwards, Charles Dognin, Bajibabu Bollepalli, Maneesh Kumar Singh, and Verisk Analytics. 2020. Multiscale system for alzheimer's dementia recognition through spontaneous speech. In *Interspeech*, pages 2197–2201.

Shahla Farzana, Ashwin Deshpande, and Natalie Parde. 2022a. How you say it matters: Measuring the impact of verbal disfluency tags on automated dementia detection. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 37–48,

Dublin, Ireland. Association for Computational Linguistics.

Shahla Farzana, Ashwin Deshpande, and Natalie Parde. 2022b. How you say it matters: Measuring the impact of verbal disfluency tags on automated dementia detection. In *Proceedings of the 21st workshop on biomedical language processing*, pages 37–48.

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, Jed A. Meltzer, and Frank Rudzicz. 2016a. Linguistic Features Identify Alzheimer's Disease in Narrative Speech. *Journal of Alzheimer's disease: JAD*, 49(2):407–422.

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

Harold Goodglass, Edith Kaplan, and Sandra Weintraub. 1983. BDAE: The Boston Diagnostic Aphasia Examination. *PA: Lea & Febiger*.

Anat Goren, Carol Swindell, and Arifulla Khan. 1992. Expressive language characteristics of schizophrenic subjects with different medication histories. *Journal of Neurolinguistics*, 7(1):67–90.

Dianne Gove, Murna Downs, MJFJ Vernooij-Dassen, and Neil Small. 2016. Stigma and gps' perceptions of dementia. *Aging & mental health*, 20(4):391–400.

Nori Graham, James Lindesay, Cornelius Katona, José Manoel Bertolote, Vincent Camus, John RM Copeland, Carlos A de Mendonça Lima, Michel Gaillard, Marie Christine Gély Nargeot, John Gray, et al. 2003. Reducing stigma and discrimination against older people with mental disorders: a technical consensus statement. *International journal of geriatric psychiatry*, 18(8):670–678.

Dean Anthony Granitsas. 2020. All laughter is nervous: An anxiety-based understanding of incongruous humor. *HUMOR*, 33(4):625–643.

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

Xingwei He, Zhenghao Lin, Yeyun Gong, A-Long Jin, Hang Zhang, Chen Lin, Jian Jiao, Siu-Ming Yiu, Nan Duan, and Weizhu Chen. 2024. Annollm: Making large language models to be better crowdsourced annotators. In *The North American Chapter of the Association for Computational Linguistics*.

D. B. Hier, K. Hagenlocker, and A. G. Shindler. 1985. Language disintegration in dementia: Effects of etiology and severity. *Brain and Language*, 25(1):117–133.

David W Hosmer Jr, Stanley Lemeshow, and Rodney X Sturdivant. 2013. *Applied logistic regression*. John Wiley & Sons.

Loukas Ilias and Dimitris Askounis. 2022. Explainable identification of dementia from transcripts using transformer networks. *IEEE Journal of Biomedical and Health Informatics*, 26(8):4153–4164.

William Jarrold, Bart Peintner, David Wilkins, Dimitra Vergryi, Colleen Richey, Maria Luisa Gorno-Tempini, and Jennifer Ogar. 2014. Aided diagnosis of dementia type through computer-based analysis of spontaneous speech. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 27–37.

Frank Jessen, Rebecca E Amariglio, Rachel F Buckley, Wiesje M van der Flier, Ying Han, José Luis Molinuevo, Laura Rabin, Dorene M Rentz, Octavio Rodriguez-Gomez, Andrew J Saykin, et al. 2020. The characterisation of subjective cognitive decline. *The Lancet Neurology*, 19(3):271–278.

Sweta Karlekar, Tong Niu, and Mohit Bansal. 2018a. Detecting Linguistic Characteristics of Alzheimer's Dementia by Interpreting Neural Models. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 701–707, New Orleans, Louisiana. Association for Computational Linguistics.

Sweta Karlekar, Tong Niu, and Mohit Bansal. 2018b. Detecting linguistic characteristics of alzheimer's dementia by interpreting neural models. In *Proceedings of NAACL-HLT*, pages 701–707.

Gitit Kavé and Ayelet Dassa. 2018. Severity of Alzheimer's disease and language features in picture descriptions. *Aphasiology*, 32(1):27–40.

Gitit Kavé and Yonata Levy. 2003. Morphology in picture descriptions provided by persons with alzheimer's disease. *Journal of speech, language, and hearing research*, 46(2):341–352.

Susan Kemper and Cheryl Anagnopoulos. 1989. Language and aging. *Annual review of applied linguistics*, 10:37–50.

Daniel Kempler and Mira Goral. 2008. Language and dementia: Neuropsychological aspects. *Annual review of applied linguistics*, 28:73–90.

Daniel Kempler and Elizabeth M Zelinski. 1994. Language in dementia and normal aging. *Dementia and normal aging*, pages 331–365.

Shunsuke Koga, Nicholas B Martin, and Dennis W Dickson. 2024. Evaluating the performance of large language models: Chatgpt and google bard in generating differential diagnoses in clinicopathological conferences of neurodegenerative disorders. *Brain Pathology*, 34(3):e13207.

Yash Kumar, Piyush Maheshwari, Shreyansh Joshi, and Veeky Baths. 2022. Ml-based analysis to identify speech features relevant in predicting alzheimer's disease. In *Proceedings of the 8th International Conference on Computing and Artificial Intelligence*, pages 207–213.

Alyssa M Lanzi, Anna K Saylor, Davida Fromm, Houjun Liu, Brian MacWhinney, and Matthew L Cohen. 2023. Dementiabank: Theoretical rationale, protocol, and illustrative analyses. *American Journal of Speech-Language Pathology*, 32(2):426–438.

Atif Latif and Jihie Kim. 2024. Evaluation and analysis of large language models for clinical text augmentation and generation. *IEEE Access*.

Vanessa Lawrence, Joanna Murray, Sube Banerjee, et al. 2009. "out of sight, out of mind": a qualitative study of visual impairment and dementia from three perspectives. *International psychogeriatrics*, 21(3):511–518.

Changye Li, Jacob Solinsky, Trevor Cohen, and Serguei Pakhomov. 2024. A curious case of retrogenesis in language: Automated analysis of language patterns observed in dementia patients and young children. *Neuroscience Informatics*, 4(1):100155.

Rumeng Li, Xun Wang, and Hong Yu. 2023. Two directions for clinical data generation with large language models: data-to-label and label-to-data. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2023, page 7129.

Shir Lissak, Yaakov Ophir, Refael Tikochinski, Anat Brunstein Klomek, Itay Sisso, Eyal Fruchter, and Roi Reichart. 2024. Bored to death: Artificial intelligence research reveals the role of boredom in suicide behavior. *Frontiers in psychiatry*, 15:1328122.

Ming Liu, Richard Beare, Taya Collyer, Nadine Andrew, and Velandai Srikanth. 2023. Leveraging natural language processing and clinical notes for dementia detection. In *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pages 150–155.

Jordan J Louviere, David A Hensher, and Joffre D Swait. 2000. *Stated choice methods: analysis and applications*. Cambridge university press.

Saturnino Luz, Fasih Haider, Sofia de la Fuente, Davida Fromm, and Brian MacWhinney. 2021a. Detecting cognitive decline using speech only: The adresso challenge. In *INTERSPEECH 2021*. ISCA.

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

Israel Martínez-Nicolás, Thide E Llorente, Francisco Martínez-Sánchez, and Juan José G Meilán. 2021. Ten years of research on automatic voice and speech analysis of people with alzheimer's disease and mild cognitive impairment: a systematic review article. *Frontiers in Psychology*, 12:620251.

