# Adversarial Text Generation using Large Language Models for Dementia Detection

**Youxiang Zhu, Nana Lin, Kiran Sandilya Balivada,**
**Daniel Haehn, Xiaohui Liang**
University of Massachusetts Boston, Boston, MA, USA
{youxiang.zhu001, nana.lin002, kiran.sandilya001,
daniel.haehn, xiaohui.liang}@umb.edu

## Abstract

Although large language models (LLMs) excel in various text classification tasks, regular prompting strategies (e.g., few-shot prompting) do not work well with dementia detection via picture description. The challenge lies in the language marks for dementia are unclear, and LLM may struggle with relating its internal knowledge to dementia detection. In this paper, we present an accurate and interpretable classification approach by Adversarial Text Generation (ATG), a novel decoding strategy that could relate dementia detection with other tasks. We further develop a comprehensive set of instructions corresponding to various tasks and use them to guide ATG, achieving the best accuracy of 85%, >10% improvement compared to the regular prompting strategies. In addition, we introduce feature context, a human-understandable text that reveals the underlying features of LLM used for classifying dementia. From feature contexts, we found that dementia detection can be related to tasks such as assessing attention to detail, language, and clarity with specific features of the environment, character, and other picture content or language-related features. Future work includes incorporating multi-modal LLMs to interpret speech and picture information.

## 1 Introduction

Large Language Models (LLMs), such as GPT-4 (Achiam et al., 2023) and Llama3 (Touvron et al., 2023), have demonstrated powerful general capabilities in traditional NLP tasks like rewriting and summarization (Pu et al., 2023). They possess two notable advantages: First, they can easily generalize to unseen tasks and specific domains using only a few in-context samples without the need for fine-tuning (Brown et al., 2020). Second, emergent abilities such as chain-of-thought (CoT) (Wei et al., 2022) enhance LLM capability by learning to derive the final answer through intermediate steps from training or in-context examples, and offer better interpretability compared to the smaller size of language models like BERT (Devlin et al., 2018).

Despite its powerful capabilities, LLMs do not perform well in dementia detection with regular prompting strategies like few-shot or CoT. Dementia detection via picture description aims to infer dementia status by analyzing speech recordings or transcripts (Becker et al., 1994). Typical accuracy of LLM on dementia detection lies in the range of 55-75% in our experiments and previous works (Bang et al., 2024), even worse than fine-tuning BERT-liked models with around 80% accuracy (Balagopalan et al., 2020; Zhu et al., 2021b). The challenge lies in the intermediate steps of dementia detection not being well-defined, and even human experts do not have a clear understanding of what kinds of language markers could be used to detect dementia accurately. Without a clear understanding, humans can not write effective demonstrations of intermediate steps, which results in LLMs struggling to learn how to detect dementia from training or in-context examples. As such, LLMs may struggle to relate their internal knowledge with dementia detection. In addition, regular prompting strategies are limited by the context window length and long context understanding capability of LLMs (Liu et al., 2024b), resulting in LLMs are not able to fully understand and effectively learn from the training set.

To bridge this gap, we propose **Adversarial Text Generation** (ATG) to relate dementia detection with other tasks that LLM may be capable of, with the guidance of the training set. ATG is a perplexity-based decoding strategy inspired by previous studies using perplexity for dementia detection (Fritsch et al., 2019; Cohen and Pakhomov, 2020; Li et al., 2022). As shown in Figure 1, given a training set and an instruction, ATG generates a human-understandable **Feature Context**, which could be used for perplexity-based classification.



*Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21918–21933
November 12-16, 2024 

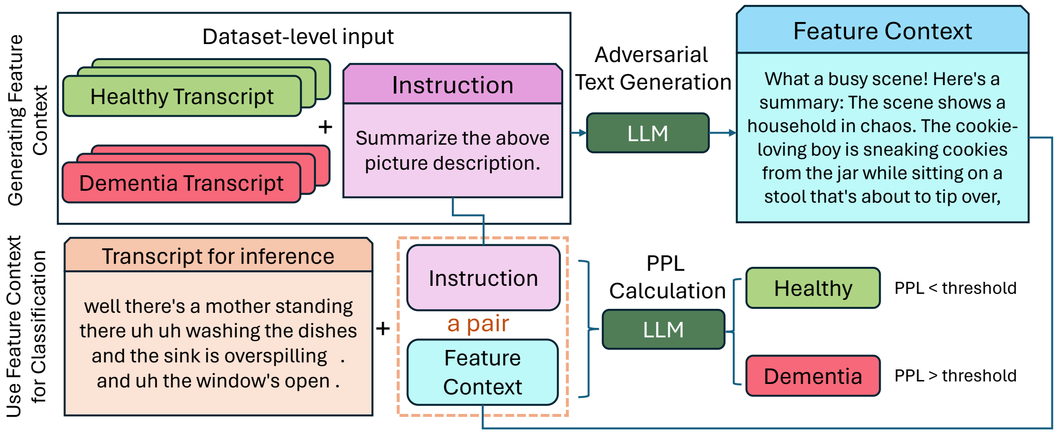


Figure 1: We propose adversarial text generation (ATG) to generate feature context based on a training set and an instruction using LLMs. Then, the instruction and feature context are concatenated with the transcript for perplexity-based classification.

The feature context is considered to be matched with healthy transcripts while unmatched with dementia transcripts, measured by the perplexity. To find out the best tasks related to dementia detection, we introduce comprehensive **Instruction guiding strategies**, which guide ATG to generate distinguishable feature contexts. In experiments, we show that these improvements ensure the generated feature contexts are task-specific and distinguishable, thus enhancing classification performance. Additionally, the feature context provides an interpretable background, facilitating further studies of the explicit features underlying dementia detection. Our contributions are three-fold:

First, we propose adversarial text generation, a perplexity-based decoding strategy that could relate dementia detection with other tasks. ATG generates a feature context that diversifies the perplexity of healthy and dementia transcripts, and then the perplexity can be used for classification.

Second, we introduce five types of instructions based on the LLM instruction learning and dementia domain knowledge. We observe that effective feature contexts emphasize the differences between healthy and dementia transcripts regarding picture contents, whereas ineffective ones do not.

Third, we introduce a difference-based instruction generation pipeline, achieving the best accuracy of 85.42% and AUC of 88.37% and reveal that dementia detection can be related to tasks including assessing attention to detail, language, and clarity, with features of environment, character, and other features related to picture contents and languages that were not included in previous studies.

## 2 Background

**Dementia Detection.** Detecting dementia via Picture Description Task (PDT) speech is a low-cost and non-invasive method that can be widely accessible to a large population for early detection of dementia. It has been studied for more than 30 years (Becker et al., 1994). In the PDT, participants describe the same picture using spontaneous speech, and researchers aim to detect whether participants have dementia or not by analyzing speech recordings or transcripts. It is considered to be challenging since the labels come from separate cognitive assessments, while even medical professionals could hardly produce accurate inferences from speech or transcripts to dementia status.

**Perplexity.** Perplexity measures the fitness of text to a language model. Formally, given a tokenized text sequence $X = [x_1, x_2, \ldots, x_n]$, the perplexity is defined as follows:

$$\mathrm{PPL_M}(X) = \exp\left\{-\frac{1}{n}\sum_{i=1}^{n}\log\left(M(x_i \mid x_{<i})\right)\right\}$$

where $M(x_i \mid x_{<i})$ are the output probability of $i$-th token of the model $M$. If the text is less common to the knowledge of $M$, the perplexity is larger; if the text is more common to the knowledge of $M$, the perplexity is smaller. We consider the language models are trained on texts mostly generated by humans without dementia or cognitive problems; the healthy transcripts from the PDT task should fit the language models better than the dementia transcripts. Thus, the perplexity score of dementia transcripts tends to be larger than that of healthy

transcripts. However, classification using such perplexity differences and a threshold produced limited performance. Additional training is needed for fully exploring LLMs and perplexity for dementia detection (Fritsch et al., 2019; Cohen and Pakhomov, 2020; Li et al., 2022).

