Too Big to Fail: Larger Language Models are Disproportionately Resilient to Induction of Dementia-Related Linguistic Anomalies

Changye Li[1], Zhecheng Sheng[1], Trevor Cohen[2], and Serguei Pakhomov[1]

[1]University of Minnesota
[2]University of Washington
[1]{lixx3013, sheng136, pakh0002}@umn.edu
[2]{cohenta}@uw.edu

Abstract

As artificial neural networks grow in complexity, understanding their inner workings becomes increasingly challenging, which is particularly important in healthcare applications. The intrinsic evaluation metrics of autoregressive neural language models (NLMs), perplexity (PPL), can reflect how "surprised" an NLM model is at novel input. PPL has been widely used to understand the behavior of NLMs. Previous findings show that changes in PPL when masking attention layers in pre-trained transformer-based NLMs reflect linguistic anomalies associated with Alzheimer's disease dementia. Building upon this, we explore a novel bidirectional attention head ablation method that exhibits properties attributed to the concepts of cognitive and brain reserve in human brain studies, which postulate that people with more neurons in the brain and more efficient processing are more resilient to neurodegeneration. Our results show that larger GPT-2 models require a disproportionately larger share of attention heads to be masked/ablated to display degradation of similar magnitude to masking in smaller models. These results suggest that the attention mechanism in transformer models may present an analogue to the notions of cognitive and brain reserve and could potentially be used to model certain aspects of the progression of neurodegenerative disorders and aging.

1 Introduction

Alzheimer's disease (AD) dementia is a currently incurable neurodegenerative condition that leads to a progressive and irreversible decline in cognitive function. Due to the challenging nature of early diagnosis of this condition, there is a pressing need for efficient and cost-effective screening tools (Bradford et al., 2009) to mitigate the negative consequences of delayed or absent diagnosis (Stokes et al., 2015). Previous studies have demonstrated that changes in cognitive status can be reflected in spoken language and spontaneous speech (Giles et al., 1996; Almor et al., 1999; Hier et al., 1985). Automated analysis of such speech, employing supervised machine learning models, shows its potential as an early screening tool. These models can be trained to identify subtle linguistic anomalies associated with dementia from transcripts of both healthy individuals and those with dementia. Recent advances in machine learning, such as deep learning models and the transformer with attention architecture (Vaswani et al., 2017), have mediated remarkable performance on this downstream task (for a review, see Shi et al. (2023)). Deep learning models, inspired by the human brain, are artificial neural networks (ANNs) that process vast amounts of data and learn complicated patterns, making them well-suited for analyzing subtle linguistic patterns. The transformer architecture, in particular, has advanced performance on natural language processing (NLP) tasks by enabling models to capture long-range dependencies more effectively via the attention mechanism (Vaswani et al., 2017).

As ANNs get larger and more complicated, it becomes even harder to interpret their inner workings. The performance of autoregressive neural language models (NLMs) (e.g., predicting the next word given the context) is frequently estimated with a single somewhat interpretable feature, perplexity (PPL), which has shown to be a suitable measurement for evaluating cognitive impairment from spontaneous speech (Fritsch et al., 2019; Cohen and Pakhomov, 2020). As the name "perplexity" suggests, it can be considered as an indicator of how "surprised" a model is by novel (i.e., not used in model's training) input. The more different the input is from a particular model's training data, the "harder" it is for the model to predict, resulting in higher PPL. Therefore, it is reasonable to hypothesize that PPL may have some degree of diagnostic utility, as an indicator of patterns of language use that fall outside the scope of the typical

Findings of the Association for Computational Linguistics: ACL 2024, pages 6363–6377
August 11-16, 2024

language used to train a model. In the context of AD, changes in language and cognitive function often manifest as differences in language complexity, with individuals experiencing difficulty in forming coherent sentences and selecting appropriate words. As AD progresses, the language used by patients with dementia becomes more unpredictable and less coherent, leading to higher PPL with models trained on language from individuals presumed to be cognitively healthy.