Daniel McFadden. 1972. Conditional logit analysis of qualitative choice behavior.

Alison Milne. 2010. The 'd'word: Reflections on the relationship between stigma, discrimination and dementia.

Kimberly D. Mueller, Rebecca L. Koscik, Bruce P. Hermann, Sterling C. Johnson, and Lyn S. Turkstra. 2018. Declines in connected language are associated with very early mild cognitive impairment: Results from the Wisconsin Registry for Alzheimer's Prevention. *Frontiers in Aging Neuroscience*, 9(JAN).

Kimberly Diggle Mueller, Rebecca L Koscik, Lyn S Turkstra, Sarah K Riedeman, Asenath LaRue, Lindsay R Clark, Bruce Hermann, Mark A Sager, and Sterling C Johnson. 2016. Connected language in late middle-aged adults at risk for alzheimer's disease. *Journal of Alzheimer's Disease*, 54(4):1539–1550.

Laura L. Murray. 2010. Distinguishing clinical depression from early Alzheimer's disease in elderly people: Can narrative analysis help? *Aphasiology*, 24(6-8):928–939. Publisher: Routledge _eprint: https://doi.org/10.1080/02687030903422460.

Omer Nahum, Nitay Calderon, Orgad Keller, Idan Szpektor, and Roi Reichart. 2024. Are llms better than reported? detecting label errors and mitigating their effect on model performance. *CoRR*, abs/2410.18889.

Trang Nguyen and Xiaoming Li. 2020. Understanding public-stigma and self-stigma in the context of dementia: A systematic review of the global literature. *Dementia*, 19(2):148–181.

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

Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Sylvester O Orimaye, Jojo SM Wong, Karen J Golden, Chee P Wong, and Ireneous N Soyiri. 2017. Predicting probable alzheimer's disease using linguistic deficits and biomarkers. *BMC bioinformatics*, 18:1–13.

David Ortiz-Perez, Pablo Ruiz-Ponce, David Tomás, Jose Garcia-Rodriguez, M. Flores Vizcaya-Moreno, and Marco Leo. 2023. A Deep Learning-Based Multimodal Architecture to predict Signs of Dementia. *Neurocomputing*, 548:126413.

Nels Oscar, Pamela A Fox, Racheal Croucher, Riana Wernick, Jessica Keune, and Karen Hooker. 2017. Machine learning, sentiment analysis, and tweets: An examination of alzheimer's disease stigma on twitter. *Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, 72(5):742–751.

Andrea Panzavolta, Andrea Arighi, Emanuele Guido, Luigi Lavorgna, Francesco Di Lorenzo, Alessandra Dodich, Chiara Cerami, et al. 2025. Patient-related barriers to digital technology adoption in alzheimer disease: Systematic review. *JMIR aging*, 8(1):e64324.

Raghavendra Pappagari, Jaejin Cho, Laureano Moro-Velazquez, and Najim Dehak. 2020. Using state of the art speaker recognition and natural language processing technologies to detect alzheimer's disease and assess its severity. In *Interspeech*, pages 2177–2181.

Lotem Peled-Cohen and Roi Reichart. 2024. A systematic review of nlp for dementia-tasks, datasets and opportunities. *arXiv preprint arXiv:2409.19737*.

Alexander Pilozzi and Xudong Huang. 2020. Overcoming alzheimer's disease stigma by leveraging artificial intelligence and blockchain technologies. *Brain sciences*, 10(3):183.

Anna Pompili, Alberto Abad, David Martins de Matos, and Isabel Pavão Martins. 2020. Pragmatic aspects of discourse production for the automatic identification of alzheimer's disease. *IEEE Journal of Selected Topics in Signal Processing*, 14(2):261–271.

Charlene Pope and Boyd H Davis. 2011. Finding a balance: The carolinas conversation collection.

Martin Prince, Renata Bryce, Cleusa Ferri, et al. 2011. *World Alzheimer Report 2011: The benefits of early diagnosis and intervention*. Alzheimer's Disease International London.

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

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Brian Roark, Margaret Mitchell, and Kristy Hollingshead. 2007. Syntactic complexity measures for detecting mild cognitive impairment. In *Biological, translational, and clinical language processing*, pages 1–8.

Paul Rottger, Bertie Vidgen, Dirk Hovy, Janet Pierrehumbert, et al. 2022. Two contrasting data annotation paradigms for subjective nlp tasks. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.

Frank Rudzicz, Leila Chan Currie, Andrew Danks, Tejas Mehta, and Shunan Zhao. 2014. Automatically identifying trouble-indicating speech behaviors in alzheimer's disease. In *Proceedings of the 16th International ACM SIGACCESS Conference on Computers & Accessibility*, ASSETS '14, pages 241–242, New York, NY, USA. Association for Computing Machinery.

Benjamin S Runde, Ajit Alapati, and Nicolas G Bazan. 2024. The optimization of a natural language processing approach for the automatic detection of alzheimer's disease using gpt embeddings. *Brain Sciences*, 14(3):211.

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

Ernest S Shtatland, Ken Kleinman, and Emily M Cain. 2002. One more time about r2 measures of fit in logistic regression. *NESUG 15 proceedings*, 15:222–226.

John R. Sims, Jennifer A. Zimmer, Cynthia D. Evans, Ming Lu, Paul Ardayfio, JonDavid Sparks, Alette M. Wessels, Sergey Shcherbinin, Hong Wang, Emel Serap Monkul Nery, Emily C. Collins, Paul Solomon, Stephen Salloway, Liana G. Apostolova, Oskar Hansson, Craig Ritchie, Dawn A. Brooks, Mark Mintun, Daniel M. Skovronsky, and TRAILBLAZER-ALZ 2 Investigators. 2023. Donanemab in Early Symptomatic Alzheimer Disease: The TRAILBLAZER-ALZ 2 Randomized Clinical Trial. *JAMA*, 330(6):512–527.

Kairit Sirts, Olivier Piguet, and Mark Johnson. 2017. Idea density for predicting alzheimer's disease from transcribed speech. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 322–332.

Jeff Smith. 2025. The future of senior care may be ai. `https://www.retirementliving.com/aging-with-ai`. [Accessed 2025-05-08].

Thomas J Smith and Cornelius M McKenna. 2013. A comparison of logistic regression pseudo r2 indices. *Multiple Linear Regression Viewpoints*, 39(2):17–26.

Aradhana Soni, Benjamin Amrhein, Matthew Baucum, Eun Jin Paek, and Anahita Khojandi. 2021. Using verb fluency, natural language processing, and machine learning to detect alzheimer's disease. In *2021 43rd annual international conference of the IEEE engineering in medicine & biology society (EMBC)*, pages 2282–2285. IEEE.

Gabriela M. Stegmann, Shira Hahn, Julie Liss, Visar Berisha, and Kimberly D. Mueller. 2021. Comparison of remote and in-person digital speech-based

measures of cognition. *Alzheimer's & dementia : the journal of the Alzheimer's Association*, 17:e056438.

Gregor Stiglic, Primoz Kocbek, Nino Fijacko, Marinka Zitnik, Katrien Verbert, and Leona Cilar. 2020. Interpretability of machine learning-based prediction models in healthcare. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(5):e1379.

Kate Swaffer. 2014. Dementia: Stigma, language, and dementia-friendly.

Greta Szatloczki, Ildiko Hoffmann, Veronika Vincze, Janos Kalman, and Magdolna Pakaski. 2015a. Speaking in Alzheimer's Disease, is That an Early Sign? Importance of Changes in Language Abilities in Alzheimer's Disease. *Frontiers in Aging Neuroscience*, 7:195.