**Regular text generation** aims to choose the next token to minimize the perplexity of the whole text sequence. Specifically, in the case of greedy search, the next token $x_{n+1}$ is chosen based on the minimum value of $\mathrm{PPL}_{\mathrm{M}}(X||x_{n+1})$, where $||$ is the concatenate operation. We denote the process of regular text generation as $Z = \mathrm{RTG}_{\mathrm{M}}(X)$, where $X$ and $Z$ are the input and output text sequences.

# 3 Method

In this section, we introduce the implementation of ATG for dementia detection. We first introduce a **perplexity-based classifier**, which enables the use of ATG for classification. It classifies a PDT transcript into dementia or healthy classes based on the perplexity of an input of a transcript, an instruction, and a feature context. A high perplexity score implies the transcript is from a dementia patient, and a low perplexity score implies it is from a healthy control. Then, we introduce the two **objectives** for using ATG for generating feature context: perplexity polarization and text coherence. The former ensures that the feature context fits the healthy transcripts and unfits the dementia transcripts, while the latter ensures the feature context is meaningful. Lastly, we introduce **instruction-guiding strategies** for ATG to enhance the utility of feature context and help interpret features for dementia detection.

## 3.1 Perplexity-based classifier

A perplexity-based classifier $\mathcal{C}_M$ takes inputs of a transcript, an instruction $I$, and a feature context $C$ and makes an inference on dementia or health. The derivation of $C$ and $I$ will be discussed in the later sections of adversarial text generation and instruction generation. Denote a training set of the PDT transcripts as $D_{train} = \{X_1, X_2, \ldots, X_l\}$, and denote healthy and dementia transcripts of the training set as $D^h_{train}$ and $D^d_{train}$, respectively. We calculate the perplexity scores for all transcripts $\mathrm{PPL}_{\mathrm{M}}(D_{train}, I, C) = \{\mathrm{PPL}_{\mathrm{M}}(X_i||I||C)|1 \leq i \leq l\}$. Following the previous work (Li et al., 2022), we choose a perplexity threshold $th$ at the equal error rate (EER) of the training set using the training label. The classifier is described below:

$$\mathcal{C}_M(X, I, C) = \begin{cases} \text{dementia}, & \mathrm{PPL}_{\mathrm{M}}(X||I||C) \geq th \\ \text{healthy}, & \mathrm{PPL}_{\mathrm{M}}(X||I||C) < th \end{cases}$$

## 3.2 Objectives for ATG

The ATG generates a feature context $C = \mathrm{ATG}_{\mathrm{M}}(D_{train}, I)$ using the training set $D_{train}$ and an instruction $I$ as inputs. The instruction $I$ will be discussed in the next section. ATG has two objects: perplexity polarization and text coherence.

### 3.2.1 Perplexity polarization

We choose the next token with a maximum perplexity-based metric (denoted as PPL metric). Specifically, given an existing context $C_n$ with $n$ token, for each possible next token $c_{n+1}$, we calculate a set of perplexity scores for the training set $\mathrm{PPL}_{\mathrm{M}}(D_{train}, I, C_n||c_{n+1})$. Then, we consider four metrics in two categories: performance-based metrics (ACC, AUC) and distance-based metrics (PPL-F, PPL-S). These metrics are used to select the next token in the text generation process.

**ACC value.** For each $c_{n+1}$, we calculate an ACC value using three steps: 1) set $C = C_n||c_{n+1}$; 2) develop a perplexity-based classifier and obtain a threshold $th$ according to §3.1; and 3) calculate accuracy according to the threshold $th$, the perplexity scores and labels of the training set.

**AUC value.** For each $c_{n+1}$, we calculate the value of the area under the receiver-operator characteristic curve (AUC) using the perplexity scores and labels of the training set.

**PPL-F value.** For each $c_{n+1}$, we calculate the PPL-F value as the difference of the mean perplexity score of two classes, i.e., $\mathrm{mean}(\mathrm{PPL}_{\mathrm{M}}(D^d_{train}, I, C_n||c_{n+1})) - \mathrm{mean}(\mathrm{PPL}_{\mathrm{M}}(D^h_{train}, I, C_n||c_{n+1}))$.

**PPL-S value.** We define the PPL-S as Cohen's d (Cohen, 2013) between dementia and healthy transcripts. Specifically, for each $c_{n+1}$, we calculate PPL-S value as PPL-F$(c_{n+1})/s$, where $s$ is the pooled standard deviation of the perplexity scores of the healthy and dementia transcripts.

**Comparing different PPL metrics.** We first discuss performance-based metrics. ACC has the advantage of considering the precision-recall balance since it is calculated using EER. However, multiple next tokens could have the same ACC value. In comparison, the AUC value is more fine-grained than the ACC, as the next tokens are unlikely to have the same AUC value. A common problem with performance-based metrics (ACC and AUC)

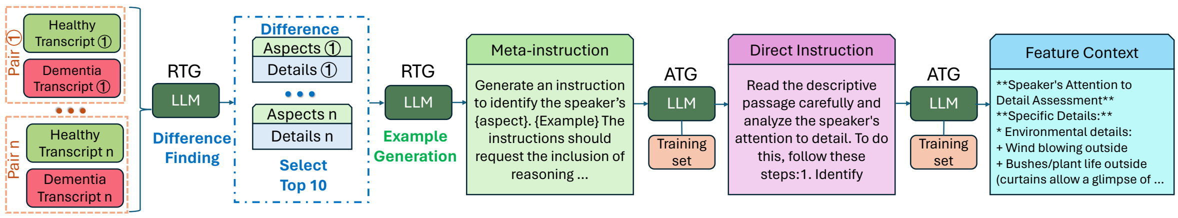


Figure 2: Instruction guiding strategies. We could 1) use human-defined direct instructions for generating feature contexts, 2) use human-defined meta-instructions to generate direct instructions, and 3) use LLM-generated difference-based information to construct meta-instructions (i.e., difference-based instructions).

is that once a sample is correct at a given threshold, it will no longer influence the selection of the next token. This problem can be overcome by distance-based metrics. PPL-F is defined using mean values, which makes it susceptible to the influence of extreme values. This may result in only certain samples with extreme PPL-F values being considered when choosing the next token. As such, we may potentially have large PPL-F values but are still underfitting. In comparison, PPL-S considers both mean and standard deviation so that it is less susceptible to extreme values and unlikely to underfit.

**Combining metrics.** We can take advantage of different PPL metrics by adding two or more of them and using the added values to rank tokens. We chose to use the PPL-S combined with ACC for our main experiments since the PPL-S works better with optimization than PPL-F, and ACC considers the precision-recall balance. We also discuss the effect of single and more combinations of PPL metrics in (§5.4).

#### 3.2.2 Coherence

While choosing the next token based on perplexity polarization, the model may not generate semantically coherent text. The incoherence of the text will limit the interpretability and make the generation process suboptimal (discussed in §5.4). Thus, we introduce a top-$p$ (nucleus) sampling method to ensure the coherence of the text. Specifically, for each $c_{n+1}$, we calculate the average output logits $\frac{1}{l}\sum_{X\in D^h_{train}} M(c_{n+1} \mid X||I||C_n)$ over the healthy transcripts of the training set. Then, we proceed with regular top-$p$ sampling that takes the softmax for all the possible next tokens and selects ones with the highest probability until the added probability is larger than a threshold $p$. This token sampling process ensures the coherence of the text.

