Extending deliberate degradation of an artificial neural language model to induce dementia-like deficits

Anonymous ACL submission

Abstract

Recent advances in speech and language technologies aim to leverage clinical information embedded in a person's language abilities to automatically assess cognitive health and function. In this work, we investigate possible perturbations of large language models that could lead to behaviors compatible with those observed in clinical conditions. In particular, we perturb GPT-2 to observe the impact on a generation task used to assess Alzheimer's dementia (AD). Our work achieves statistically significant degradation of the model, and additional classification experiments demonstrate that lexico-syntax is the most impacted linguistic apparatus during deliberate degradation of GPT-2. These findings could inform diagnostic pathways and medical interventions of AD.

1 Introduction

By 2050, the global population of people aged 60 years and older is expected to double to 2.1 billion people (of Economic and Division, 2017). For Alzheimer's dementia (AD), age is the strongest known risk factor (Organization et al., 2017), as the brain becomes more damaged over time, and this necessitates improved strategies for detection to provide timely interventions for the best outcomes possible (Porsteinsson et al., 2021). AD is a clinical condition that leads to cognitive impairment and decline. Subtle changes in a person's speech and language can offer insights into the nature of such decline, particularly in cognitive-linguistic structures and their function in the brain. The battery of tests employed during diagnosis entails a significant speech and language assessment component which can be leveraged therein (Hernández-Domínguez et al., 2018; Sanborn et al., 2022).

In this context, computational methods can offer a framework to simulate cognitive decline and approximate or simulate the linguistic deficits that arise in patients diagnosed with AD (Borge-Holthoefer et al., 2011; Li et al., 2022). For instance, neural deep learning (DL) models, which have proven to be useful on classification tasks among others (de la Fuente Garcia et al., 2020), have also been investigated in the context of classifying clinical conditions, such as in the Alzheimer's Dementia Recognition through Spontaneous Speech (ADReSS) Challenge (Luz et al., 2020). Insights from such investigations have potential for deriving knowledge that may guide clinical directions (Mota et al., 2012). However, this requires bespoke approaches that take into account the characteristics of the base model. For instance, using DL models can be challenging due to the quantity and quality of domain-specific data to clinical conditions, which require novel methodologies. Moreover, the particularities of DL architectures may also play a role in the results obtained.

In this work, we investigate a method of degrading LMs to understand the impact on language use and the linguistic apparatuses that underlie them, building on the approach proposed by (Li et al., 2022). Although the brain is extremely complex and we cannot yet align the inner workings of the brain exactly to computational models, to explore how cognitive decline affects linguistic apparatuses in those diagnosed with AD, we simulate this decline through deliberate degradation of a generative LM. Evaluation of how this degradation impacts specific linguistic abilities of the LMs focuses on syntactic and semantic tasks. We also investigate the impact of degradation of different parts of the architecture on performance, concentrating on transformer-based models, given their wide adoption for language tasks. In particular we aim to answer the following core research questions:

- Given their opacity, how might we effectively compare the degradation in deep neural models and the brain?
- To what extent this method to simulate cog-

nitive decline in neural LMs reflects the way in which such decline manifests in humans diagnosed with AD?

- What linguistic apparatus is most likely to be affected in AD?
- Are the effects specific to parts of the architecture? Or are they uniform and robust?

This paper joins other computational evaluations of early detection of cognitive decline leading to AD (Hernández-Domínguez et al., 2018). It starts with a discussion of related work (§2) and of the methods adopted (§3). The results (§4) are discussed (§5) along with conclusions and future work (§6). Insights from these studies may inform diagnostic pathways and therapeutic treatments involving language for those diagnosed with AD.

2 Related Work

Reported changes in language use due to cognitive decline affect different apparatuses of language, including aspects of syntax and semantics. For example, a decrease in lexical diversity and longitudinal changes in lexical choices preceding an AD diagnosis were reported by Berisha et al. (2015) and corroborated by Aramaki et al. (2016); Kavé and Dassa (2018); Vincze et al. (2022); Lira et al. (2014). From a computational perspective, these changes have been modeled with machine learning (ML) classifiers trained on features derived from language samples from the target groups. For instance, different dementia types were classified on the basis of semantic verbal fluency tasks and on features derived from word embeddings (Paula et al., 2018) or from speech graphs modeling speech as a series of nodes (representing the words) and edges (representing the temporal sequence in which the words were spoken) (Bertola et al., 2014).

Moreover, in certain neurological disorders, semantic memory can be impaired, and, for instance, people with AD often find it increasingly difficult to categorize and name items as their memory deficits worsen, which is one known behavior attributed to a word finding difficulty (Almor et al., 1999). The network theory of semantic memory (Collins and Loftus, 1975) has formed a basis for computational modeling. Degradation across the semantic network causes particular difficulty on explicit semantic tasks, such as picture naming and word-picture matching (Altmann and McClung, 2008), and an unexpected "hyperpriming" effect has been known to occur in people with AD (Chertkow et al., 1989; Rogers and Friedman, 2008). Using percolation theory, (Borge-Holthoefer et al., 2011) modeled a form of cognitive degradation of the semantic memory to simulate this abnormal semantic priming effect by using semantic, free association networks created from psycholinguistic tests (Nelson et al., 1998). The consequences of this global degradation are an impoverished network, where some relationships are reinforced and other weaker links disappear altogether, corroborated by its effect in humans (Chertkow et al., 1989). In another study, data from participants responding to a virtual, onscreen agent regarding questions about their memory and well-being could be used to distinguish between AD and Mild Cognitive Impairment groups using a fully automated classification system (CognoSpeak, O'Malley et al. (2021)). These innovative projects may help define new diagnostic pathways to address a lack of accessibility to screening services for cognitive decline, accelerating waiting times and clinical directions, among other benefits.