Greta Szatloczki, Ildiko Hoffmann, Veronika Vincze, Janos Kalman, and Magdolna Pakaski. 2015b. Speaking in alzheimer's disease, is that an early sign? importance of changes in language abilities in alzheimer's disease. *Frontiers in aging neuroscience*, 7:195.

Amir Abbas Tahami Monfared, Yaakov Stern, Stephen Doogan, Michael Irizarry, and Quanwu Zhang. 2022. Stakeholder insights in alzheimer's disease: natural language processing of social media conversations. *Journal of Alzheimer's Disease*, 89(2):695–708.

Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024. Large language models for data annotation and synthesis: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 930–957.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Argonde C van Harten, Michelle M Mielke, Dana M Swenson-Dravis, Clinton E Hagen, Kelly K Edwards, Rosebud O Roberts, Yonas E Geda, David S Knopman, and Ronald C Petersen. 2018. Subjective cognitive decline and risk of mci: The mayo clinic study of aging. *Neurology*, 91(4):e300–e312.

Malvika Verma and Robert J Howard. 2012. Semantic memory and language dysfunction in early alzheimer's disease: a review. *International journal of geriatric psychiatry*, 27(12):1209–1217.

Viswan Vimbi, Noushath Shaffi, and Mufti Mahmud. 2024. Interpreting artificial intelligence models: a systematic review on the application of lime and shap in alzheimer's disease detection. *Brain Informatics*, 11(1):10.

Vimbi Viswan, Noushath Shaffi, Mufti Mahmud, Karthikeyan Subramanian, and Faizal Hajamohideen. 2024. Explainable artificial intelligence in alzheimer's disease classification: A systematic review. *Cognitive Computation*, 16(1):1–44.

Sebastian Wankerl, Elmar Nöth, and Stefan Evert. 2017. An n-gram based approach to the automatic diagnosis of alzheimer's disease from spoken language.

Adina Williams, Ryan Cotterell, Lawrence Wolf-Sonkin, Damián Blasi, and Hanna Wallach. 2021. On the relationships between the grammatical genders of inanimate nouns and their co-occurring adjectives and verbs. *Transactions of the Association for Computational Linguistics*, 9:139–159.

Eric Williams, Catherine Theys, and Megan McAuliffe. 2023. Lexical-semantic properties of verbs and nouns used in conversation by people with Alzheimer's disease. *PLOS ONE*, 18(8):e0288556.

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.

Jonathan Adrian Zegarra-Valdivia, Myrthe Gwen Rijpma, Tal Shany-Ur, Joel H. Kramer, Bruce L. Miller, and Katherine P Rankin. 2023. Cognitive and emotional theory of mind in dementia. Impact on real life behaviors. *Alzheimer's & Dementia*, 19(S4):e067855.

Luke Zhou, Kathleen C Fraser, Frank Rudzicz, et al. 2016. Speech recognition in alzheimer's disease and in its assessment. In *Interspeech*, volume 2016, pages 1948–1952.

A Full Feature List and Prompts

Following is the complete list of 38 binary features we defined, divided per category. For each feature, we present its relevant prompt, provided to GPT-4o-2024-08-06. For all model prompting (GPT, LLaMA, Gemini), we used "temperature": 1.0. We instructed the LLM to respond only with "yes" or "no." We then extracted the distribution over the first generated token and computed the probabilities P("yes") and P("no") by summing over any case-insensitive variants of "yes" or "no." The final prediction for each feature or perception is determined by whether P("yes")>P("no"). This process yields deterministic predictions, as we only consider the first token and do not perform sampling.

A.1 Linguistic Features

These are features representing the speaker's use of language.

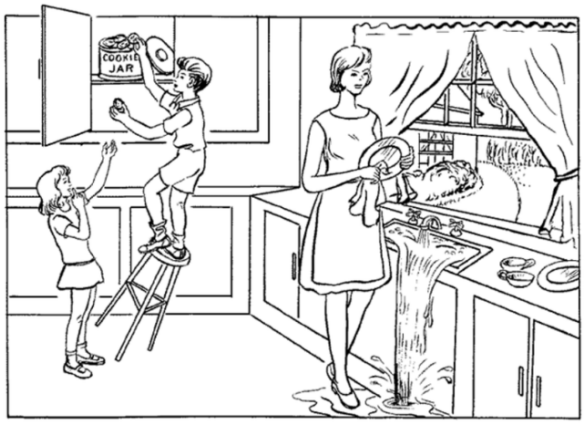

Figure 4: Cookie Theft Picture.

Circumlocution: "You are analyzing a transcribed description of the Cookie Theft Picture from the Boston Diagnostic Aphasia Examination (BDAE). Your task is to determine whether the speaker shows signs of circumlocution excessive or indirect speech that makes it harder to follow the message. Answer "yes" only if the text contains noticeable wordiness or overly long explanations that make the message less clear or harder to follow. This includes: * Describing something in a roundabout way without clearly naming it (e.g., "the thing you use to dry stuff" instead of "towel"). * Over-explaining simple observations (e.g., "and then I guess what she's doing is maybe she's holding something and it looks like it could be" instead of directly saying "she's holding a plate"). Disregard normal conversational markers like- * Slight elaboration or descriptive phrasing. * Minor hesitations or informal language. Please answer yes or no only, no explanations.

Grammatical inaccuracies: "Does the text contain noticeable grammatical errors (e.g., "the boy going", "they is washing dishes", "the boy does climbing")? Does not qualify: (1) Incomplete sentences without clear content, (2) Hesitations and disfluencies (e.g., "the boy went, uh, boy") unless they result in a grammatical mistake. Please answer yes or no only, no explanations."

Introduction: "Is the first sentence a general introduction to the picture? Qualifies: (1) A statement that sets up the scene before describing specific elements (e.g., "This is a family scene.", "This is a mess."). (2) A general remark about the picture, even if opinionated or informal (e.g., "What a mess in this kitchen!"). Does not qualify: Starting with fillers ('uh. the boy is...') or discourse markers ('Well, this here is a cookie jar.'). (2) Jumping straight into the description without an introduction (e.g., "Okay, the mother is cooking.") (3) Making statements about the task ('I'll start now.') (4) asking questions/clarifications directed at the interviewer (e.g., "Can I start?", "You wanted only actions?"). Please answer yes or no only, no explanations."

Naming characters: "Does the speaker call any of the characters by any name? (e.g., "Johnny here is stealing cookies") please answer with "yes" or "no" only, no explanation."

Non-specific language: "You are analyzing a transcribed description of the Cookie Theft Picture from the Boston Diagnostic Aphasia Examination (BDAE). Your task is to determine whether the speaker exhibits a strong reliance on non-specific language–including both non-specific nouns (e.g., excessive pronoun use or vague descriptions) and non-specific verbs (e.g., "doing something", "going over there")–in a way that suggests difficulty retrieving specific words. A description should be flagged as "yes" only if: There is a strong and persistent pattern of avoiding specific words, making the description vague or unclear. Example: "He's on that thing, and she's doing something over there." Example: "The thing is falling, and she's making the water go." The avoidance applies to both nouns and verbs, reducing clarity significantly. The pattern is excessive, indicating possible word retrieval difficulty rather than normal variation in speech. Examples (Flag as "yes"): "That one is up on the thing, and the other one is doing something." "She's making it go while he's messing with that over there." "They're all doing things, and it's happening." Examples (Flag as "no"): "The boy is standing on the stool, and the mother is washing dishes." (Clear noun and verb use.) "She is drying the dishes while he is reaching for cookies." (Some pronoun use but still specific enough.) Response Format: Return "yes" if the description contains an excessive reliance on non-specific language (both nouns and verbs) to the point that it strongly suggests word retrieval difficulty. Return "no" if pronoun or vague verb use is within normal variation or does not significantly impair clarity. Please return "yes" or "no" only, no intermediate calculations or explanations."