### 3.3 Instruction guiding strategies

We explore three instruction generation methods to guide the ATG in generating effective feature contexts: direct, meta, and difference-based instructions, with the latter building upon the former, as shown in Figure 2. We first introduce five types of direct instructions that are defined by existing knowledge, including LLM instruction learning and dementia domain knowledge. These instructions are human-defined and directly apply to ATG to generate feature context. We further introduce meta-instructions, incorporating the knowledge found by LLMs through ATG. Specifically, we use meta-instruction to generate a new direct instruction via ATG. Lastly, we introduce difference-based instructions, which further incorporate the pair-wise difference knowledge found by LLMs through RTG into meta-instructions.

#### 3.3.1 Direct instructions

**Empty instruction.** We leave the instruction empty so that the ATG receives no guidance and generates text based on the training set only.

**Common instructions.** We use instructions from the instruction tuning datasets, including self-instruct (Wang et al., 2022) and alpaca (Taori et al., 2023). We consider the most common root verbs as instructions, including "rewrite," "summarize," "identify," and "suggest." We use 31 common instructions as shown in Table 5 in Appendix.

**Freestyle instructions.** The freestyle instruction allows LLMs to discover anything from the transcripts they want to talk about. We test two freestyle instructions: i) "Discuss anything notable in the above text. Include as much detail as possible"; ii) "Ask $n$ questions about the above picture description and answer them."

**Information-unit instruction.** Information units are a set of human-defined subjects, places, objects, actions, and relations in the cookie theft picture (Yancheva and Rudzicz, 2016). Previous

| Name | Content |
|---|---|
| Template $P_1$ | Text 1: {Text 1} Text 2: {Text 2} Find out the difference between text 1 and text 2. Discuss the differences in a list of aspects. |
| Template $P_2$ | Extract the values of {aspect} mentioned in the above text using one sentence. Start with "For example, the {aspect} could be" |
| Template $P_3$ | Generate an instruction to identify the speaker's {aspect}. {Example} The instructions should request the inclusion of reasoning steps, followed by a conclusion drawn from these steps. Output the instruction only. |

Table 1: Prompt templates for difference-based instruction generation.

works show that healthy controls mention more information units than dementia ones. As such, we define the information unit instruction as "Discuss the mention of the following subject, places, objects, action, etc." This instruction includes a total of 35 information units (Table 6 in Appendix)

**Linguistic-based instruction.** 33 linguistic features are selected from the previous dementia works (TaghiBeyglou and Rudzicz, 2024) (based on the $eval$ command in the CLAN package (MacWhinney, 2017)). The instruction examples are "Extract the following features: Total number of utterances in the transcript" and "Mean Length of Utterances." The complete instruction includes all features listed in Table 7 in Appendix).

### 3.3.2 Meta instructions

A meta-instruction $I_m$ could be used to generate new a direct instruction $I_d = \mathrm{ATG}_\mathrm{M}(D_{train}, I_m)$ via ATG. Meta instructions could introduce multiple steps specified in a complicated instruction according to the training set, while humans may not easily handle the many details of the content and features of such instruction. We construct meta-instructions for information-unit instruction or linguistic-based instruction because they contain many details of the content and features of the picture. We add a "Generate an instruction to" prefix to the original instructions and use ATG to generate new ones. We expect the newly generated instructions to include organized steps using information units or linguistic-based features.

### 3.3.3 Difference-based instructions

We introduce three steps of using difference-based instructions: Difference finding, example generation, and meta and direct instruction generation.

**Difference finding.** To find out the main difference between healthy and dementia transcripts, we first generate all the pairs of healthy and dementia transcripts $(X^h, X^d)$. Then, following the prompt template $P_1$ in Table 1, we use the RTG of an LLM to generate the difference $d = \mathrm{RTG}_\mathrm{M}(P_1(X^h, X^d))$ for each pair. The output difference includes a set of aspects $a$ and corresponding details. Aspects are a set of words or phrases (e.g., "attention to detail"), and the details are the difference in corresponding aspects (e.g., text 1 is superficial while text 2 is nuanced). Given the differences from all pairs, we count the number of each unique aspect and then obtain the top 10 aspects. We only keep the difference in these aspects for the next step.

**Example generation.** Then, for each of the top 10 aspects, we use $P_2$ to extract the examples $e$ of details using the first 10 of the difference items $e = \mathrm{RTG}_\mathrm{M}(P_2(([d_1, \dots d_{10}, a)))$.

**Meta and direct instruction generation.** Based on each aspect $a$ and corresponding examples $e$, we construct the meta-instruction $P_3(a, e)$ using template $P_3$ and generate an direct instruction $I_a = \mathrm{ATG}_\mathrm{M}(D_{train}, P_3(a, e))$. The instruction $I_a$ is expected to include a step-by-step guide of feature extraction for dementia detection.

## 4 Data and implementation details

We used three speech datasets collected via the PDT task and the cookie theft picture, which has been publicly available for dementia research. **ADReSS-2020** (Luz et al., 2021b) include 108 samples for training and 48 samples for testing. Human transcripts are provided in this dataset. We present our main results using this dataset, considering it is the only dataset with standard train test split, human transcription, and balanced numbers of samples for each class, age, and gender. **ADReSSo-2021** (Luz et al., 2021a) include 166 samples for training and 71 for testing. It also has a standard train test split and balanced numbers of samples for each class, age, and gender. However, it doesn't have human transcription. We transcribe the speech samples using Whisper ASR (large-v3) (Radford et al., 2023). **Pitt** (Becker et al., 1994) dataset includes 548 samples, including 243 healthy and 305 dementia transcripts. It provides human transcriptions but does not have a standard train/test split and balanced numbers of samples for each class, age, and gender, and each participant may have multiple samples. We

| Instruction | Training | | | Testing | | |
|---|---|---|---|---|---|---|
| | P-S | ACC | AUC | P-S | ACC | AUC |
| **Baselines (Regular prompting, no ATG used)** | | | | | | |
| 0-shot | - | - | - | - | 64.58 | - |
| 1-shot | - | - | - | - | 66.66 | - |
| 5-shot | - | - | - | - | 54.17 | - |
| 0-shot-CoT | - | - | - | - | **72.92** | - |
| **Empty instructions** | | | | | | |
| Empty | 1.35 | 75.93 | 83.64 | 1.27 | 79.17 | 83.33 |
| **Common instructions (Top-5 Train PPL-S)** | | | | | | |
| Detect | 1.73 | 83.33 | 89.64 | 1.46 | 77.08 | 85.76 |
| Describe | 1.54 | 83.33 | 87.62 | 1.34 | 77.08 | 83.33 |
| Evaluate | 1.54 | 77.77 | 86.83 | 1.27 | 81.25 | 83.33 |
| Rewrite | 1.53 | 79.63 | 87.14 | 1.32 | 79.17 | 82.81 |
| Explain | 1.51 | 78.70 | 86.80 | 1.31 | 70.83 | 82.64 |
| **Information units instructions** | | | | | | |
| Direct | 0.84 | 60.19 | 70.10 | 1.00 | 70.83 | 75.87 |
| Meta | 1.85 | 85.19 | 91.05 | 1.60 | **83.33** | **88.19** |
| **Linguistic-based instructions** | | | | | | |
| Direct | 0.43 | 51.85 | 57.37 | 0.30 | 54.17 | 57.12 |
| Meta | 0.51 | 50.92 | 59.60 | 0.35 | 54.17 | 58.51 |
| **Free-style instructions** | | | | | | |
| Discuss anything notable | 2.02 | 88.88 | 92.46 | 1.51 | **83.33** | **87.85** |
| Ask 5 questions | 1.47 | 75.93 | 86.56 | 1.28 | 75.00 | 83.16 |
| **Difference-based instructions (Top-5 Train PPL-S)** | | | | | | |
| Attention to detail | 2.06 | 87.04 | 93.66 | 1.58 | 81.25 | 87.50 |
| Language | 2.06 | 84.26 | 93.42 | 1.67 | 77.08 | **88.37** |
| Focus | 1.86 | 87.96 | 90.84 | 1.66 | 79.17 | 88.71 |
| Description of the scene | 1.82 | 85.19 | 91.87 | 1.64 | 81.25 | 87.67 |
| Clarity | 1.70 | 86.11 | 89.71 | 1.64 | **85.42** | 86.63 |