The availability of data from initiatives like the Alzheimer's Dementia Recognition using Spontaneous Speech (ADReSS) challenge (Luz et al., 2020) (which has become the most commonly used dataset for AD detection (Ševčík and Rusko, 2022)), has enabled a wealth of new research examining the applicability of advances in natural language processing (NLP) and speech processing techniques. For instance, in response to methodological challenges of using DL models on limited data Li et al. (2022) presents a novel approach to deliberate degradation, perturbing DL transformer models by modifying parameters in the architecture, approaching state-of-the-art performance (SOTA) on ADReSS data using a paired perplexities approach.

In this work, we extend the methodology of Li et al. (2022) in a novel way to investigate how modifying additional parameters in the DL transformer models' structure impacts its performance on a text generation task. We aim to elucidate model degradation as a future avenue for exploring the impacts of cognitive decline on linguistic function. We investigate the effects on generation and on semantic tasks to determine if these are compatible with empirical data. We also examine vulnerabilities of different parts of the architecture and how these perturbations affect performance.

3 Methods

The methods described in this work extend a technique by Li et al. (2022) of deliberate degradation of GPT-2 (Radford et al., 2019) to understand how damaging linguistic apparatuses impacts text generation. We compare the results to the impact that AD has on a person's performance on a speech elicitation task: the Cookie Theft Description Task (Goodglass et al., 2001).

Two versions of GPT-2 are used for evaluation, following Li et al. (2022): the off-the-shelf GPT-2, taken as the "control" model, and the degraded and impaired versions of GPT-2 as "GPT-D". They are probed to generate text based on a synthetic Cookie Theft Picture Description narrative created by Bird et al. (2000). As neural LMs are sensitive to lexical frequency (Cohen and Pakhomov, 2020), lexical frequency and type-to-token ratio (TTR) are calculated, and a two-sided Welch's t-test is used to obtain p-values. The results are further evaluated by classifying the text generated using BERT (Devlin et al., 2018) fine-tuned on the ADReSS dataset and the Corpus of Linguistic Acceptability (CoLA) (Warstadt et al., 2019).

3.1 Datasets

ADReSS (Luz et al., 2020) is a fully balanced dataset in terms of age and gender containing responses from participants with and without a diagnosis of AD to the Cookie Theft Picture Description Task. We use the transcriptions available in the CHAT transcription format (MacWhinney, 2009).

Additional classifiers were trained on the tokenized in-domain set of the **Corpus of Linguistic Acceptability (CoLA)**, which contains sentences sampled from published linguistic works and annotated for grammatically (Warstadt et al., 2019).

As using additional descriptions of the Cookie Theft Picture Description Task seems to improve classification performance (Guo et al., 2021), we use the descriptions from the **CognoSpeak dataset**, which includes 41 control and 24 dementia transcripts across a variety of ages and gender groups. These include both manual and automatically recognised speech transcriptions.

To test semantic understanding we use the **LAnguage Modeling Broadened to Account for Discourse Aspects** (LAMBADA) dataset (Paperno et al., 2016), consisting of narrative passages that humans can complete given the rest of the passage, as such, models should predict the final word of a passage. LAMBADA was used to evaluate language understanding in the original GPT2 (Radford et al., 2019) and decent performance was shown.

3.2 Degrading a transformer model

To modify and degrade GPT-2 to explore impact on its text generation abilities, we extend the method of Li et al. (2022). However, our motivation for using the same transformer model (GPT-2) diverges: while Li et al. (2022) motivate the use of GPT-2 for experimentation because it was found to be arguably the most cognitively plausible transformer model, in this work we do not explore the cognitive plausibility argument from (Schrimpf et al., 2021).

3.2.1 GPT-2 Impairment

GPT-2, a generative transformer model pre-trained on English data (Radford et al., 2019), is used to generate additional text based off a synthetic Cookie Theft Picture description (Bird et al., 2000). From GPT-2 (simple) several impairment configurations are created by breaking the attention heads at a number of different layers within the self-attention mechanism.[1] "Impairment" here refers to masking, or zeroing, the values in different patterns which "degrades" the model, and the impairment patterns were informed by Vig and Belinkov (2019) who analyzed the interaction between attention in transformer models and syntax. We hypothesize that breaking the attention heads using various styles and combinations of layers will affect the text generated from the resulting model. In other words, it removes its access to values in the attention layers and heads. By impairing the internal structures that store specific kinds of linguistic information, we investigate how the loss of such information imbued in the layers, caused by zeroing the values, may lead to generated text that resembles the speech of those diagnosed with AD.

3.2.2 Artificial Impairment: Locations

To determine the portions of values and locations at which we will perform the artificial impairment, we follow Li et al. (2022), who found that the impairment of 50% of the values (out of 25%, 50%, 75% and 100%) at the corresponding locations, yielded the best results. However, unlike Li et al. (2022),

[1] Functionalities for these experiments are from Li et al. (2022) available in https://github.com/LinguisticAnomalies/hammer-nets/