Rich vocabulary: "You are analyzing a spoken or written description of the Cookie Theft Picture

from the Boston Diagnostic Aphasia Examination (BDAE). Your task is to determine whether the text demonstrates vocabulary richness. Respond with "yes" if the description includes specific and varied content words, such as precise nouns, vivid verbs, coumpund nouns or relatively unique phrases (e.g., "overflowing", "spigot", "faucet", "summertime", "drapes"). These indicate the speaker is using a rich and varied vocabulary beyond basic object naming. Respond with "no" if the description is dominated by vague, repetitive, or basic words, lacks elaboration, or relies heavily on simple naming without further detail. Frequent disfluencies or abandoned thoughts may also suggest limited vocabulary use. Return "yes" or "no" only. No explanations."

Short sentences: "Does the speaker primarily use short, independent clauses with minimal subordination or conjunctions? Responses should favor simple sentence structures (e.g., 'The boy is in the cookie jar. The mother spilled the water. The window is open.') over complex or compound sentences with multiple dependent clauses (e.g., 'The young fellow is standing on a stool which is getting ready to fall while he's handing the top of the cookie jar.'). Please answer 'yes' or 'no' only, no explanations."

Starts with interjection: "Does this text start with discourse markers (e.g., 'well', 'like', 'you know', 'I mean', 'okay'), verbal fillers (e.g., 'let me think...', 'the thing is...', 'how do I put this...'), or repetition (e.g., 'so, so...', 'you know, you know...')? Does not qualify: (1) starting with empty speach and fillers (e.g., 'uh', 'umm'), (2) starting with content words like "cookie jar". Please answer yes or no only, no explanation."

Disfluencies: "You are analyzing a transcribed description of the Cookie Theft Picture from the Boston Diagnostic Aphasia Examination (BDAE). Your task is to determine whether the text contains a noticeable amount of disfluencies. These include: (1) Repeated starts (e.g., "the girl... uh... the mother") signaling a change of mind or confusion. (2) Incomplete or abandoned utterances (e.g., "the girl went the boy"). (3) Unusual or excessive fillers and hesitations (e.g., many uses of ="uh" or "um") suggesting trouble forming thoughts rather than conversational rhythm. Remember that this is transcribed speech. So you should disregard normal spoken features, such as: (1) Occasional "uh" or "um" used naturally. (2) Minor self-corrections or restarts that do not break sentence coherence. Answer only with "yes" or "no", with "yes" reserved for texts presenting excessive disfluencies."

Self corrections: "Does the text include self-corrections? i.e., the speaker says something, identify they have a mistake, then immediately corrects themselves (e.g., "a kid uh, getting in the... falling off the stool", "the girl, uh, mother"). Hesitations and fillers by themselves (e.g., "the girl, uh, uh... is climbing") do not count. Please answer yes or no only, no explanations."

Actions over objects: "Does the speaker focus on actions rather than objects, showing a tendency toward describing movement and processes over static elements (e.g., "reaching", "tipping", "putting" instead of "cookie jar", "stool", "apron", "shoes"). Please answer yes or no only, no explanations."

A.2 Objective Interpretation features

These features represent whether the speaker refers to elements that are indisputably visible in the picture.

Boy explicitly mentioned: "This text describes a picture including a boy in the context of climbing a stool and reaching for a cookie jar. Is this boy explicitly, unmistakable mentioned? It qualifies in both direct terms (e.g., 'boy', 'lad', 'man', 'youngster', 'junior') and indirect terms ('he', 'him', 'Johnny'), provided that the context is clear. If the speaker describes children in plural (e.g., 'children', 'kids'), it qualifies. Please answer "yes" or "no" only, no explanations. If you are uncertain if the speaker refers to the boy, please say "no"."

Girl explicitly mentioned: "This text describes a picture including a girl in the context of standing next to a boy, plotting to steal cookies, putting a finger to her lips, or reaching for the boy. Is this girl explicitly, unmistakably mentioned? It qualifies in both direct terms (e.g., 'girl', 'sister') and indirect terms ('her', 'she'), provided that the context unambiguously refers to the girl in a way that is evident in the picture. If the speaker describes children in plural (e.g., 'children', 'kids'), it qualifies. If you are uncertain whether the speaker refers to the girl, say "no". Otherwise, answer "yes" only."

Kitchenware attention: "Your task is to determine whether the speaker pays attention to kitchenwear items (e.g., tap, towels, cups, plates) Posi-

tive example: ”cups on the counter and dish on the counter”, “washing this plate”. Mentioning “dishes” as a general term, or as part of the mother’s (she’s drying dishes) does not count, since it does not focus on which dishes. “cookie jar” does not qualify as kitchenwear items. Please answer with “yes” or “no” only, no explanation.”

Mother explicitly mentioned: “This text is a description of the Cookie Theft Picture. The picture includes a mother in the context of washing and drying dishes around a sink with spilling water. She is wearing a dress and shoes, handling dishes, her feet are in water, and she is looking outside the window. Is this character explicitly, unmistakably mentioned? Both direct terms (mother, lady, woman, model) and indirect terms (she, her) qualify, given that the context of water, dishes, etc is clear. Even if the speaker portrays uncertainty (e.g., “I guess”), explicitly stating “mother” still qualifies. Any female characters mentioned in the contexts of dishes or standing in water is the mother. Any female mentioned in context of cookie jar is not the mother so does not qualify.”

Outside mentioned: “Does the speaker mention the physical outside surroundings of the house? (i.e. bushes, trees, path, sun) Weather conditions (e.g. wind, summer) do not count. Please answer with “yes” or “no” only, no explanation.”

A.3 Subjective Interpretation features

These features represent whether the speaker presents higher-level interpretations that may or may not be accurate.

Assumes sound: “Does the speaker explicitly describe speaking, or any noises being made (such as speaking, shushing, birds chirping, a kettle whistling, or other sounds)? Qualifies: mentions of sounds OR sound-inducing actions (e.g., talking, laughing) produced by characters or objects in the scene. Do not assume a sound is present based on an event (e.g., dropping a plate) unless the speaker explicitly states that a noise occurred. Please answer with “yes” or “no” only, no explanation.”

Cause and effect: “Does the text contain an inference or prediction about something probable that is happening now or is about to happen, based directly on the scene? Answer “yes” if the text includes a logical inference or prediction stemming clearly from the picture (e.g., “soon there will be a mess”, “the boy is going to fall”, “the boy is falling”). The inferred event must be a direct consequence of what is visible in the image. Answer “no” if the statement (1) does not contain inference or prediction, (2) makes assumptions about imagined events unrelated to the immediate scene (e.g., “the husband will be back from work”), or (3) is too unclear without an obvious cause-and-effect link to the picture. Respond with “yes” or “no” only. No explanation.”

Criticise characters : “Does the speaker explicitly criticize the characters in the scene? (e.g., “she is irresponsible”, “I would spank that girl”, “they are reckless”). Criticism includes negative judgment or disapproval directed at what the characters are doing, how they are behaving, or the choices they are making (e.g., “that looks dangerous!”). Does not qualify: general frustration towards the task itself (“this is stupid”, “can I leave now” etc.) please answer ’yes’ or ’no’ only, no explanations.”