While training data from cognitively healthy individuals is plentiful, language data produced by patients with dementia is much more impractical to obtain in sufficient quantity to train a large NLM. In hyperdimensional computing (Kanerva, 2009), high-dimensional vector representations are manipulated using operators that alter their distance from other learned representations. A prior work inspired by this concept (Li et al., 2022) demonstrates that masking the attention sub-modules of pre-trained transformer-based NLMs and thereby artificially increasing PPL on text from cognitively healthy individuals, can provide an effective solution to the challenge of limited data availability. By strategically altering these sub-modules and introducing controlled perturbations in the NLMs' attention layers, the degraded NLMs induce the linguistic anomalies and unpredictability associated with dementia.

Recent work in neuroscience using functional magnetic resonance imaging (fMRI) and electrocorticography (ECoG) has demonstrated that NLM's PPL is associated with predicting neural activation patterns during language comprehension tasks in the human brain (Schrimpf et al., 2021; Hosseini et al., 2024). This suggests a potential connection between the predictive capabilities of these models and understanding human information processing. In particular, one of the less well-understood phenomena in how neurodegeneration affects the human brain is the notion of cognitive and brain reserve. This notion is hypothesized to be responsible for findings that indicate individuals with higher innate abilities and/or aspects of life experience, such as educational and professional attainment, are able to mask the effects of dementia longer than those without these characteristics (Stern, 2002, 2009, 2012; Scarmeas and Stern, 2004, 2003; Snowdon et al., 1996). In some cases, the notion of cognitive and brain reserve may even allow individuals to revert from initial signs of cognitive impairment to normal function (Iraniparast et al., 2022).

Building upon these findings, our study seeks to further explore the potential of probing pre-trained GPT-2 family models (Radford et al., 2019) to simulate cognitive impairment observed in patients with dementia with a specific focus on the cognitive reserve hypothesis. Using a set of transcripts from a widely-used "Cookie Theft" picture description cognitive task, we propose that the impaired information processing as the disease progresses can be simulated by masking a certain share of attention heads in a pre-trained GPT-2 model. Specifically, we follow the previously established paired-perplexity paradigm (Li et al., 2022) using a pair of unmasked ("control") and masked ("dementia") NLMs. In this approach, the difference between PPLs produced by these two NLMs is used to discriminate between picture descriptions by patients with dementia and healthy controls. We hypothesize that larger GPT-2 models with more attention heads will exhibit greater resilience to masking (i.e., a proxy for neural degeneration), necessitating a larger share of attention heads to be masked to achieve comparable classification performance to smaller models. We evaluate this hypothesis by targeting two subsets of attention heads that are a) *most* important, and b) *least* important to representation of the content of the "Cookie Theft" task, in which the degree of importance is ranked by the gradient changes in each attention head during fine-tuning of a pre-trained GPT-2 model to the content of the "Cookie Theft" transcripts.

The contributions of this work can be summarized as follows: a) we provide preliminary evidence suggesting that the concept of cognitive reserve observed in human cognition appears to have an analog in ANNs; and b) our attention masking approach achieves comparable classification performance to another approach developed in prior work that directly artificially degrades NLM parameters (Li et al., 2022), and the state-of-the-art (SOTA) model trained from scratch (TaghiBeyglou and Rudzicz, 2024) with *significantly fewer* trainable parameter masking/fitting.[1]

2 Background

2.1 Cognitive Reserve

The notions of brain plasticity in the human brain and "graceful degradation" in ANNs have been

[1]The code to reproduce the results presented in this paper is available at GitHub. The data are also publicly available but cannot be redistributed and must be obtained directly from Dementia Bank.

extensively investigated in the neuroscientific literature demonstrating, for example, that a large proportion (over 80%) of the connections in an ANN trained to simulate the motor cortex to generate signals directing body movement have to be ablated before the model's performance begins to collapse (Lukashin et al., 1994; Lukashin and Georgopoulos, 1994). The concepts of cognitive and brain reserves are closely related to brain plasticity applied to observations in neurodegenerative diseases as illustrated in Figure 1. One of the earlier observations of this phenomenon comes from the Nun Study which found that low linguistic ability early in life (possibly due to innate abilities or educational attainment) is predictive of poor cognitive function and AD later in life (Snowdon et al., 1996). The concept of cognitive reserve was further developed based on observations of the individual differences in effects of brain damage or pathology on clinical manifestations of cognitive function (Stern, 2002, 2009, 2012). A multi-site study (Esiri et al., 2001) reported that up to 25% older adults without signs of cognitive impairment during neuropsychological testing meet all the histopathological criteria for AD (amyloid plaques and tau protein tangles) prior to their death. While this study did not assess brain volume, another similar study did find that a subgroup of 10 study participants who had both AD pathology and preserved mental status had greater brain weights and number of neurons in their brains (Katzman et al., 1988).