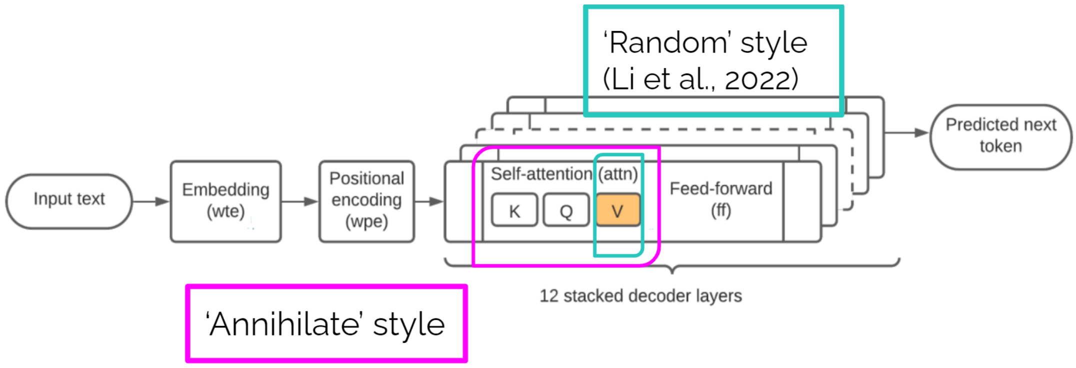

Figure 1: GPT-2 architecture and GPT-D impairment styles

we focus exclusively at the self-attention mechanism instead of other areas of the GPT-2 transformer architecture, since they found that patterns of artificial impairment at other locations, namely the embeddings and feed-forward network components, did not yield the expected impact on producing different discourse.

3.2.3 Artificial Impairment: Patterns

Within the 12 layers and 12 attention heads per layer, we followed a number of different combinations of impairing the layers and attention heads. The self-attention mechanism in GPT-2 contains concatenated Query-Key-Value matrices that precede a feed-forward layer. We use the **'random' (RAN)** masking style (Li et al., 2022), in which the values in the attention heads are randomly set to zero exclusively at the Values matrix, "as their parameters directly determine the content of the vectors that are passed onto the subsequent feed-forward layer" (Li et al., 2022).

We extend this by impairing the parameters of the entire concatenated Query-Key-Value matrices under a new type of impairment pattern called **"annihilate" (ANH)**. We want to explore how much the generated text will be affected when both the amount of attention directed towards items in the sentence sequence and to where the mechanism is directing its attention are impaired. The ANH pattern considers all possible parameters in the concatenated Query-Key-Value matrices for masking instead of just those at the Value matrix, which can be seen to provide information pertaining to the impact of each token on a given token's representation. As such, zeroing out the parameters would remove the impact of that token in calculating the self-attention of the other tokens. The two impairment styles are indicated in Figure 1.

Similarly, we hypothesize that zeroing out the parameters of the attention matrix corresponding to the Key and the Query will additionally divert the self-attention away from the tokens that would ordinarily be used in calculating a token's representation. This is because the Query and the Key provide the relative importance of different tokens in calculating the representation of a given token.

The impairment patterns were further motivated by analyses of the structure of attention in transformers, focusing on different properties of syntax and its interplay with attention at different layer depths (Vig and Belinkov, 2019). We frame our investigation of observing the impact of AD by adopting a division between syntax and semantics. As such, the patterns of impairment are as follows:

- **Layers 1-6**, seem to align syntactic dependencies with attention most strongly (Vig and Belinkov, 2019), and we expect that masking the parameters at these layers will produce the most impacted text generated from the GPT-D model(s) in terms of syntactic correctness and grammaticality.

- **Layers 6-12** seem to capture the longest-range relationships and semantic information (Vig and Belinkov, 2019; Belinkov et al., 2018), and we expect that masking at these layers will impact the generated text differently than at layers 1-6, with less of an effect on syntactic correctness and grammaticality.

3.2.4 Evaluation and Metrics

We measure the effect of impairing attention layers by using the generated text from GPT-2 and GPT-D to calculate the p-value using a two-sided Welch's t-test. The p-value measures the statistical significance in the difference between GPT-2 and various

	Lexical frequency		Type-to-Token Ratio		p-values
Impairment configurations	*GPT-2*	*GPT-D*	*GPT-2*	*GPT-D*	
Layers 1-6 (RAN)	0	0.25	0.71	0.40	0.083
Layers 7-12 (RAN)	0	0.08	0.72	0.64	0.159
Layers 1-6 (ANH)	**0**	**0.4**	**0.74**	**0.55**	**0.005***
Layers 7-12 (ANH)	0	0	0.73	0.52	NaN

Table 1: Results of p-values calculated from two-sided Welch's t-test, lexical frequency values, and type-to-token ratio (TTR) values

GPT-D models (**p < 0.05). This p-value score captures two key word repetition metrics, lexical frequency and TTR (Li et al., 2022), which have shown to draw parallels with linguistic patterns produced by those with AD. For this framework, this serves as the measure to determine if a text is "dementia-like." Although previous research has found that those with AD tend to exhibit word repetitions (Bucks et al., 2000; Berisha et al., 2015), suggesting it as a linguistic anomaly that may be indicative of dementia-like speech, there have also been conflicting findings about the effects of AD on language use (Altmann and McClung, 2008).

4 Results

4.1 GPT-2 and GPT-D Impairment

The control GPT-2 and the degraded GPT-D models are probed with a beam search to generate the next best, non-empty 20 tokens following a synthetic Cookie Theft picture description (Bird et al., 2000).[2] We use the p-value as a measure of statistical significance between the control and the degraded models, to evaluate the impact of the impairment experiments. Based on the results by Li et al. (2022), in accordance with the linguistic deficits that occur in those with dementia, the generated text from GPT-D is expected to have higher lexical frequency values and lower TTR values, and the statistical significance to be observed more saliently in the impairment configurations that take place in the initial layers of the model. The results in Table 1 mostly align with these expectations.

While the TTR values are consistently lower for the GPT-D than for the GPT-2 counterparts as expected, there is no pattern for the effect on the initial 6 layers for the TTR values. There is, though, a pattern of higher lexical frequency values for the initial 6 layers in both the **RAN** and **ANH** styles.