Empathy: “Does the speaker use words to communicate empathy toward the characters in the picture? Answer ’yes’ if the text includes ANY explicit expressions of emotional concern, sympathy, or understanding (e.g., ’poor mama,’ ’that must hurt,’ ’she looks worried,’ or ’the kids are nicely dressed’). Only direct emotional expressions count! Does not count: descriptive statements (e.g., ’the boy is going to fall and hurt himself’). Please answer ’yes’ or ’no’ only, no explanations.”

Improbable interpretation: “Does this text include anything entirely unrelated to the Cookie Theft image? Return “yes” if you are confident that the description includes elements that does not correspond to anything that could reasonably be inferred from the scene, like an impossible or completely fabricated item or thing.”

Mentioning many objects: “Does the text describe the scene primarily by listing a large number of inanimate objects (e.g., shoes, clothes, dishes, sink, cookie jar, curtains, windows, cupboard, lid) rather than focusing on people and their actions? Answer “yes” if the description is dominated by objects, details, and static elements rather than actions or interactions. Answer “no” if the focus is primarily on people and their actions, even if some objects are mentioned. Respond with “yes” or “no” only. No explanation.”

Other characters mentioned: "Does the text include human characters that are related to the scene but are not the mother (explicitly or implied), girl (explicitly or implied), or boy (explicitly or implied)? A mention qualifies even if the speaker later questions, dismisses, or negates the relationship (e.g., 'This wouldn't be the grandmother' and 'I don't see a husband' still qualify). Please answer with 'yes' or 'no' only, no explanations."

Probable speculation: "Does the speaker describe something that is not explicitly visible in the picture but is a reasonable interpretation based on the scene? Answer "yes" if the speaker makes an assumption that cannot be directly proven or disproven but makes sense in context (e.g., "it's a nice day outside" because the window is open and people are in short sleeves, or "the husband is at work" in a family setting). Answer "no" if: (1) the text is strictly based on what is visible in the image, without adding interpretations beyond the scene, (2) the text is completely unclear, (3) the inference is completely nonsensical. Respond with "yes" or "no" only. No explanation."

ToM boy: "Does the speaker use Theory of Mind when describing the boy? Includes: Assuming his mood or intentions (e.g., "poor boy", "he's generous"), Subjective interpretations of his actions (e.g., assuming he is stealing, meaning it's forbidden), Inferring readiness or preparation to act (e.g., "he's about to give her a cookie"), References to the boy explicitly or by other terms ('Johnny', 'man', 'kid', 'him', 'brother', etc.), Plural references that apply ToM to both children ("they are stealing") Does not qualify: Objective, indisputable descriptions of the boy's actions (e.g., "taking/reaching for a cookie") Speculations about future occurrences (e.g., "he is about to fall") that do not involve intent or mental states."

ToM girl: "Does the speaker use Theory of Mind (ToM) when describing the girl? This includes projecting her mood, mental state, intentions, or subjective interpretations of her actions (e.g., assuming she is happy, sneaky, has a certain motivation, is saying something). It also qualifies if ToM is applied to both kids in plural ("they are stealing"). Does not qualify: ToM applied to the mother. If 'she' or 'her' is used, determine from context whether it refers to the mother or the girl. please answer with "yes" or "no" only, no explanation."

ToM mother: "Does the speaker use Theory of Mind (ToM) when describing the mother? This includes projecting her mood, mental state, intentions, or subjective interpretations of her actions (e.g., assuming she is absent-minded, inattentive, unconcerned about her children, or responsible for the mess). ToM includes descriptions implying the mother is thinking about or focusing on something specific (e.g., "she is looking outside", "she is unaware of the water") rather than purely describing physical actions indisputibly portrayed in the picture. ToM applied to the girl does not count. If 'she' or 'her' is used, determine from context whether it refers to the mother or the girl."

Weather conditions mentioned: "Does the speaker mention any weather conditions or seasons of the year? (i.e. summer, windy, a hot day, etc.) Please answer with "yes" or "no" only, no explanation."

A.4 Human Experience:

These features represent whether the speaker expresses their own state or emotions.

Checking previously said: "Does the speaker explicitly check if they've already mentioned something in their description? Qualify: direct questions or statements such as "Did I already say...?", that indicate they are actively verifying whether they have said something before. Does not qualify: Self-corrections or rewording a previous statement. Does not qualify: reiterating ideas ("I already told you that..."). Only checking whether something was already said. Please answer with "yes" or "no" only, no explanations."

Continuing after saying done: "A contradiction is when someone claims they are done describing the picture (e.g., "that's all", "I'm done", "that's all I can see") but afterwards continues to provides more knowledge about the picture (e.g. continues talking about the jar, house, etc). If you note such a contradiction, return "yes". Otherwose, return "no". Return only "yes" or "no", no explanations or intermediate calculations."

Hesitation: "Does the speaker explicitly, verbally express hesitation in their own description? Qualifies: (1) Direct phrases such as "I'm not sure", "I don't know." (2) Fragmented, self-questioning speech (e.g., "uh... which one?" "this... thing?"). (3) Self-directed questioning about how to describe

something (e.g., "now what would I say about them?" "how do I put this?"). Does not qualify: (1) Speculative statements (e.g., "could be", "he will probably fall"). (2) Words expressing a certain level of certainty like "probably", "I guess", "I think" (3) Statements of inability (e.g., "I can't see anything else", "I can't make out any action"). (4) Questions to the interviewer (e.g., "is that it?" "did I do it right?"). Please answer "yes" or "no" only, no explanation."

Irritability: "Does the speaker show indisputable irritation? please answer with "yes" or "no" only, no explanation."

Lightheartedness: "Does the speaker explicitly convey humor or lightheartedness through their tone or choice of words? This includes the speaker laughing, joking, playful expressions (e.g., "gee whiz", "oh boy", "golly", "oh goody"), or an amused tone. The description of a chaotic or absurd scene does not qualify unless the speaker's words or manner explicitly suggest they find it humorous or lighthearted. please answer with "yes" or "no" only, no explanation."

Reiterating idea: "Does the speaker repeat the same idea multiple times within the text? i.e., reintroducing a previously mentioned element after discussing other details (e.g., idea A, idea B, idea C, idea A) please answer with "yes" or "no" only, no explanation."

Sad depressed despaired: "Does the speaker sound depressed, sad or despaired? please answer with "yes" or "no" only, no explanation."

Self limitations: "Does the speaker explicitly express a personal limitation in their ability to complete the task, like self-deprecation? This includes statements indicating that they cannot see, understand, or describe something (e.g., "I can't see anything else", "I don't know what that is", "I can't do this", "I give up."). Do not qualify: (1) Expressions of frustration without implying a personal limitation (e.g., "this is dumb", "this is annoying") (2) questions or requests to stop (e.g., "Can I stop now?" "Do I have to keep going?") (3) general finishing statements like "that's all", "I guess that's it". please answer with "yes" or "no" only, no explanation)."

Vision difficulties: "Can it be understood from the text that the speaker has vision impairment? Either explicitly ("where are my glasses") or implicitly ("what does it say there?") please answer with "yes" or "no" only, no explanation."

Variable	Value
N	27
Sex	Female: 16 (59.3%) Male: 11 (40.7%)
Age (years)	26.6 ± 3.56
Years of education	16.55 ± 1.82
Mother Tongue	Hebrew: 25 (92.6%) Russian: 1 (3.7%) Spanish: 1 (3.7%)
Familiarity with dementia	(A) None: 5 (18.5%) (B) Basic awareness via friends/media: 11 (40.7%) (C) Personal interaction with relatives: 11 (40.7%)

Table 2: Participant Demographics and Background.