Table 2: Main results of ADReSS-2020 dataset. We report the PPL-S (P-S), ACC(%), and AUC(%) for both training and testing.

used 5-fold cross-validation for this dataset without participant overlap.

We used Llama 3 8B Instruct[1] as the LLM for both ATG and perplexity calculation. We analyzed all parameters of the PPL metrics and used PPL-S + ACC as the main PPL metric. In the ATG, we stop the text generation at the "eos" token and then truncate the sequence at the peak PPL metrics. All experiments were done with a single A100 of 40 GB memory using less than 3 hours per instruction.

We consider the following regular prompting strategies as baselines: 0-shot, 1-shot, 5-shot, and 0-shot-CoT. For 1-shot and 5-shot, we used the first one/five samples in the training set as demonstration examples. We do not consider few-shot-CoT since we cannot come up with accurate CoT demonstrations. The detailed promoting templates are shown in Table 4 in the Appendix.

## 5 Results

### 5.1 Analysis of direct and meta-instructions

We present the baselines, five types of direct and meta-instructions in Table 2 with the following observations.

**Baseline results.** For few-shot prompting, the best result is 0-shot with 64.58% accuracy, and the worst is 5-shot with 54.17% accuracy. More examples may not lead to better results, which indicates the LLM may not effectively learn dementia-related features from in-context examples. The best regular prompting result is 0-shot-CoT with 72.92% accuracy. In comparison, with ATG, even the empty instructions outperform the best regular prompting. The best ATG result of 85.42% shows a 12.5% improvement compared to the best regular prompting results.

**Low overfitting of ATG.** We first compare the performance difference between training and testing. We found that for all of the instructions, the performance difference between training and testing is less than 6% for AUC and less than 8% for ACC, demonstrating the low overfitting of ATG. This is significantly different from the previous fine-tuning-based method, where the training accuracy is easily reached 100% due to the small size of the training set.

**Low-performance instructions.** For the instructions with testing $<$ 80% AUC, including all linguistic-based instructions and direct information unit instructions, we found these instructions failed to improve the PPL metrics after 100-200 newly generated tokens. The corresponding feature contexts show limited meaningful information requested by the instruction. Specifically, the feature context of direct information unit instruction only repeats the subjects, and the feature context of the linguistic-based instructions either includes a lot of "not applicable" descriptions or clarification questions to those features.

**High-performance instructions.** For the instructions with $>$ 80% AUC, including the empty instruction, top-5 common instructions, meta information unit instruction, and freestyle instructions, we found they successfully improved the PPL metrics with 200-700 newly generated tokens. The corresponding feature contexts mainly discuss the picture contents. Specifically, the empty instruction generates a feature context that summarizes the scene into points (e.g., the stool falling over) and asks follow-up questions at the end. The "detect" instruction from common instructions considers detecting a crime scene (cookie theft) with features of suspects, motives, opportunities, obstacles, and challenges. The meta information unit instruction generated 6 questions to discuss, includ-

[1]https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct

ing family dynamics, kitchen chaos, window scene, cookies and secrecy, maternal oversight, objects, and consequences. The “ask 5 questions” instruction produces questions related to the picture (e.g., why stool falling). The “discuss anything notable” instruction considers the picture-relevant features, including kitchen mayhem, sink overflowing, the wind outside, kids’ actions, mother’s neglect, and summer puddled insight. By using feature contexts related to picture contents, healthy transcripts fit these contexts, while dementia transcripts would not fit. We consider such contexts to emphasize the difference between healthy and dementia transcripts in terms of picture contents. Among these instructions, the meta information unit instruction and “discuss anything notable” instruction achieved the best performance, with both ACC of 83.33% and AUC of 88.19% and 88.37%, respectively. We consider the meta information unit instruction to guide the LLM in extracting the features with domain knowledge, while the “discuss anything notable” instruction gives the LLM freedom to extract the features.

### 5.2 Analysis of difference-based instructions

We show the top 5 difference-based instructions in Table 2. We found that the feature context of these instructions has common features related to picture contents, including actions, events, settings, characters, etc. Such features could be considered as the main difference between healthy and dementia transcripts, ensuring the good performance of difference-based instructions (all have AUC > 86%) that outperform the direct and meta instructions. In addition to picture contents, some of the instructions also focus on features related to language. For example, the “language” instruction also discusses vocabulary, sentence structure, and tone as features. The “clarity” instruction also identifies some problematic sentences or phrases and their impact on clarity. We consider features related to language to also contribute to the performance that the “language” instruction achieves the best AUC of 88.37% while the “clarity” instruction achieves the best ACC of 85.42%. Also, compared to the previous works that only use text modality (Balagopalan et al., 2020; Li et al., 2022), our work achieves similar or better performance. Overall, we conclude that difference-based instructions have successfully identified the difference between healthy and dementia transcripts and achieved better performance than the direct and meta-instructions.

| Instruction | ADReSSo-2021 ACC | ADReSSo-2021 AUC | Pitt ACC | Pitt AUC |
|---|---|---|---|---|
| **Baselines (Regular prompting, no ATG used)** | | | | |
| 0-shot | **74.65** | - | **60.4(4.83)** | - |
| 1-shot | 69.01 | - | 59.49(4.79) | - |
| 5-shot | 56.34 | - | 48.36(5.01) | - |
| 0-shot-CoT | 60.56 | - | 57.12(6.65) | - |
| **ATG instructions** | | | | |
| Empty | 69.01 | 72.22 | 70.09(5.81) | 76.57(4.65) |
| Discuss anything notable | 64.79 | 72.46 | 70.99(4.87) | 78.1(4.86) |
| Attention to detail | **73.24** | **79.92** | 71.9(4.02) | **80.2(3.56)** |
| Language | 61.97 | 70.56 | 69.35(4.35) | 77.39(4.31) |
| Focus | 64.79 | 75.87 | 70.62(2.98) | 77.89(1.98) |
| Description of scene | 67.61 | 76.51 | 69.35(5.09) | 79.5(5.26) |
| Clarity | 64.79 | 74.44 | **72.27(5.57)** | 78.37(5.17) |

Table 3: Results of ADReSSo-2021 (testing) and Pitt (5-fold cross-validation, mean, standard deviation).