4.2 Dementia Evaluation

While the findings on the p-value metric is consistent with those by Li et al. (2022), perhaps statistical significance in word repetition (i.e., lexical frequency and TTR) is not the only characteristic affected in those with AD. We investigate this for the p-value metric by fine-tuning BERT classifiers on other datasets to see if BERT can accurately classify speech from a 'control' group versus a 'dementia' group of participants in the ADReSS dataset.

Following (Li et al., 2022), we experimented on BERT and DistilBERT, a lighter, distilled version of BERT that retains 40% of the parameters while still retaining 95% accuracy of BERT models (Sanh et al., 2019). [3] Each participant response to the Cookie Theft picture description task averaged 445 words and was fed into the model as one sample for fine-tuning. The results of these fine-tuning experiments are detailed in Table 8 in the appendix section. Our best model on the evaluation accuracy (T5) on the BERT ('bert-base-uncased') model approaches SOTA classification performance using the ADReSS test set by (Balagopalan et al., 2020).

What is particularly surprising is that GPT-D output probabilities for the dementia label were classified as from the 'control' group, even though our best BERT classifier, fine-tuned on ADReSS, approaches SOTA performance on the test set shown in Table 2. We acknowledge that the GPT-D output probabilities are marginally higher than those of GPT-2, except for the impairment configuration "Layers 7-12 (ANH)."

To this end, we verify the viability of this BERT classifier by feeding our BERT classifier the tran-

[2]Additional information about the beam search for this language generation can be found in (Li et al., 2022) and the text generation scripts in https://github.com/LinguisticAnomalies/hammer-nets/

[3]Pre-trained models were publicly available through OpenAI and the huggingface library and fine-tuned (Wolf et al., 2020).

	Probability of Dementia Classification	
Impairment configurations	*GPT-2 Outputs*	*GPT-D Outputs*
Layers 1-6 (RAN)	33.77 %	35.17 %
Layers 7-12 (RAN)	34.92 %	38.16 %
Layers 1-6 (ANH)	34.17 %	37.13 %
Layers 7-12 (ANH)	34.11 %	32.72 %

Table 2: Dementia evaluation of GPT-2/GPT-D outputs

scripts from the CognoSpeak dataset. As these transcripts are also in response to the Cookie Theft picture description task, they are comparable to the ADReSS data and therefore can effectively measure the viability of our classification task.

While it is unsurprising that the control transcripts were classified as 'dementia' with only a 3.83% probability (Table 3), we were surprised to see the dementia transcripts classified as 'dementia' with a percentage well below chance at 26.68%. While this is still greater than the probability with which the control transcripts were classified, it is still not high enough to find our BERT classifier as a viable way to distinguish speech from dementia participants or verify the p-value metric findings. To this end, we conclude that a BERT model fine-tuned on ADReSS data cannot sensibly classify text as 'control' or 'dementia.' We look to explore an additional BERT classification task fine-tuned on a different dataset to verify them instead.

Probability of Dementia Classification	
Control Transcripts	*Dementia Transcripts*
3.83 %	26.68%

Table 3: Dementia evaluation of CognoSpeak data

4.3 Syntactic Evaluation

The generated outputs from GPT-D were found to be different and "dementia-like" in comparison to those of GPT-2 with statistical significance, particularly in regards to the lexico-syntax apparatus. As such, we assess the grammaticality, or syntactic correctness, of the outputs to support this result.

We fine-tune a BERT model on CoLA (Warstadt et al., 2019) and report the cumulative results in Table 9 of the appendix. The best performing model, T3, achieves an accuracy of 83.9% on the validation dataset, and 84.21% accuracy on the test set. Table 4 shows that our GPT-D model, impaired at the initial 6 layers using the RAN and ANH styles, produced outputs that are found to be only 2.51% and 3.43% linguistically acceptable, respectively, which aligns with expectations.

	Percentage of Linguistic Acceptability	
Impairment configurations	GPT-2 Outputs	GPT-D Outputs
Layers 1-6 (RAN)	96.36 %	**2.51 %**
Layers 7-12 (RAN)	98.22 %	96.68 %
Layers 1-6 (ANH)	99.99 %	**3.43 %**
Layers 7-12 (ANH)	99.99 %	93.71 %

Table 4: Linguistic acceptability: GPT-2/GPT-D outputs

As a final measure, we use the CoLA classifier on the ADReSS and CognoSpeak data themselves to see if its findings align with our hypotheses on how AD may impact the syntax apparatus. In contrast with our expectations, as shown in Table 5, the control transcripts in both datasets are classified as less linguistically acceptable than the dementia transcripts.

	Percentage of Linguistic Acceptability	
Dataset	Control Transcripts	Dementia Transcripts
ADReSS	5.79 %	8.08 %
CognoSpeak	23.45 %	18.54 %

Table 5: Linguistic acceptability: ADReSS & CognoSpeak

4.4 Semantic Evaluation

To evaluate the effect on semantic understanding we employ the same impairment framework used in the previous tasks on the LAMBADA dataset. We also introduce 2 other variants on the RAN strategy, RAN-Q and RAN-K, which impair the Query and Key matrices respectively, rather than the Value matrix.

The results show that impairment in the lower layers of the model (1-6) has the highest effect on the performance in this semantic understanding task across all impairment configurations, contrary to the suggestions of (Vig and Belinkov, 2019). We also see that RAN-Q impairment has a larger impact on performance than RAN-K and RAN impairment, with the level of degradation similar to that of the ANH impairment configuration.