A.5 Interview Context Features

This feature represents whether the text includes references to the interview situation itself.

Clarification required: "Please create a list of sentences from this text that contain questions or exchange of words with someone. For each, rank 1-5 if it makes sense that this was said to the interviewer, or spoken to oneself. If any sentence received over 4, please return yes. Otherwise please return no. Return only yes/no, no explanations or intermediate results."

B Dementia Perception Annotation

B.1 Human Annotators

B.1.1 Annotation Guidelines

"You will be presented with 180 descriptions of the Cookie Theft Picture (figure 4 attached).

The speakers (either healthy or with dementia) were asked to describe the picture with as many details as they can. Their responses were recorded, then transcribed from voice to text. The transcripts may include repetitions, incomplete words, nonverbal gestures (laughter, cough. . .) etc. For each transcript you read, state your intuition: Does this text feel like it was spoken by a Dementia patient, or a Healthy Control? Given what you know / heard / imagine about Dementia / Alzheimer's patients, please mark "0" if this description sounds like it's from a healthy person, or "1" if you sense it came from a dementia patient.

There is no right or wrong- we are interested in your instincts. We are examining how cognitive decline is perceived by the general population.

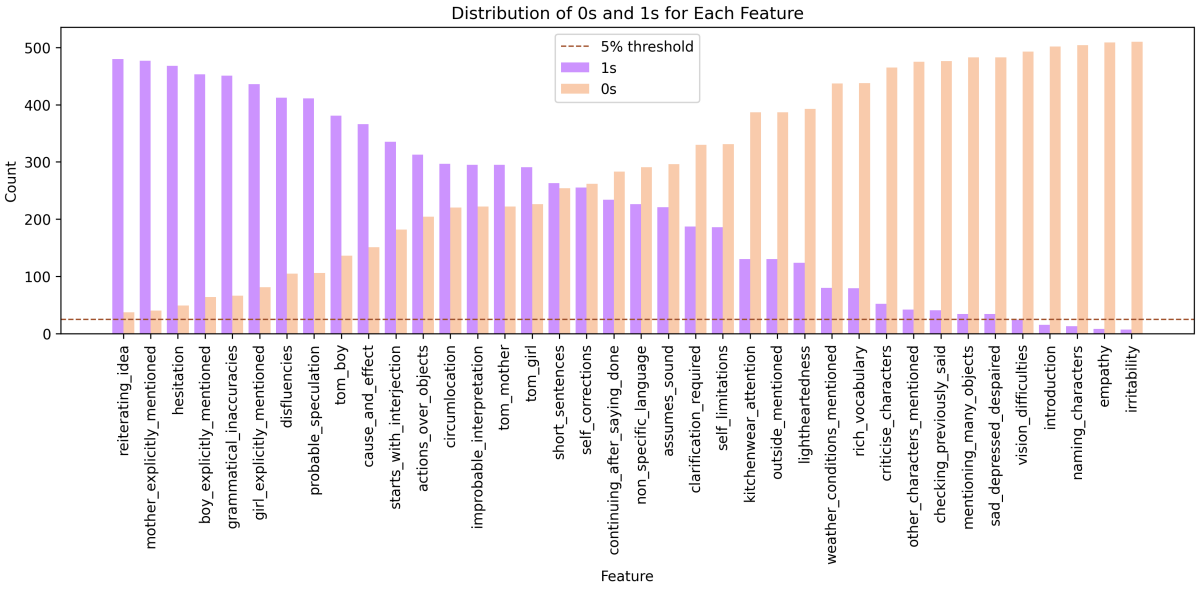

Figure 5: Value distribution across features. Features marked as 'yes' in fewer than 5% of samples (indicated by the dotted line) were excluded from our analyses.

This can influence people to notice warning signs in their loved ones, prompting them to seek medical attention. It can also reveal biases around the disease. Remember, we are not asking for a clinical diagnosis- just your opinion. Good Luck!"

B.1.2 Human Annotator Demographics

Table 2 presents the full demographic details of our human annotators. The group includes 16 females and 11 males, aged 22-36. Most are enrolled students (BSc, MSc, or PhD), and one is a postdoctoral researcher. None are native English speakers, though all have knowledge of English as a second or third language and rate themselves as proficient. None of the participants have prior experience as caregivers for individuals with dementia.

B.1.3 Recruitment and Consent Procedure

Permission to recruit annotators was granted by the author's academic institution (IRB). Annotators were recruited through an ad distributed in student WhatsApp groups. A trained member of the research team conducted individual phone conversations to explain the task, during which verbal informed consent was obtained. Following consent, annotators received the data and annotation guidelines. Upon completion of the task, they were compensated with $50 or academic credit.

B.2 LLM Perception Prompt

"You are analyzing a spoken description of the Cookie Theft Picture from the Boston Diagnostic Aphasia Examination (BDAE). The speakers (either Dementia patients or Healthy Controls) were asked to describe the picture with as many details and/or actions as they can. Their responses were recorded, then transcribed from voice to text. The transcripts may include repetitions, incomplete words, non-verbal gestures (laughter, cough, etc). Given what you know / heard / imagine about Dementia / Alzheimer's patients, please mark whether this description sounds like it's from a healthy person, or from a Dementia patient. Please use your best judgment. Remember, we are not asking for a clinical diagnosis- just your perception. Please return "dementia" if dementia, "healthy" if healthy. Please return "dementia" or "healthy" only, no explanations or intermediate calculations."

C Data Analysis

C.1 Feature Distribution

Figure 5 presents the number of occurrences of "1" or "0" for each feature. Five features were marked as positive in fewer than 5% of the samples (i.e., fewer than 25 samples): *vision difficulties, introduction, naming characters, empathy*, and *irritability*, and were therefore cleaned from our dataset and disregarded throughout our analyses.

C.2 Feature-Feature Correlations

Figure 6 presents Pearson correlations within our 38 features. No extreme correlations are observed, although the Theory of Mind features (ToM boy,