A distinction can be made between the closely related notions of cognitive reserve and brain reserve. Cognitive reserve refers to the efficiency of brain networks, which manifests as greater educational and professional attainment. Brain reserve, on the other hand, refers to the physical properties of the brain, such as a larger number of neurons in biological neural network(s). This can manifest, for example, as a higher intelligence quotient. These two notions are difficult to disentangle due to their significant interdependence (Steffener and Stern, 2012). The properties of these notions have also been described using passive or active models that correspond to the notions of brain and cognitive reserves, respectively. Passive models (Katzman, 1993; Satz, 1993) measure the cognitive reserve by the size of the brain or the count of neurons in the brain. Passive models hypothesize that there is a threshold for brain reserve capacity - once an individual passes the "point of no return", the manifestation of neurodegenerative disease, such as AD, will occur regardless. Contrary to passive models, active models (Stern, 2002) hypothesize that there is a neural compensatory effect for brain damage. This effect consists of the brain compensating for the damage by activating other biological neural network(s) to perform cognitive task-related activities. In this case, patients of similar brain impairment but with more cognitive reserve may be more resilient to the disease's progression before the clinical manifestations of neurodegeneration become apparent. Quantitatively, there is no clear difference between the passive and active models of cognitive reserve, as both of them rely on the physiologic basis of biological neural networks in the brain. This provides an opportunity to evaluate the underlying mechanisms that contribute to cognitive reserve across various neurological conditions computationally.

To avoid any potential confusion between these terms referring to different types of resilience and to avoid any inadvertent conflation between artificial and human brain networks, in the remainder of this paper we will refer to the phenomenon of resilience to damage that we observe in ANNs specifically as "artificial neural reserve" and use the terms "cognitive/brain reserve" to refer exclusively to human brain networks.

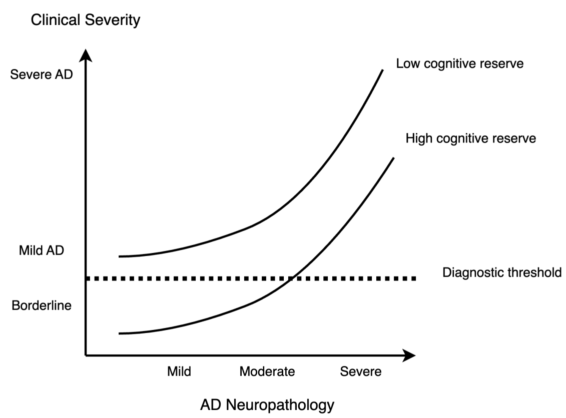


Figure 1: A theoretical illustration of cognitive reserve and its mediation effect between AD neuropathology (x-axis) and clinical outcome (y-axis). Illustration derived from Stern (2002, 2009). As the disease progresses (i.e., with more impairment), individuals with higher cognitive/brain reserve would be more resilient to the effects, resulting in a lower level of clinical severity.

2.2 Probing the Neural Network

The ablation of connections in ANNs is also referred to as probing in NLMs. This is a growing field aimed at understanding the inner workings of large-scale transformer-based NLMs by probing the mechanism (i.e., attention weights, hidden states) to better understand the linguistic structure and representations encoded by such models. Similarly to the early findings of Lukashin et al. (1994) and Lukashin and Georgopoulos (1994), more recent work on transformers (i.e., Michel et al. (2019); Prasanna et al. (2020); Sanh et al. (2020)) demonstrates that a large percentage of attention heads or sub-modules can be removed at inference time without significantly impacting performance.

2.3 Linguistic Anomalies in AD

AD is a neurodegerative disease, and progressively worsening linguistic deficits often accompany its progression (Kempler and Goral, 2008; Altmann and McClung, 2008). A widely-used diagnostic task to capture such linguistic anomalies is the "Cookie Theft" picture description task from the Boston Diagnostic Aphasia Examination (Goodglass and Kaplan, 1983). In this task, participants are asked to describe everything they see going on in Figure 2. Previous studies have demonstrated that dementia patients tend to overuse pronouns (Almor et al., 1999) and tend to perseverate (Hier et al., 1985) when describing the "Cookie Theft" picture.