Impairment configurations	Accuracy
Layers 1-6 (RAN)	20.7%
Layers 1-6 (RAN-K)	9.93%
Layers 1-6 (RAN-Q)	3.62%
Layers 1-6 (ANH)	4.05%
Layers 7-12 (RAN)	33.04%
Layers 7-12 (RAN-K)	20.03%
Layers 7-12 (RAN-Q)	10.90%
Layers 7-12 (ANH)	11.76%

Table 6: Accuracy on the LAMBADA Dataset, averaged across 10 runs.

5 Discussion

5.1 GPT-2 Impairment Evaluation

The p-value metric calculation from (Li et al., 2022) determines if generated text from GPT-2 can be said to reflect the linguistic anomalies that occur in the speech of those with AD. Our experiments impairing GPT-2 into various degraded configurations of GPT (GPT-D) indeed generated text based off of the synthetic Cookie Theft picture description task (Bird et al., 2000) that was distinct from GPT-2.

Table 1 details the lexical frequency, TTR, and p-value measures of GPT-D's generated text following impairment. ANH in layers 1-6 is the only impairment pattern that shows a significant difference between GPT-2 and GPT-D. However, the p-value is still lower for both impairment styles in layers 1-6 compared with RAN in layers 7-12. The implication of these results is two-fold:

- **Degradation is more impactful in the first 6 layers than the last 6 layers of attention.** The GPT-D model impaired with the ANH pattern in the first 6 layers of attention produced the highest lexical frequency value, which aligns with increased word repetitions.

The next highest lexical frequency used the RAN impairment pattern in layers 1-6, with lower lexical frequency from impairments in layers 7-12. This may indicate a difference between impairment at layers 1-6 and 7-12 for RAN, but the results do not show a statistically significant difference between GPT-2 and GPT-D for the p-value. The p-value for the deeper 6 layers was inconclusive.

To explain this, the first 6 layers of attention have been found to be most strongly aligned with syntactic dependencies, according to the dependency alignment metric established by (Vig and Belinkov, 2019). We see a significant difference in generated outputs between GPT-2 and GPT-D, implying the impairment at layers 1-6 has effectively damaged the model's syntactic apparatus.

Additionally, we observe a lack of p-value significance when damaging the last 6 layers, with the impaired model producing text more similar to the control GPT-2 model. Researchers have found that while the initial layers in the self-attention mechanism encode lower-level syntactic structures of language, the deeper layers may be more responsible for encoding higher-level syntactic information and even semantics (Vig and Belinkov, 2019), suggested to be due to the 'global perspective' afforded to them (Belinkov et al., 2018).

- **Attention is more impacted using the ANH pattern in comparison to the RAN pattern of impairment.** Additionally impairing the self-attention mechanism at the Query and Key matrices produces the most impactful difference in the linguistic apparatuses encoded within attention.

The RAN masking style was originally designed to perturb the Value matrix. Multiplying by the Value matrix is thought to "generate a semantic representation of each token" (Li et al., 2022), and so, zeroing out it's parameters would remove the impact of a given token's representation in calculating the self-attention of the other tokens.

Similarly, we speculated that because the key and query matrices provide the relative importance of each token to the attention calculation, zeroing out these parameters would divert self-attention away from the ordinarily used tokens. This zeroing strategy - which we call **ANH** – was expected to cause the greatest impact. This was confirmed by our results and may be attributed to the self-attention mechanism's ability to formulate representations of words at lower-levels of language, including syntax. This would reflect the changes in language use in terms of lexical richness and grammatical structure in adults with AD, as demonstrated by (Bucks et al., 2000).

5.2 Syntactic Evaluation

Our dementia evaluation experiments yielded mixed results. While the best performing BERT model fine-tuned on ADReSS approached those of other baseline and SOTA models (Meghanani et al., 2021; Balagopalan et al., 2020), it was not able to accurately classify text as 'control' or 'dementia.' The p-value metric suggests that the nature of the degradation may not resemble dementia in a lexico-syntactic way, supported by previous findings that the syntactic abilities of "mildly or moderately demented" patients remain relatively intact (Murdoch et al., 1987) even in written language (Kemper et al., 1993), though such work may have other implications beyond the scope of the spontaneous speech data we used. Degradation may instead extend further into the apparatuses responsible for storing semantic memory (Hier et al., 1985; Nebes, 1989; Almor et al., 1999; Altmann and McClung, 2008).

This is not to contest that there are statistically significant differences in lexico-syntactic measures of word repetition in the text generated by both GPT-2 and GPT-D models, mirrored in the type of decline found in the speech of those with AD, particularly in terms of lexical diversity and richness and syntactic complexity (Bucks et al., 2000; Berisha et al., 2015), correlating further with the Mini-Mental State Examination (Hernández-Domínguez et al., 2018; Kavé and Dassa, 2018). By fine-tuning BERT on the CoLA dataset, the classifier verifies this difference and predicts that 96-97% of the generated outputs from the GPT-D models impaired at the initial 6 layers are deemed linguistically *unacceptable*. This is in contrast to the classification results on outputs produced from all other impairment configurations, which our classifier finds to be linguistically acceptable.

However, in contrast with our expectations, the CoLA classifier found the control transcripts in both ADReSS and CognoSpeak datasets to be less linguistically acceptable than the dementia transcripts. It is important to note this difference in classification findings on the human data in ADReSS and CognoSpeak from the findings on data generated by GPT-2. This suggests a fundamental difference in how degradation transpires in the human brain versus that which can be induced in a LM generating experimental data, aligning with our stance of not adopting GPT-2 as a proxy to the human brain. The investigations in this work explore deliberate degradation of an artificial LM and the deficits induced as a consequence of such perturbations, which are importantly from a LLM perspective. Nevertheless, such work can inspire possible avenues for exploring the impact cognitive decline on linguistic function, as increasingly more advanced AI and language technologies emerge.