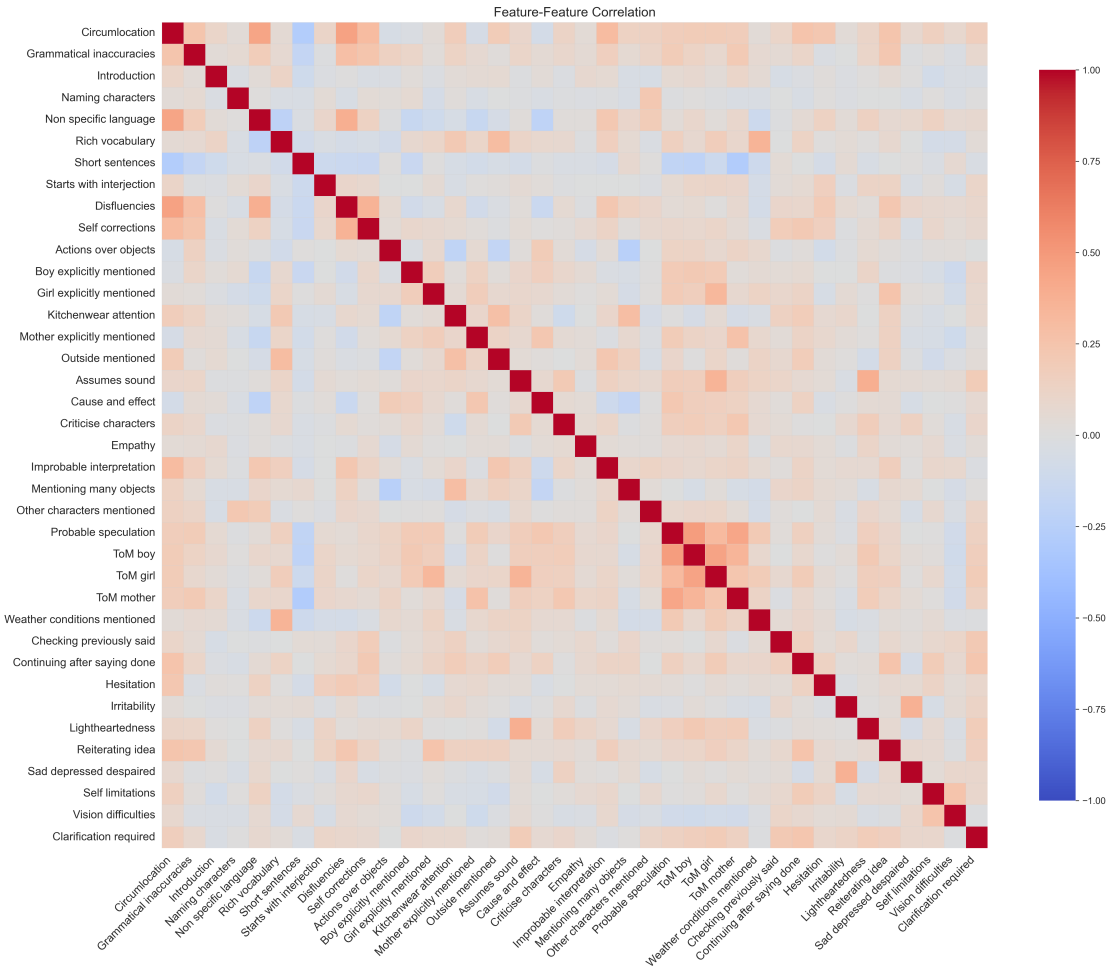

Figure 6: Heatmap of feature correlations (Pearson): red indicates positive correlation, and blue indicates negative.

ToM girl, and ToM mother) appear somewhat correlated with one another, as well as with probable speculation. This pattern is intuitive: probable speculation, i.e., interpreting something not indisputably visible in the picture, may overlap with describing emotions or motivations. Also, a tendency to apply ToM to one character may reflect an inclination to do so across multiple characters.

C.3 Feature-Judgment Correlations

Figure 7 presents Pearson correlations between our 38 features and the three judgment types: human perception (top row), LLM perception (middle row), and clinical diagnosis (bottom row). Linguistic features (*disfluencies, non-specific language*, etc.) consistently correlate with the "Dementia" label across all three judgments. LLMs show correlations with a broader range of cues than either humans or clinicians, correlating with Human Experience features like *lightheartedness* and *negative sentiment*. Additionally, they show strong associations ($p < 0.001$) with 13 features and additional correlations ($p < 0.05$) with 6 more– nearly double those seen in human perception. LLM judgments also correlate with all five feature categories, suggesting they draw on a wide spectrum of cues. In contrast, human perception is narrower: linguistic cues suggest dementia, while Objective Interpretation features such as mentioning the girl, mother, or outdoors align with perceived healthiness. Clinical diagnosis aligns most with Linguistic and Objective features, with some sensitivity to Subjective elements like weather mentions. Interestingly, weather, although not shown in the image, is significantly correlated across all three sources, making it the only Subjective Interpretation feature shared in this way. Clinical diagnosis also identifies short sentences as a dementia marker, consis-

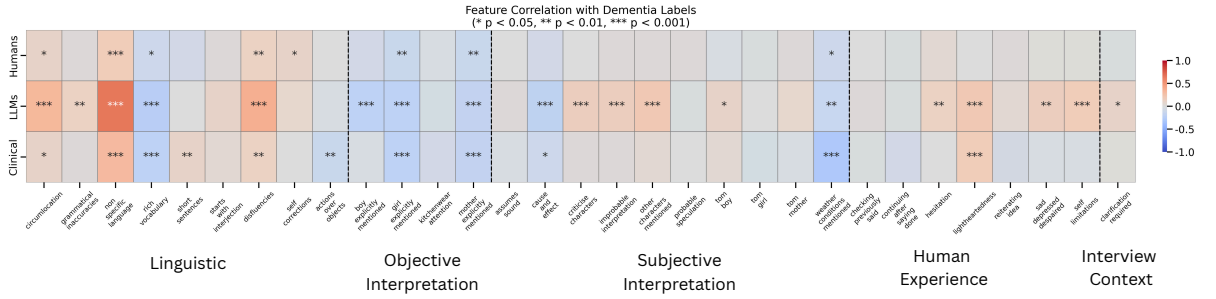

Figure 7: Heatmap of Pearson correlations between features and Human perceptions, LLM majority-vote perceptions, and clinical diagnoses. Red indicates positive correlation with the 'Dementia' label; blue with the 'Healthy' label.

tent with prior findings (Hier et al., 1985; Blanken et al., 1987; Murray, 2010; Li et al., 2024; Agmon et al., 2024). Finally, clinical judgments show fewer correlations with emotional and behavioral cues compared to LLMs, likely because clinicians rely on multimodal signals (tone, expressions, body language) during in-person interviews, cues that LLMs must infer from language alone.

D Perception Modeling- Sup. Material

D.1 Implementation

Our logistic regression models were implemented in R version 4.5.1 using the glm() and step() functions. The glm() function was configured with the binomial family for logistic regression, a tightened convergence tolerance (epsilon = 1e-8) to ensure numerical stability, and an increased cap of 25 iterations to allow for complex feature sets. Model selection was performed via the step() function with bidirectional stepwise selection (direction = "both"), guided by the Akaike Information Criterion (AIC). The search space was defined from a null model (intercept only) to a full model including all candidate features, with up to 1000 steps.

D.2 All Significant Coefficients

Table 3 shows all significant coefficients for the "Dementia" and "Healthy" labels in the human perception, LLM perception, and clinical diagnosis models. Positive coefficients indicate association with "Dementia" while negative coefficients indicate association with "Healthy". These results are also visualized in Figure 3.

D.3 Prediction Evaluation

Section 6 described the stepwise logistic regression models used to analyze coefficients linked to human perception, LLM perception, and clinical diagnoses. Since our primary goal is to interpret the learned coefficients and derive meaningful insights, we initially trained the models using the entire dataset. However, to comply with NLP standards and evaluate predictive performance, we then train models using only the statistically significant features for each judgment group (Section 7.1). For example, to predict human perception, we train a stepwise logistic regression model using only the features found significant for that group.

We use a 5-fold cross-validation with test folds of 103-104 instances each, reporting average performance on the held-out test sets. To prevent data leakage, we ensure that no patient with multiple samples (due to the longitudinal structure of the Pitt corpus) appears in both training and test sets.

Table 4 presents accuracy, precision, recall, F1 score, and ROC-AUC for each model. Performance is notably strong when predicting LLM perception, possibly due to the broader range of features they rely on (13 features, compared to 8 and 4 for clinical diagnosis and human perception, respectively).