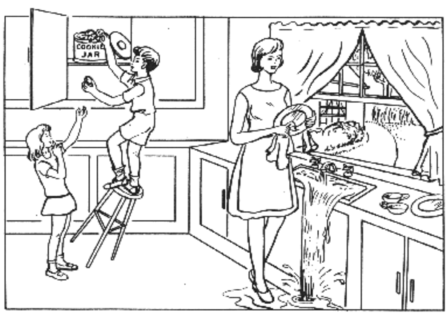


Figure 2: The "Cookie Theft" picture description stimuli.

There is a rich body of evidence that supervised machine learning and deep learning methods can learn to distinguish the subtle linguistic characteristics between healthy individuals and people with dementia. However, such models present a danger of overfitting, and hinder interpretability of model predictions, which are both critical concerns for clinical artificial intelligence (AI) applications (Graham et al., 2020). Alternatively, PPL is an easily interpretable measure used to evaluate model performance. With dementia, the difference of the *paired-perplexity* paradigm from a "healthy control" NLM and a "dementia" NLM provides a diagnostically useful summary value that distinguishes language samples produced by dementia patients (Fritsch et al., 2019; Cohen and Pakhomov, 2020). Prior work (Li et al., 2022) has shown that the difference of PPLs from a pre-trained GPT-2 paired with an artificially degraded version of itself approximates SOTA classification performance *without* requiring a data set from dementia patients of comparable size to its comprehensive training data. However, this approach requires evaluating thousands of masking patterns in order to investigate the effects of masking various combinations of attention heads exhaustively. In the current work we obviate this requirement for extensive experimentation by using targeted masking (guided by the changes in gradients during training) of two subsets of attention heads that are a) *most* "important", and *least* "important" with respect to the content of the "Cookie Theft" picture. We show that the resulting masked models can effectively identify transcripts from dementia patients with *significantly fewer* trainable parameters while exhibiting comparable classification performance to previous studies (Li et al., 2022; TaghiBeyglou and Rudzicz, 2024).

3 Methods

3.1 Data

We use two publicly available datasets that contain responses to the "Cookie Theft" picture description task: a) AD Recognition through Spontaneous Speech (ADReSS) Challenge[2] (Luz et al., 2020), and b) the Wisconsin Longitudinal Study (WLS)[3] (Herd et al., 2014). Table 1 shows basic characteristics of datasets used in this study. ADReSS is a subset of the Pitt corpus (Becker et al., 1994) designed to address the absence of a standardized train/test split in prior work. It is specifically matched on age and gender to reduce potential confounding effects. The WLS is a longitudinal study of 694 men and 675 women who graduated from Wisconsin high schools in 1957. The participants were interviewed up to 6 times between 1957 and 2011. The "Cookie Theft" picture description task was administered in the later round of interviews. In particular, we restricted the original WLS dataset to a total of 102 participants who a) agreed to partici-

[2] https://dementia.talkbank.org/ADReSS-2020/

[3] https://dementia.talkbank.org/access/English/WLS.html

pate in the "Cookie Theft" picture description task, and b) had either a clinical diagnosis of dementia or were deemed healthy in follow-up interviews conducted in 2020. This information was obtained through phone interviews and assessments by advanced practice providers. Subsequently, the collected data was presented to a panel of clinicians to obtain the diagnosis.

Dataset		Dementia	Healthy Controls
		# of participants (n)	
ADReSS	Train	54	54
	Test	24	24
	Total	78	78
WLS		29	73

Table 1: The characteristics of ADReSS and WLS.

We perform verbatim transcripts pre-processing using TRESTLE (**T**oolkit for **R**eproducible **E**xecution of **S**peech **T**ext and **L**anguage **E**xperiments) (Li et al., 2023) by removing utterances that do not belong to the participants, unintelligible words, and speech and non-speech artifacts event descriptions (i.e., "laughs", "clear throat").

3.2 Modeling and Evaluation

We follow a similar masking strategy to that proposed by Michel et al. (2019) to mask attention heads of the GPT-2 small, medium, large, and XL models via the rank of their importance to the task. We focus on the GPT-2 family models to minimize the variability that would result from multiple modeling architectures. The task-importance of attention heads in each model is determined by the gradient changes during the fine-tuning for subsequent word prediction task using transcripts of "Cookie Theft" picture descriptions in the training portion of the ADReSS dataset. Intuitively, if the gradient change of an attention head is large, this attention head is likely important with respect to predicting the language to which the model is being fine-tuned, and vice versa.