### 5.3 Results on other datasets

We provide the results on larger (Pitt) and speech-based (ADReSSo-2021) datasets in Table 3. We consider the “empty”, “discuss anything notable” and all top 5 difference-based instructions. We found overall, the “attention to detail” generalized well in the larger and speech-based dataset, with ACC of 73.24% and 71.9%, and AUC of 79.92% and 80.2%, respectively. We consider “attention to detail” as the main shared difference between healthy and dementia transcripts across different datasets, which mainly focus on the features related to picture content. Other instructions have some level of performance drop with larger or speech datasets. By checking the corresponding feature context, we found there are variations in the descriptions of language-related features. For example, considering the “language” instruction with the first feature of “vocabulary,” there are variations in the descriptions of the “vocabulary,” in the feature contexts of different datasets: “common everyday word” (fold 1 Pitt), “complex vocabulary” (fold 2 Pitt), “everyday” (ADReSS-2020), “neither overly formal nor simplistic” (ADReSSo-2021). Such variations indicate bias across different datasets, which means a feature that works well in a dataset may not generalize to others. Larger-scale data are needed to find robust features that could be consistent in different data distributions. Also, compared to baseline results, ATG outperform the text-based dataset (Pitt) while not outperforming the best baseline of the speech-based dataset. We consider ATG use perplexity as measurement may be sensitive to ASR errors, which may be addressed by future available speech large language models. To conclude, we found features related to picture content to be generalized better than features related to language across different datasets.

### 5.4 Parameter analysis

To understand the effect of different parameters, we compare the different PPL metrics and top-$p$s in Figure 3. For comparing different PPL metrics, we set the top-$p$ to 0.9. To compare different $p$, we use PPL-S + ACC. We use the AUC as the performance metric (i.e., the y-axis in Figure 3).

**PPL metrics.** As shown in the left two sub-figures of Figure 3, we found PPL-S + ACC achieves the best train and test performance, as expected. Some metrics, including PPL-S, AUC, PPL-S + ACC + AUC, and PPL-S + AUC, also achieve comparable performance. For generated feature context, we found mentions of similar features, including characters, actions, and other observations like window and weather conditions, etc., despite being in a different order and organization. Other metrics, including PPL-F, ACC, PPL-F + AUC, and PPL-F + ACC, do not produce good performance. By checking the generated feature context, we found these PPL metrics did not generate coherent content. For ones with PPL-F, they start to generate misspelled words at around 200 tokens, and then the performance starts to decrease at that point. For ACC, we found it starts generating random symbols at around 100 tokens, with only a little performance increase after that. We conclude that well-performed metrics could generate relevant and coherent feature context, while bad-performed metrics can not.

**Top-$p$s.** As shown in the right two sub-figures of Figure 3, we found the 0.9 top-$p$ achieves the best training and testing performance of 0.92 and 0.87, respectively. 1.0 (no top-$p$ sampling) achieves a little worse performance than 0.9, while the other top-$p$ values achieve limited performance. By checking the generated feature context, we found 0.9 top-$p$ generated the coherent contents with no misspelling error. 1.0 top-$p$ has some incoherent content, such as little misspelling errors, indent errors (e.g., misplaced tabs), and mess up with languages (generated some tokens in languages other than English), indicating the necessity of coherence objective. Other low top-$p$ values tend to generate short sentences with limited details, which results in limited performance. Overall, we conclude that top-$p$ sampling is necessary for coherence and performance, while lower top-$p$ values may result in limited details and low performance.

## 6 Discussion

**Improving speech task design of dementia detection.** Our findings may contribute to the design of better speech tasks for dementia detection in the future. As shown in Figure 4, 5 and, 6, most parts of our feature contexts are related to the picture information. This indicates the most effective features for dementia detection are task-dependent. In contrast, other task-independent features may be less effective. This finding highlights the importance of task design for effective dementia detection. It also suggests that future speech task design improvement may need to prioritize the search under highly controlled settings (e.g., ask all participants to talk about the same picture, topic, etc.) to effectively elicit the difference in speech and language between dementia and healthy participants.

## 7 Related work

Speech analysis is a non-invasive and low-cost method for dementia classification (Vigo et al., 2022). Various speech tasks are studied by researchers such as telephone interview (Konagaya et al., 2007), linguistic features (Rentoumi et al., 2017), picture descriptions (Hernández-Domínguez et al., 2018; Guo et al., 2021), speech and writing (Gkoumas et al., 2021) and voice assistants (Liang et al., 2022). Recent studies explore automatic ways (Yang et al., 2022) to analyze spoken language to achieve fast, accurate, and eco-economical tools for dementia detection. There is sufficient evidence showing that machine learning has the ability to distinguish between dementia patients and healthy controls via speech performance (Warnita et al., 2018; Vázquez-Romero and Gallardo-Antolín, 2020; Roshanzamir et al., 2021).

Pre-trained and large language models, such as BERT (Devlin et al., 2018), GPT-3 (Floridi and Chiriatti, 2020), and LLaMA (Touvron et al., 2023), have achieved state-of-the-art performance on a wide range of NLP tasks. Recently, researchers have used these models in dementia detection. (Balagopalan et al., 2020) observed that fine-tuned BERT models outperform feature-based approaches on the dementia detection task. (Li et al., 2022) proposed a new method, GPT-D, using pre-trained GPT-2 paired with an artificially degraded version of itself to compute the ratio of the perplexities in language from dementia and healthy participants. It showed perplexity could be used for dementia detection by introducing impair-

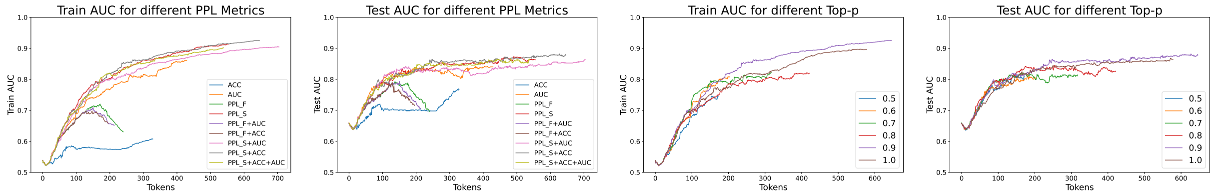


Figure 3: Parameter analysis using "discuss anything notable" instruction. The left two figures compare the different perplexity-based metrics, while the right two figures compare the different top-$p$ values.

ment to the LLMs. (Agbavor and Liang, 2022)'s work suggested that GPT-3 based text embedding is a viable approach for dementia detection from speech transcripts and has the potential to improve early diagnosis of dementia. In-context learning allows language models to learn tasks given only a few examples in the form of demonstration (Dong et al., 2022). It can improve the ability of the model (Wang et al., 2023) to predict the probability distribution of the next word in a sequence based on the context of the previous words. Researchers have used in-context learning to improve the performance of text classification LLMs by helping the model to identify context-specific patterns and features that are relevant to the classification task (Brown et al., 2020). Our work in this paper is the first one applying in-context learning and LLMs to enhance both performance and interpretability in dementia detection.

## 8 Conclusion

In this paper, we propose adversarial text generation, which relates dementia detection with existing well-defined tasks. We first introduce a perplexity-based classifier, which classifies a text sequence using perplexity, enabling the ATG for classification. Then, we introduce the objectives for ATG, including perplexity polarization and coherence. We further incorporate a variety of instructions to guide the ATG in generating effective feature context. We found high-performance instructions successfully reveal the difference between healthy and dementia transcripts, while low-performance ones fail to do so. The main features that contribute to the high performance are related to the picture contents, including the environment, characters, etc, while the language-related features may provide additional performance gain. ATG could be further enhanced with multi-modal LLMs and could probably applied to other classification tasks with limited explicit features.