Our findings can support the potential for clinicians to utilize speech elicitation tasks during assessment and diagnosis that target grammaticality to assess the cognitive health, an approach that has been supported by findings in research (Hernández-Domínguez et al., 2018). Research has also demonstrated the utility of correlating clinicians' assessments of speech and language to automated analyses conducted using NLP techniques (Yeung et al., 2021). Speech therapies may also aim to reinforce skills in grammar and syntax as a result.

6 Conclusions

The present work sought to validate an effective way to simulate degradation in generative transformer-based LMs that is comparable to the cognitive decline of AD. The deliberate degradation approach introduced by (Li et al., 2022) allows for experimentation on and probing of computational models to generate language that may otherwise be inaccessible in real-life clinical settings with patients. Our novel extension provides insight into which linguistic apparatuses may be impacted during cognitive decline, and joins other computational methods to elucidate the linguistic apparatuses that are most severely impacted in those with AD. The main contributions of this work include:

- Creation of a new impairment style called 'annihilate' building upon (Li et al., 2022), which yields more significant results on the linguistic apparatuses
- Corroboration with existing literature regarding linguistic deficits that occur during cognitive decline, further demonstrating the potential utility for the degradation approach

The value of such work lies in its potential for informing clinical directions preceding a diagnosis of AD and/or other forms of cognitive impairment, and the therapeutic treatments that follow.

7 Limitations

7.0.1 Datasets

The use of the datasets involving human participants utilized in this work, namely ADReSS and CognoSpeak, received full ethics approval. As the data in the ADReSS and CognoSpeak datasets consist of responses provided by human participants, the data were fully anonymized and cannot be linked back to the individuals who provided the responses. While the ADReSS data can be made publicly available, access to this data must be requested from the organizers of the challenge.

7.0.2 Methodology

Our methodology operates under the assumption that our experiments do not attempt to or suggest the idea of replacing professional medical advice and evaluation that is required to receive any clinical diagnosis. Computational modeling should not be used to determine or diagnose human clinical conditions. While advances in computational psychiatry and ML models have become extremely powerful in the tasks they can perform in terms of human language, they are purely models in themselves. They are not exact or fully accurate models of the human brain, nor do they fully begin to capture the extremely complex inner workings and structures of the human brain, which neuroscientists, researchers, and other professionals have yet to fully understand. These models allow us to perform a variety of experimental tests that we understand are prone to error and human biases that are derived from the data and engineers that are involved in the training process. Therefore, these models are unable to draw definitive inferences in real-world, clinical settings. Testing on computational models to understand complex neurodegenerative change is not ideal, but we hope that they may give us clues into structural deterioration. They serve as an alternative to otherwise costly and potentially time-consuming methods of studying the apparatuses that impact language use.

We clarify that while we previously utilized and will henceforth utilize the term "control" to refer to model and data associated with language generated and derived from neurotypical individuals, we do not claim that this group of individuals is comparatively "normal." The term "control" is simply a way for us to define a standard within the framework of our experiments to how we expect language to be produced so that we may be able to compare and contrast linguistic anomalies. These linguistic anomalies may uncover various types of cognitive and neurological degradation that we otherwise may or may not otherwise associate with cognitive disorders, and give us insight into how we can possibly help guide the direction of clinical assessments of cognitive health.

The aim of this work is to potentially develop a pipeline or framework that allows us to study linguistic phenomena and explore changes in language use when the linguistic apparatuses of LMs are altered. These LMs have been specifically designed and trained on human language tasks, which make them an interesting entrypoint into understanding changes in human language use.

Our study could potentially inform the work of clinicians in how they run human subject-oriented tests that have been well-established in the diagnostic pipeline. We hope that our work would help determine better, more informed ways to assess and treat individuals so that they are able to access necessary medical interventions and treatment as soon as possible.

References

Amit Almor, Daniel Kempler, Maryellen C MacDonald, Elaine S Andersen, and Lorraine K Tyler. 1999. Why do alzheimer patients have difficulty with pronouns? working memory, semantics, and reference in comprehension and production in alzheimer's disease. *Brain and language*, 67(3):202–227.

Lori JP Altmann and Jill S McClung. 2008. Effects of semantic impairment on language use in alzheimer's disease. In *Seminars in speech and language*, volume 29, pages 018–031. © Thieme Medical Publishers.

Eiji Aramaki, Shuko Shikata, Mai Miyabe, and Ayae Kinoshita. 2016. Vocabulary size in speech may be an early indicator of cognitive impairment. *PloS one*, 11(5):e0155195.

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*.

Yonatan Belinkov, Lluís Màrquez, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, and James Glass. 2018. Evaluating layers of representation in neural machine translation on part-of-speech and semantic tagging tasks. *arXiv preprint arXiv:1801.07772*.

Visar Berisha, Shuai Wang, Amy LaCross, and Julie Liss. 2015. Tracking discourse complexity preceding alzheimer's disease diagnosis: a case study comparing the press conferences of presidents ronald reagan and george herbert walker bush. *Journal of Alzheimer's Disease*, 45(3):959–963.

Laiss Bertola, Natália B Mota, Mauro Copelli, Thiago Rivero, Breno Satler Diniz, Marco A Romano-Silva, Sidarta Ribeiro, and Leandro F Malloy-Diniz. 2014. Graph analysis of verbal fluency test discriminate between patients with alzheimer's disease, mild cognitive impairment and normal elderly controls. *Frontiers in aging neuroscience*, 6:185.

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

Javier Borge-Holthoefer, Yamir Moreno, and Alex Arenas. 2011. Modeling abnormal priming in alzheimer's patients with a free association network. *PloS one*, 6(8):e22651.