Lower scores were observed for the prediction of clinical diagnosis. This is expected, given our approach of representing Cookie Theft picture descriptions using a small number of binary, high-level features. State-of-the-art studies aiming to optimize prediction on the Pitt corpus report accuracies around 85%; however, they rely on large transformer architectures that are hyperparameter-tuned on the full raw text, and sometimes audio as well (Ilias and Askounis, 2022). Studies more similar in nature to ours – i.e., those that represent Cookie Theft transcripts using relatively few, well-defined extracted features – report predictive metrics more in line with our results (e.g., F1 scores around 70%) (Sirts et al., 2017; Wankerl et al., 2017). Accordingly, our regression model for predicting clinical diagnoses, based solely on eight high-level features, is not optimized for high-accuracy dementia detec-

Judgment	Feature Name	β	SE	Wald Z	p-value	Significance
Clinical Diagnosis	weather conditions mentioned	-1.567	0.332	-4.71	$p < 0.0001$	***
	non specific language	1.091	0.217	5.024	$p < 0.0001$	***
	girl explicitly mentioned	-1.165	0.329	-3.542	0.0003	***
	lightheartedness	0.934	0.26	3.592	0.0003	***
	mother explicitly mentioned	-1.467	0.528	-2.778	0.0054	**
	short sentences	0.598	0.215	2.783	0.0053	**
	actions over objects	-0.589	0.216	-2.72	0.0065	**
	probable speculation	0.724	0.275	2.63	0.0085	**
LLM Perception	non specific language	2.957	0.318	9.278	$p < 0.0001$	***
	girl explicitly mentioned	-1.821	0.439	-4.141	$p < 0.0001$	***
	outside mentioned	-1.467	0.371	-3.953	$p < 0.0001$	***
	disfluencies	2.166	0.577	3.754	0.0001	***
	criticise characters	1.811	0.494	3.662	0.0002	***
	rich vocabulary	-1.467	0.496	-2.956	0.0031	**
	boy explicitly mentioned	-1.537	0.466	-3.299	0.0009	***
	lightheartedness	0.829	0.338	2.446	0.0144	*
	self limitations	0.697	0.289	2.409	0.0159	*
	actions over objects	-0.76	0.317	-2.393	0.0166	*
	sad depressed despaired	1.252	0.545	2.298	0.0215	*
	tom boy	0.839	0.374	2.241	0.025	*
	other characters mentioned	1.065	0.527	2.022	0.0431	*
Human Perception	non specific language	0.568	0.19	2.99	0.0027	**
	girl explicitly mentioned	-0.572	0.263	-2.17	0.0299	*
	short sentences	-0.571	0.192	-2.965	0.003	**
	mother explicitly mentioned	-0.758	0.372	-2.037	0.0416	*

Table 3: Coefficient analysis for logistic regression models predicting each judgment type. The regression coefficient is denoted by β, with larger absolute values indicating a stronger influence. Positive β values suggest an association with the *Dementia* label, and negative values with *Healthy*. *SE* denotes the standard error. The Wald statistic tests whether $\beta = 0$. p-values reflect each feature's significance in predicting perceived or diagnosed dementia.

	Clinical	LLM Perc.	Human Perc.
Accuracy	67.3	84.1	60.1
Precision	69.9	80.7	58.1
Recall	75.4	81.0	42.9
F1 Score	71.1	80.5	48.7
ROC-AUC	78.4	92.1	62.5

Table 4: Evaluation metrics for the predictive performance of logistic regression perception models across all three judgment types.

tion. However, it is designed to provide insight into how different groups perceive dementia.

Finally, when looking at human perception, it is no surprise that it is difficult to predict. Throughout our work, we have repeatedly demonstrated the inherent inconsistency, subjectivity, and confusion in human judgments; even the annotators themselves lacked clear insight into their decision process.

E Analysis of Human-provided Rationale

We received reports from 18 of the 27 annotators, describing the patterns they noticed and believed they relied on to assess whether the speaker appeared cognitively impaired or healthy. We translated these responses into English and manually reviewed them. Raw annotator responses are presented in Appendix E.1. We distilled these responses into concepts, some align with features from our predefined expert-guided list, and others representing new, emergent patterns. We then tallied the number of times each of these concepts/features was mentioned. The results of this analysis are shown in Table 5. The most frequently mentioned features were *reiterating ideas* (10 mentions) and *disfluencies* (7 mentions).

Notably, some concepts introduced by annotators were broader in scope than our predefined categories. Among the emergent concepts, *level of*

	Feature Name	Num Mentions
Fully correspond to our features	reiterating ideas	10
	disfluencies	7
	improbable interpretation	5
	checking previously said	3
	circumlocution	2
	clarification required	2
	naming characters	2
	rich vocabulary	2
	short sentences	2
	self limitations	1
	boy explicitly mentioned	1
	girl explicitly mentioned	1
	mother explicitly mentioned	1
Additional features mentioned	level of detail	9
	incoherence	5
	total length	3
	talk about self	1
	assigning emotions	1
	laughter	1
	confusion	1

Table 5: Concepts reported by human annotators.

detail was the most frequently cited possibly due to its encompassing nature, covering a range of specific content features. Similarly, some annotators referred generally to the attribution of emotions to characters, without identifying which characters. A particularly nuanced observation regarding *level of detail* is that some associated it with signs of cognitive impairment, while other interpreted it as indicative of cognitive intactness.

Some features also spanned multiple interpretive categories. For example, *confusion* could be understood as a linguistic, interpretive, or contextual cue, and *laughter occurrence* may relate both to the speaker's emotional state and the dynamics of the interview setting. This suggests that annotators often focused on broader impressions rather than specific, fine-grained elements.

E.1 Raw Annotators Responses

- Fragmented sentences, repetitions, confusion, and self-questioning like "Did I already say that?"
- Marked as "Dementia" if there was reference only to part of the picture, mentions of things not shown, overly short or discontinuous descriptions, many questions, incoherence, or excessive focus on tiny details.
- Repetition of the same sentence or phrase, messy or incoherent flow, or exact repetition of a sentence triggered a dementia label.
- Explicit self-checks (e.g., "Did I already say that?") and topic jumps were viewed as dementia indicators.
- Sentences that faded mid-way or had abrupt topic changes; exaggerated imagination or personal tangents were seen as red flags.
- Initially thought humming might be a sign, but later decided it wasn't a reliable cue.
- Looked for attention patterns, whether patients noticed different details, but found no clear pattern.
- Hesitated over emotional expressions like "oh no" or laughter but didn't use them as indicators.
- Often relied on gut feeling: either excessive rambling or overly clipped responses missing key details stood out.
- Confusion between characters, fading thoughts (not normal stuttering), assigning names, and projecting emotions/inferences were red flags.
- Excessive ellipses or repetitive statements like "I don't know, I don't know" were used as indicators.
- Less about content and more about delivery, the words themselves.
- Looked for repetition, vocabulary level, length of response, and level of detail (with detailed = healthy).
- Scattered responses or inclusion of elements not present in the image were marked as dementia.
- Naming characters or imagining beyond the image was a red flag.
- Repetition and lack of detail were key; detail often equated with cognitive health.
- Confused or incorrect responses (e.g., interpreting cookie quantity emotionally) raised suspicion of decline.
- Clear general descriptions were often seen as indicative of intact memory.

- Noticed differences in attention to detail across participants, some highly detailed, others gave almost nothing.
- Repetition spaced out in the text and incoherent sentence flow with emotional inserts were used as dementia cues.
- Requests to repeat the task, especially mid-description, were noted as potential signs.
- Strange repetition or failure to find words (e.g., calling a stool a chair) stood out.
- Incoherent phrasing, incomplete ideas, and inability to retrieve specific words were viewed as markers.
- People who laughed were perceived as healthy.
- Repetition of actions, lack of detail, or unrelated storytelling triggered a dementia classification.
- Jumping between characters inconsistently (e.g., finishing with the children, moving to the mother, then back) was noted.