## Limitations

The current version of ATG only considers the information from the transcripts and doesn't consider the information from speech. Despite the fact that text modality generally outperformed speech modality for PDT, incorporating speech information also helps improve performance (Cummins et al., 2020; Koo et al., 2020; Zhu et al., 2021a). This can be addressed using speech LLMs (Hu et al., 2024; Zhang et al., 2023). Similarly, ATG could benefit from incorporating picture information (Zhu et al., 2023) using vision LLMs (Liu et al., 2024a). Also, ATG may benefit from future open-sourced LLMs with stronger reasoning capability, producing more discriminative differences and more reasonable features. Moreover, the current ATG only considers single-turn conversations with LLMs, which could possibly extend to multi-turn for further enhancement. At last, we also note that ATG could possibly be a general framework that applies to many tasks without a clear definition of intermediate steps to gain both performance and interpretability.

## Ethics Statement

We note that ATG could possibly be a pre-screen for dementia instead of a formal diagnosis. Users should proceed cautiously when using the result in the real world. In addition, the feature ATG finds only reflects the distribution of the training data, so we need to be cautious when considering this as medical findings.

## Acknowledgement

This research is funded by the US National Institutes of Health National Institute on Aging, under grant No.1R01AG067416, in part by the College of Science and Mathematics Dean's Doctoral Research Fellowship through fellowship support from Oracle, project ID R20000000025727.

## References


Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Felix Agbavor and Hualou Liang. 2022. Predicting dementia from spontaneous speech using large language models. *PLOS Digital Health*, 1(12):e0000168.

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

Jeong-Uk Bang, Seung-Hoon Han, and Byung-Ok Kang. 2024. Alzheimer's disease recognition from spontaneous speech using large language models. *ETRI Journal*.

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.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Jacob Cohen. 2013. *Statistical power analysis for the behavioral sciences*. Routledge.

Trevor Cohen and Serguei Pakhomov. 2020. A tale of two perplexities: sensitivity of neural language models to lexical retrieval deficits in dementia of the alzheimer's type. *arXiv preprint arXiv:2005.03593*.

Nicholas Cummins, Yilin Pan, Zhao Ren, Julian Fritsch, Venkata Srikanth Nallanthighal, Heidi Christensen, Daniel Blackburn, Björn W Schuller, Mathew Magimai-Doss, Helmer Strik, et al. 2020. A comparison of acoustic and linguistics methodologies for alzheimer's dementia recognition. In *Interspeech 2020*, pages 2182–2186. ISCA-International Speech Communication Association.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. 2022. A survey for in-context learning. *arXiv preprint arXiv:2301.00234*.

Luciano Floridi and Massimo Chiriatti. 2020. Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30:681–694.

Julian Fritsch, Sebastian Wankerl, and Elmar Nöth. 2019. Automatic diagnosis of alzheimer's disease using neural network language models. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5841–5845. IEEE.

Dimitris Gkoumas, Bo Wang, Adam Tsakalidis, Maria Wolters, Arkaitz Zubiaga, Matthew Purver, and Maria Liakata. 2021. A longitudinal multi-modal dataset for dementia monitoring and diagnosis. *arXiv preprint arXiv:2109.01537*.

Yue Guo, Changye Li, Carol Roan, Serguei Pakhomov, and Trevor Cohen. 2021. Crossing the "cookie theft" corpus chasm: applying what bert learns from outside data to the adress challenge dementia detection task. *Frontiers in Computer Science*, 3:642517.

Laura Hernández-Domínguez, Sylvie Ratté, Gerardo Sierra-Martínez, and Andrés Roche-Bergua. 2018. Computer-based evaluation of alzheimer's disease and mild cognitive impairment patients during a picture description task. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, 10:260–268.

Shujie Hu, Long Zhou, Shujie Liu, Sanyuan Chen, Hongkun Hao, Jing Pan, Xunying Liu, Jinyu Li, Sunit Sivasankaran, Linquan Liu, et al. 2024. Wavllm: Towards robust and adaptive speech large language model. *arXiv preprint arXiv:2404.00656*.

Yoko Konagaya, Yukihiko Washimi, Hideyuki Hattori, Akinori Takeda, Tomoyuki Watanabe, and Toshiki Ohta. 2007. Validation of the telephone interview for cognitive status (tics) in japanese. *International Journal of Geriatric Psychiatry: A journal of the psychiatry of late life and allied sciences*, 22(7):695–700.

Junghyun Koo, Jie Hwan Lee, Jaewoo Pyo, Yujin Jo, and Kyogu Lee. 2020. Exploiting multimodal features from pre-trained networks for alzheimer's dementia recognition. *arXiv preprint arXiv:2009.04070*.

Changye Li, David Knopman, Weizhe Xu, Trevor Cohen, and Serguei Pakhomov. 2022. Gpt-d: Inducing dementia-related linguistic anomalies by deliberate degradation of artificial neural language models. *arXiv preprint arXiv:2203.13397*.

Xiaohui Liang, John A Batsis, Youxiang Zhu, Tiffany M Driesse, Robert M Roth, David Kotz, and Brian MacWhinney. 2022. Evaluating voice-assistant commands for dementia detection. *Computer Speech & Language*, 72:101297.

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

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024b. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Saturnino Luz, Fasih Haider, Sofia de la Fuente, Davida Fromm, and Brian MacWhinney. 2021a. Detecting cognitive decline using speech only: The adresso challenge. *arXiv preprint arXiv:2104.09356*.

Saturnino Luz, Fasih Haider, Sofia de la Fuente Garcia, Davida Fromm, and Brian MacWhinney. 2021b. Alzheimer's dementia recognition through spontaneous speech. *Frontiers in computer science*, 3:780169.

Brian MacWhinney. 2017. Tools for analyzing talk part 2: The clan program. *Talkbank. Org*, (2000).

Xiao Pu, Mingqi Gao, and Xiaojun Wan. 2023. Summarization is (almost) dead. *arXiv preprint arXiv:2309.09558*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR.

Vassiliki Rentoumi, George Paliouras, Eva Danasi, Dimitra Arfani, Katerina Fragkopoulou, Spyridoula Varlokosta, and Spyros Papadatos. 2017. Automatic detection of linguistic indicators as a means of early detection of alzheimer's disease and of related dementias: A computational linguistics analysis. In *2017 8th IEEE international conference on cognitive infocommunications (CogInfoCom)*, pages 000033–000038. IEEE.

Alireza Roshanzamir, Hamid Aghajan, and Mahdieh Soleymani Baghshah. 2021. Transformer-based deep neural network language models for alzheimer's disease risk assessment from targeted speech. *BMC Medical Informatics and Decision Making*, 21:1–14.

Behrad TaghiBeyglou and Frank Rudzicz. 2024. Context is not key: Detecting alzheimer's disease with both classical and transformer-based neural language models. *Natural Language Processing Journal*, 6:100046.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Adrián Vázquez-Romero and Ascensión Gallardo-Antolín. 2020. Automatic detection of depression in speech using ensemble convolutional neural networks. *Entropy*, 22(6):688.

Inês Vigo, Luis Coelho, and Sara Reis. 2022. Speech- and language-based classification of alzheimer's disease: A systematic review. *Bioengineering*, 9(1):27.

Xinyi Wang, Wanrong Zhu, and William Yang Wang. 2023. Large language models are implicitly topic models: Explaining and finding good demonstrations for in-context learning. *arXiv preprint arXiv:2301.11916*.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022. Self-instruct: Aligning language models with self-generated instructions. *arXiv preprint arXiv:2212.10560*.