In contrast to the approach by Michel et al. (2019), which prunes the *least* important attention heads during testing, we anticipate that the *most* important attention heads are those relevant for predicting the text of the "Cookie Theft" task. This idea is supported by Yorkston and Beukelman (1980) and Berube et al. (2019), who found that the number of content units represented – a measure of how much relevant information is conveyed in the description – is sensitive to linguistic deficits often observed in individuals with neurodegenerative disease. However, we also reason that the *least* important attention heads may represent subtle differences in linguistic structure and representations that may distinguish between dementia patients and healthy controls. We also test the possibility that the semantic impairment observed in AD (Huff et al., 1986; Giffard et al., 2001; Hodges and Patterson, 1995) could be potentially simulated by masking a certain share of the columns in the pre-trained NLMs' token embedding matrix, where each column contributes to the representation of the meaning of each token in the model's vocabulary. Thus, masking columns in the embedding matrix leads to degrading the representation of *all* vocabulary items vs. degrading or deleting specific tokens from the otherwise intact vocabulary by operating on the rows of the embedding matrix.

Following these considerations, we design the masking strategies as follows: a) we fine-tune each of the GPT-2 models with a language model head layer as the top layer on the ADReSS training set to get the corresponding ranking of importance for each of the attention heads; b) we iteratively mask a small share ($n\%$) of ranked attention heads *bidirectionally*, which consists of the $\frac{n}{2}\%$ *most* important attention heads and the $\frac{n}{2}\%$ *least* important attention heads, then gradually increase the percentage of attention heads for masking, and c) we iteratively mask columns of the word embedding matrix in reverse order, moving from right to left, and gradually increase the percentage of word embedding columns for masking[4].

We examine the artificial neural reserve hypothesis using two evaluation approaches. The first approach consists of simply estimating the PPL of the progressively degraded NLMs based on healthy individuals' transcripts from an independent dataset containing the same type of picture descriptions as the dataset that was used to rank attention heads by their importance to the dementia classification task. We use the WLS dataset and select only those WLS participants that remained cognitively healthy over the entire study period as the independently collected dataset for log PPL estimation. Using this

[4]All experiments in this study are done with HuggingFace's transformers package (Wolf et al., 2020) on one A100 GPU.

approach, in addition to masking attention heads, we also experiment with masking model weights in the token embedding matrix to see if any observed effects are specific to the attention mechanism.

The second approach consists of evaluating the classification performance of ablated/degraded models paired with the original versions of the same GPT-2 model using the paired-perplexity paradigm (Fritsch et al., 2019; Cohen and Pakhomov, 2020; Li et al., 2022). These evaluations are conducted on the testing portion of the ADReSS dataset, with accuracy (ACC) and area under the receiver-operator characteristic (ROC) curve (AUC) as the evaluation metrics. Specifically, for the paired-perplexity paradigm, we estimate the ratio of PPLs $\frac{PPL_{control}}{PPL_{dementia}}$ of each transcript from the test set. The ACC measure is calculated as accuracy at the equal error rate (EER), where the false acceptance rate is equal to false rejection rate on the ROC curve. The intuition behind this approach is based on the expectation that successful masking of a portion of attention heads in a pre-trained NLM will result in the NLM exhibiting dementia-like behavior, which would in turn result in high AUC and ACC values of the paired-perplexity classification.

4 Results

4.1 Effects of Masking on Perplexity

As illustrated in Figure 3a, the predictive ability of smaller GPT-2 models degrades linearly with the degree of damage inflicted on the attention mechanism by masking progressively larger proportion of attention heads. The predictive ability of the larger GPT-2 models, on the other hand, degraded in a non-linear fashion where increases in log PPL were relatively flat up to 40-50% of the attention heads being masked and then began to increase exponentially. Fitting the GPT-2 small, medium, large and XL model log PPL to a linear regression line resulted in r^2 goodness-of-fit values of 0.99, 0.89, 0.91 and 0.83, respectively, whereas fitting to an exponential regression line failed to converge for the small and medium models and yielded r^2 values of 0.97 and 0.99 for the large and XL models, respectively. The results of Dunn's test further confirmed our observations, showing that the differences between log PPLs estimated by GPT-2 small and GPT-2 XL (adjusted p-value $<$ 0.01), and GPT-2 medium and GPT-2 XL (adjusted p-value $<$ 0.05) when masking attention heads are statistically significant. In contrast, *all* combinations of log PPLs were not significantly different from each other for all GPT-2 models when masking the word embedding matrix (adjusted p-value $>$ 0.05).