Romola S Bucks, Sameer Singh, Joanne M Cuerden, and Gordon K Wilcock. 2000. Analysis of spontaneous, conversational speech in dementia of alzheimer type: Evaluation of an objective technique for analysing lexical performance. *Aphasiology*, 14(1):71–91.

Howard Chertkow, Daniel Bub, and Mark Seidenberg. 1989. Priming and semantic memory loss in alzheimer's disease. *Brain and language*, 36(3):420–446.

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*.

Allan M Collins and Elizabeth F Loftus. 1975. A spreading-activation theory of semantic processing. *Psychological review*, 82(6):407.

Sofia de la Fuente Garcia, Craig W Ritchie, and Saturnino Luz. 2020. Artificial intelligence, speech, and language processing approaches to monitoring alzheimer's disease: a systematic review. *Journal of Alzheimer's Disease*, 78(4):1547–1574.

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*.

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

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.

Daniel B Hier, Karen Hagenlocker, and Andrea Gellin Shindler. 1985. Language disintegration in dementia: Effects of etiology and severity. *Brain and language*, 25(1):117–133.

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

Susan Kemper, Emily LaBarge, F Richard Ferraro, Hintat Cheung, Him Cheung, and Martha Storandt. 1993. On the preservation of syntax in alzheimer's disease: Evidence from written sentences. *Archives of neurology*, 50(1):81–86.

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*.

Juliana Onofre de Lira, Thaís Soares Cianciarullo Minett, Paulo Henrique Ferreira Bertolucci, and Karin Zazo Ortiz. 2014. Analysis of word number and content in discourse of patients with mild to moderate alzheimer's disease. *Dementia & neuropsychologia*, 8:260–265.

Saturnino Luz, Fasih Haider, Sofia de la Fuente, Davida Fromm, and Brian MacWhinney. 2020. Alzheimer's dementia recognition through spontaneous speech: The adress challenge. In *INTERSPEECH 2020*, pages 2172–2176. ISCA.

Brian MacWhinney. 2009. The childes project part 1: The chat transcription format.

Amit Meghanani, CS Anoop, and Angarai Ganesan Ramakrishnan. 2021. Recognition of alzheimer's dementia from the transcriptions of spontaneous speech using fasttext and cnn models. *Frontiers in Computer Science*, 3:624558.

Natalia B Mota, Nivaldo AP Vasconcelos, Nathalia Lemos, Ana C Pieretti, Osame Kinouchi, Guillermo A Cecchi, Mauro Copelli, and Sidarta Ribeiro. 2012. Speech graphs provide a quantitative measure of thought disorder in psychosis. *PloS one*, 7(4):e34928.

Bruce E Murdoch, Helen J Chenery, Vicki Wilks, and Richard S Boyle. 1987. Language disorders in dementia of the alzheimer type. *Brain and language*, 31(1):122–137.

Robert D Nebes. 1989. Semantic memory in alzheimer's disease. *Psychological bulletin*, 106(3):377.

Douglas L Nelson, Cathy L McEvoy, and Th A Schreiber. 1998. University of south florida free association norms. *URL: http://w3. usf. edu/FreeAssociation/(: 08.04. 2017)*.

United Nations. Department of Economic and Social Affairs. Population Division. 2017. *World population ageing: 2017 highlights*. UN.

Ronan Peter Daniel O'Malley, Bahman Mirheidari, Kirsty Harkness, Markus Reuber, Annalena Venneri, Traci Walker, Heidi Christensen, and Dan Blackburn. 2021. Fully automated cognitive screening tool based on assessment of speech and language. *Journal of Neurology, Neurosurgery & Psychiatry*, 92(1):12–15.

World Health Organization et al. 2017. Global action plan on the public health response to dementia 2017–2025.

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Ngoc Quan Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. 2016. The LAMBADA dataset: Word prediction requiring a broad discourse context. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1525–1534, Berlin, Germany. Association for Computational Linguistics.

Felipe Paula, Rodrigo Wilkens, Marco Idiart, and Aline Villavicencio. 2018. Similarity measures for the detection of clinical conditions with verbal fluency tasks. *Proceedings of NAACL-HLT 2018*, pages 231–235.

AP Porsteinsson, RS Isaacson, Sean Knox, MN Sabbagh, and I Rubino. 2021. Diagnosis of early alzheimer's disease: Clinical practice in 2021. *The Journal of Prevention of Alzheimer's Disease*, 8(3):371–386.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Sean L Rogers and Rhonda B Friedman. 2008. The underlying mechanisms of semantic memory loss in alzheimer's disease and semantic dementia. *Neuropsychologia*, 46(1):12–21.

Victoria Sanborn, Rachel Ostrand, Jeffrey Ciesla, and John Gunstad. 2022. Automated assessment of speech production and prediction of mci in older adults. *Applied Neuropsychology: Adult*, 29(5):1250–1257.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Martin Schrimpf, Idan Asher Blank, Greta Tuckute, Carina Kauf, Eghbal A Hosseini, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. 2021. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45):e2105646118.

Adam Ševčík and Milan Rusko. 2022. A systematic review of alzheimer's disease detection based on speech and natural language processing. In *2022 32nd International Conference Radioelektronika (RADIOELEKTRONIKA)*, pages 01–05. IEEE.

Jesse Vig and Yonatan Belinkov. 2019. Analyzing the structure of attention in a transformer language model. *arXiv preprint arXiv:1906.04284*.

Veronika Vincze, Martina Katalin Szabó, Ildikó Hoffmann, László Tóth, Magdolna Pákáski, János Kálmán, and Gábor Gosztolya. 2022. Linguistic parameters of spontaneous speech for identifying mild cognitive impairment and alzheimer disease. *Computational Linguistics*, 48(1):119–153.