Tifani Warnita, Nakamasa Inoue, and Koichi Shinoda. 2018. Detecting alzheimer's disease using gated convolutional neural network from audio data. *arXiv preprint arXiv:1803.11344*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Maria Yancheva and Frank Rudzicz. 2016. Vector-space topic models for detecting alzheimer's disease. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2337–2346.

Qin Yang, Xin Li, Xinyun Ding, Feiyang Xu, and Zhenhua Ling. 2022. Deep learning-based speech analysis for alzheimer's disease detection: a literature review. *Alzheimer's Research & Therapy*, 14(1):1–16.

Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. 2023. Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities. *arXiv preprint arXiv:2305.11000*.

Youxiang Zhu, Xiaohui Liang, John A Batsis, and Robert M Roth. 2021a. Exploring deep transfer learning techniques for alzheimer's dementia detection. *Frontiers in computer science*, page 22.

Youxiang Zhu, Nana Lin, Xiaohui Liang, John A Batsis, Robert M Roth, and Brian MacWhinney. 2023. Evaluating picture description speech for dementia detection using image-text alignment. *arXiv preprint arXiv:2308.07933*.

Youxiang Zhu, Abdelrahman Obyat, Xiaohui Liang, John A Batsis, and Robert M Roth. 2021b. Wavbert: Exploiting semantic and non-semantic speech using wav2vec and bert for dementia detection. In *Interspeech*, pages 3790–3794.

| Name | Content |
| --- | --- |
| $n$-shot | Classify the following text into "healthy" or "dementia". Do not output other things.<br>Text: {Example}<br>Label: healthy<br>Text: {Example}<br>Label: dementia<br>(Repeat for $n$ examples)<br>Text: {Example to inference}<br>Label: |
| 0-shot-CoT | Classify the following text into "healthy" or "dementia". You need to think step by step, and then make the conclusion.<br>Text: {Example to inference}<br>Label: |

Table 4: Template of regular prompting strategies.

| Common instructions |
| --- |
| Calculate, Classify, Complete, Construct, Convert, Correct, Create, Describe, Design, Detect, Edit, Evaluate, Explain, Find, Generate, Give, Have, Identify, Make, Name, Output, Paraphrase, Predict, Provide, Rewrite, Simplify, Suggest, Summarize, Tell, Verify, Write |

Table 5: Common instructions.

| Information units instruction |
| --- |
| Discuss the mention of the following subjects, places, objects, actions and relations:<br>Subjects: boy, girl, woman, mother<br>Places: kitchen, exterior<br>Objects: cookie, jar, stool, sink, plate, dishcloth, water, cupboard, window, cabinet, dishes, curtains, faucet, floor, counter, apron<br>Actions: boy stealing cookies, boy/stool falling over, woman washing dishes, woman drying dishes, water overflowing in sink, girl's actions towards boy, girl asking for a cookie, woman daydreaming, unaware or unconcerned about overflow, dishes already washed sitting on worktop, woman being indifferent to the children<br>Relations: brother, sister, son, daughter |

Table 6: Information units instruction.

| Linguistic-based instruction |
|---|
| Extract the following features:<br>Total number of utterances in the transcript.<br>Mean Length of Utterances, which is the average number of words per utterance.<br>Mean Length of Words, which is the average number of morphemes per word.<br>Mean Length of Morphemes, which is the average number of morphemes per utterance.<br>Number of unique word types in the transcript.<br>Number of total word tokens in the transcript.<br>Type-Token Ratio, which is the ratio of unique word types to the total number of word tokens.<br>Percentage of verbs in each utterance.<br>Percentage of word errors in the transcript.<br>Number of errors per utterance in the transcript.<br>Syntactic density, which is a measure of the ratio of the number of content words to the total number of words.<br>Percentage of nouns in the transcript.<br>Percentage of plural nouns in the transcript.<br>Percentage of verbs in the transcript.<br>Percentage of auxiliary verbs in the transcript.<br>Percentage of modal verbs in the transcript.<br>Percentage of third-person singular verbs in the transcript.<br>Percentage of first and third-person singular verbs in the transcript.<br>Percentage of past tense verbs in the transcript.<br>Percentage of past participle verbs in the transcript.<br>Percentage of present participle verbs in the transcript.<br>Percentage of prepositions in the transcript.<br>Percentage of adjectives in the transcript.<br>Percentage of adverbs in the transcript.<br>Percentage of conjunctions in the transcript.<br>Percentage of determiners in the transcript.<br>Percentage of pronouns in the transcript.<br>Ratio of nouns to verbs in the transcript.<br>Ratio of open-class words (nouns, verbs, adjectives, adverbs) to closed-class words prepositions, conjunctions, determiners, pronouns) in the transcript.<br>Number of open-class words in the transcript.<br>Number of closed-class words in the transcript.<br>Frequency of retracing (repeating part of an utterance).<br>Frequency of repetition (repeating entire utterances or parts of utterances). |

Table 7: Linguistic-based instruction

Discuss anything notable

Discuss anything notable in the above text. Include as much detail as possible.<|eot_id|><|start_header_id|>assistant<|end_header_id|>

What a chaotic scene you've described! Here are some notable points to discuss:

**Kitchen Mayhem**

* The stool is falling over, which suggests that it's either old, unstable, or has been poorly maintained. This adds to the sense of chaos in the kitchen.
* The cookies being taken out of the cookie jar by the boy and being asked by the little girl to be handed to her quietly implies that snacking is a common activity in this household.
* The fact that the stool is falling while the cookies are being handled increases the risk of mess and potential accidents.

**Sink Overflowing**

* The sink is overflowing, which means the mother is either taking a long time to wash dishes or isn't paying attention to the water level. This could be a source of stress or annoyance for her, but for now, she's choosing to dry dishes (on the assumption that it's less pressing).
* This detail also creates a sense of water-themed tension, with water running "into" and possibly outside, but no indication of what happens to it.

**Wind Outside**

* The curtains are drawn back to allow a gentle breeze into the house, and the scene appears to be on a relatively mild day, as you mentioned. The fact that you explicitly state there's nothing to note about the outside action (beyond the path, window, and building envelope) suggests that the wind is likely a gentle, harmless blow.

**Kids' Actions**

* The girl's finger to the mouth could be seen as a clever play on innocence, hiding mischief behind a cookie-stealing facade. Her reach for a cookie and the brother's slippage from the stool create a sense of partnership in their naughtiness.
* Both kids seem comfortable snacking in the kitchen, almost like it's a normalized behavior.

**Mother's Neglect (Daycare... er, laundry of attention?)**

* You cleverly contrast the overflowing sink with the mother's drying of dishes, implying a degree of obliviousness or prioritization on her part. Her feet getting soaked or compromised hygiene aren't necessarily consequences of her actions, but it does add to the impression of a household where multiple tasks are happening simultaneously.

**Summer Puddled Insight**

* Your aside about not knowing whether you meant for it to be action-worthy mentions several extraneous details, but effectively highlights the fact that the outside scene doesn't quite gel with the chaotic interior. You raise an interesting question about the importance of showing consequences or actions related to the stool tipping, overflowing water, or blowing curtains.
* The comment about shortsightedly standing in the mother's attention ("um the mother or don't see much. um mother wiping dishes ") serves as a wry commentary on the scene's pandemonium.

Overall, this vivid scene showcases a busy, potentially messy household where multiple dynamics are at play. The juxtaposition of controlled chaos (curtains flowing, kids snacking) with underlying mayhem (stool tipping, overflowing sink) creates a engaging contrast. The outside weather serves as a brief interlude, acknowledging that not everything has to be resolved or escalating inside

Figure 4: Discuss anything notable instruction and feature context.

Attention to detail

**Assessment Task: Speaker's Attention to Detail**

Read the descriptive passage carefully and analyze the speaker's attention to detail. To do this, follow these steps:

1. Identify the specific details mentioned in the passage, focusing on the minor elements such as object positions, character actions, and surroundings.
2. Categorize the types of details mentioned, such as:
    * Environmental details (e.g., wind, curtains, water overflowing)
    * Character actions (e.g., cookie-stealing, stool overturning, mother drying dishes)
    * Object positions and arrangements (e.g., stool falling, cookies in the jar)
    * Character requests (e.g., little girl asking boy to be quiet)
3. Consider the level of precision used to describe each detail. For example, the passage notes specific aspects such as the sink overflowing, but does not describe the temperature or humidity of the environment.
4. Evaluate how the speaker's attention to detail impacts the overall picture painted by the passage. For example, the detailed descriptions of character actions and surroundings create a vivid scene, but the mother's inattention to the overflow suggests that her priorities are elsewhere.
5. Based on these findings, conclude how the speaker's attention to detail affects the passage's overall effect. Is the level of detail meticulous and thorough, or does it prioritize some elements over others? What is the likely effect of the speaker's focus on detail on the reader?<|eot_id|><|start_header_id|>assistant<|end_header_id|>

**Speaker's Attention to Detail Assessment**

**Specific Details:**

* Environmental details:
    + Wind blowing outside
    + Bushes/plant life outside (curtains allow a glimpse of the outdoor view)
    + Open window/ curtains
    + Overflowing sink
    + Weather appears to be light enough to cause the stool to overturn (suggesting not too intense or turbulent)
* Character actions:
    + Girl reaching for cookies from the boy
    + Boy slipping/stalling with cookie ( cookie out of the jar, stool tipping over)
    + Mother drying dishes while oblivious to the overflowing sink
    + Daughter with finger to her mouth (talking quiet, be quiet request to brother)
* Object positions and arrangements:
    + Cookies in the jar, being transferred between boy and sister
    + Stool close enough for it to fall
    + Plates/cups (dry/ not being washed/dried on the counter
* Other details:
    + Summer (water drying her feet not suggesting soaking-wet surface)

**Categorization of details:**

The speaker pays attention to both macro-level elements (water running out of the sink, wind, cookie-snatching kids) and micro-level details (apron-wearing mom, short-handed kids, slips on socks). There's an evident effort to note all manner of movement within the space: the stools falls, girl reaching, brother giving/losing cookies, the curtains in motion ( tied and back slightly away from the breeze)

**Precision of description:**

Most of the descriptions in this passage show remarkable detail and specific location:
The overflowing sink overflowing into the top or splashing onto the mother's wet shoes wouldn't that um let alone uh plate um wiping dish sink um except outside nothing action there fine okay that's fine". Contrastingly, the outside situation lacks similar specificity ("I did say...I guess action alright.")

**Impact on the passage:**

The passage effectively conveys a chaotic and potentially disordered scene within the kitchen, while implying that the mother might not be attending to all aspects of the space (stool slippers into chaos while wiping the dish, breeze seemingly ignored). Some actions stand out as problematic (stealing, stool overturning), while others might seem relatively innocuous (curtains billowing).

**Speaker's attention to detail:**

Overall, the speaker demonstrates meticulous attention to the movements and postures within the kitchen, whereas surrounding areas (particularly outside) receive more vague or observational comments. While there are few weak links (ums, laughs, trailing thoughts), the writer skillfully captures multiple simultaneous actions, object arrangements, and sensations (action-packed vs daydreaming daughter).

The likely effect of the speaker's focus on detail is a visceral, snapshot impression of kitchen mayhem, tempered by some interpretive uncertainty (summer outside unclear, outside calm hard to describe). The careful reader will anticipate trouble erupting soon (crack on the head?), highlighting the daughter's reach into cookies amidst potential

Figure 5: Attention to detail instruction and feature context.

Clarity

**Task: Analyze the Speaker's Clarity**

**Objective:** Evaluate the clarity of the speaker's description by assessing the ease of understanding the scene.

**Steps:**

1. Break down the description into individual elements (actions, events, settings, characters).
2. Identify any sentences or phrases that are:
   * Concise and specific (Text-like)
   * Vague or open-ended (Script-like)
   * Overly descriptive (novelistic)
   * Redundant or unnecessary
3. Consider how these elements contribute to the overall clarity of the scene:
   * Do the individual elements cohere and flow smoothly, or are there jarring transitions or abrupt shifts in focus?
   * Are important details (e.g. the overflowing sink, the curtains blowing) given adequate attention, or do some elements feel secondary or minor?
   * Is the description easy to visualize and imagine, or are some elements confusing or unclear?
4. Use your observations from steps 1-3 to conclude whether the speaker's clarity is more akin to Text or Script. Be sure to support your reasoning with specific examples from the description.<|eot_id|><|start_header_id|>assistant<|end_header_id|>

**Breakdown of the description into individual elements:**

1. Actions:
   * Mother drying dishes
   * Boy stealing cookies
   * Girl asking boy to be quiet (or reaching out for cookie?)
   * Stool about to fall
   * Curtains open and outside looking unobstructed
2. Events:
   * Sink overflowing
3. Settings:
   * Inside the kitchen (mother drying dishes, cookie jar, sink, counter, window)
   * Outside the window (path, etc.)
4. Characters:
   * Mother (daydreaming/distracted)
   * Boy (on stool)
   * Girl (stealing cookie with finger to lip)
5. Descriptive details:
   * Physical settings: water running out of sink, plate, cups on counter, dress/appearance of family
   * Weather conditions: summer season implied, shrubbery possibly suggesting lush outdoor conditions
6. Style: um/xxx indicate filler words, slightly detract from the smoothness of the narrative

**Identifying problematic sentences or phrases:**

* Some phrases feel novelistic or overly descriptive: e.g. "um the wind... those might slow down the pacing or feel gratuitous.
* A few transitions could be tightened: for example, connecting sentences are separated with comma-drops ("the girl was saying be quiet...", "...the mother wiping dishes") or umbrellaed lists of descriptive phrases ("I guess that's it...")
* Small redundancies: stating that "I did say action or right" feels extraneous

**Impact on clarity:**

* Despite these minor issues, the description overall coheres well and presents a cohesive image of a scene in progress.
* Major events (stool tipping, boy/girl misbehaving, overflowing sink) receive due attention and draw the reader's attention to critical areas of the scene.
* Settings (inside and outside) feel accurately evoked, if occasionally lavishly detailed (leaving aside brevity, multiple wind/Cookie-swapping-centric closeups ensue... Ahmargo.wordpress).
* Weather/outside action (# uh do I don't ) jars ever barely blowing explicitly disrupting agency winds clarity—to blowingwater-runningoveisse.

**Conclusion:**

While the description leans slightly towards script-like excess (um-filled asides, parenthetical remarks), the writer primarily presents a detailed, action-rich scene that prioritizes concrete, specific details. Important events unfold rapidly, and individual elements interact in a plausible, chaotic sequence. Compared to typical scriptwriting standards, this scene might err on the side of showing too much ("warts and all" household pandemonium!), whereas literary (text-like) clarity often favors selection and subtlety. Therefore, I argue that this speaker's clarity trend towards **Script**, aiming to document a frenetic moment in time rather than achieving

Figure 6: Clarity instruction and feature context.