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Anthony Yeung, Andrea Iaboni, Elizabeth Rochon, Monica Lavoie, Calvin Santiago, Maria Yancheva, Jekaterina Novikova, Mengdan Xu, Jessica Robin, Liam D Kaufman, et al. 2021. Correlating natural language processing and automated speech analysis with clinician assessment to quantify speech-language changes in mild cognitive impairment and alzheimer's dementia. *Alzheimer's research & therapy*, 13(1):1–10.

A Appendix A: BERT Training Results

The BERT experiments were carried out using the Google Colab Pro infrastructure.

For the binary dementia classification task, the BERT models were fine-tuned using an 80/20 split on the training data. The ADReSS training set resulted in 86 samples for training, 22 samples for evaluation, and 48 samples for the test set. Shown in Table 8, experiments T1 to T11 employed our own fine-tuning parameters and those defined by (Devlin et al., 2018) were used for experiments T12 to T19 on the BERT and DistilBERT models. Each of the trials was run 5 times over 5 random seeds and the accuracies were averaged into a single accuracy score. Using the best BERT model and hyperparameters from T5, the generated text from GPT-2 and GPT-D were classified ('control' or 'dementia') for each impairment configuration using a softmax function.

For the linguistic acceptability classification task, the BERT models were fine-tuned with the suggested hyperparameter values from (Devlin et al., 2018). Each of the trials from T1 to T5 was run 5 times over 5 random seeds and the accuracies were averaged into a single accuracy score. The results are reported in Table 9.

Trial	# of epochs	train batch size	eval batch size	warmup steps	learning rate	weight decay	# of runs	avg eval accuracy	avg test accuracy
T1	3	16	32	2	1E-06	0	5	49.09%	50.83%
T2	10	16	32	2	1E-06	0	5	53.64%	57.50%
T3	10	16	32	2	1E-05	0	5	80.91%	78.33%
T4	20	16	32	5	1E-05	0	5	81.82%	80.83%
T5	**50**	**16**	**32**	**5**	**1E-05**	0	**5**	**86.36%**	80.00%
T6	25	16	32	5	5E-05	0	5	81.82%	79.17%
T7	25	16	32	5	1E-04	0	5	77.27%	78.75%
T8	25	16	32	8	5E-05	0	5	81.82%	79.17%
T9	25	16	16	10	5E-05	0	5	69.09%	80.42%
T10	25	16	32	10	1E-05	0	5	72.73%	80.00%
T11	25	16	16	10	1E-05	0	5	72.73%	80.00%
T12	25	16	32	2	1E-05	0.01	5	72.73%	80.00%
T13	25	16	32	2	5E-05	0.01	5	69.09%	80.42%
T14	3	16	32	2	5E-05	0.01	5	74.55%	77.50%
T15	3	16	32	2	2E-05	0.01	5	73.64%	77.50%
T16	3	16	32	2	3E-05	0.01	5	75.45%	78.33%
T17	3	16	32	2	4E-05	0.01	5	72.73%	77.08%
T18	3	16	32	2	1E-05	0.01	5	73.64%	74.58%
T19	3	32	32	2	1E-06	0.01	5	53.64%	51.25%

Table 7: Cumulative results of fine-tuning DistilBERT on ADReSS over 5 runs per trial

Trial	# of epochs	train batch size	eval batch size	warmup steps	learning rate	weight decay	# of runs	avg eval accuracy	avg test accuracy
T1	3	16	32	2	1E-06	0	5	52.73%	55.83%
T2	10	16	32	2	1E-06	0	5	69.10%	67.50%
T3	10	16	32	2	1E-05	0	5	82.73%	81.25%
T4	20	16	32	5	1E-05	0	5	85.45%	79.58%
T5	**50**	**16**	**32**	**5**	**1E-05**	0	**5**	**88.18%**	79.58%
T6	25	16	32	5	5E-05	0	5	84.55%	80.00%
T7	25	16	32	5	1E-04	0	5	82.72%	77.92%
T8	25	16	32	8	5E-05	0	5	84.54%	80.00%
T9	25	16	16	10	5E-05	0	5	84.54%	80.00%
T10	25	16	32	10	1E-05	0	5	87.27%	78.75%
T11	25	16	16	10	1E-05	0	5	87.27%	78.75%
T12	25	16	32	2	1E-05	0.01	5	87.27%	78.75%
T13	25	16	32	2	5E-05	0.01	5	84.55%	80.00%
T14	3	16	32	2	5E-05	0.01	5	71.82%	77.92%
T15	3	16	32	2	2E-05	0.01	5	66.36%	82.08%
T16	3	16	32	2	3E-05	0.01	5	65.45%	80.83%
T17	3	16	32	2	4E-05	0.01	5	66.36%	77.50%
T18	3	16	32	2	1E-05	0.01	5	65.45%	75.00%
T19	3	32	32	2	1E-06	0.01	5	48.18%	54.58%

Table 8: Cumulative results of fine-tuning BERT on ADReSS over 5 runs per trial

Trial	**# of epochs**	**train batch size**	**eval batch size**	**warmup steps**	**weight decay**	**learning rate**	**# of runs**	**avg eval accuracy**	**avg test accuracy**
T1	3	16	32	2	0.01	5E-05	5	83.83 %	84.21%
T2	3	16	32	2	0.01	2E-05	5	82.70%	83.64%
T3	**3**	**16**	**32**	**2**	0.01	**3E-05**	**5**	**83.90%**	**84.21%**
T4	3	16	32	2	0.01	4E-05	5	83.30%	83.80%
T5	3	16	32	2	0.01	1E-05	5	83.09%	84.14%

Table 9: Cumulative results of fine-tuning BERT on CoLA over 5 runs per trial