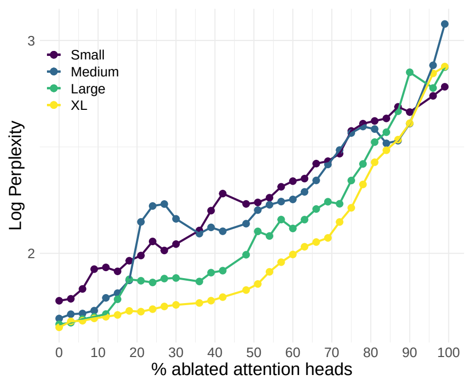
Compared to masking attention heads, with GPT-2 small, medium, large and XL model we needed to mask 93% (714 out of 768), 66% (675 out of 1024), 87% (1113 out of 1280), and 66% (1050 out of 1600) columns in the word embedding matrix to achieve ACCs of 0.75, 0.85, 0.79, and 0.81 respectively, on the ADReSS test set. Figure 3b can further support this claim, as estimated log PPLs of masking word embedding matrix show no significant statistical differences across various GPT-2 models.

4.2 Effects of Masking on Dementia Classification

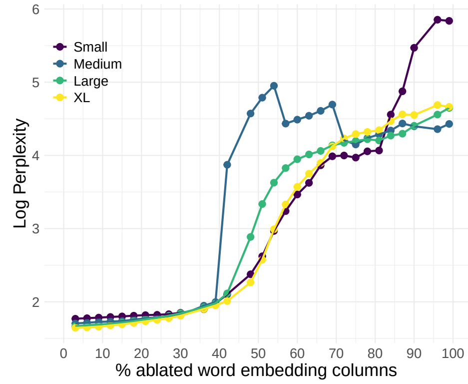
As shown in Table 2, impairing 9% of attention heads (n=12) of the GPT-2 small model (the "dementia" model) achieved an ACC of 0.83 and AUC of 0.86 when paired with the original unmasked version of itself (the "control" model) on the ADReSS test set. This is comparable to the prior work (Li et al., 2022) (ACC = 0.85, AUC = 0.89) but the masking approach uses *significantly fewer* masked parameters. Our results also show that a larger share of attention heads in the larger models must be masked to approximate a "dementia" model with the same level of performance in the paired-perplexity classification than with smaller models. Notably, masking of the word embedding matrix did not result in comparable observations. As anticipated by the results shown in Figure 3b, with GPT-2 models we needed to mask a majority portion (e.g., $>$ 50%) of the word embedding matrix to obtain similar level of classification performance regardless of model size.[5]

As illustrated in Figure 4, we observed that once the best-performing masking pattern, marked by the highest ACC, was reached, the classification performance of all GPT-2 models started to fluctuate. However, this observation did not occur with the word embedding matrix masking. As illustrated in Figure 5 in Appendix A, the classification performance exhibited fluctuations prior to the emergence of the best-performing masking pattern, indicating that masking the columns of the word embedding matrix has less impact on identifying the signs of

[5]The importance of attention heads for each model can be found in Table 3, Table 4, Table 5, and Table 6 in the Appendix A.



(a) GPT-2s with masked attention heads.



(b) GPT-2s with masked word embedding matrix.

Figure 3: Changes in model log PPL as a function of the proportion of masked attention heads across GPT-2 models of various sizes. Note: the curves in panel (a) show that GPT-2 XL model has the most non-linear/concave shape indicating that the model starts to degrade rapidly only after masking of about 50% of its attention heads, followed by the curve for the GPT-2 large model. The smaller GPT-2 models begin to degrade with proportionally less masking, and exhibit a monotonic relationship between the magnitude of attention heads masking and model performance. The curves in panel (b) show almost completely preserved model performance without differences between models up to the point at which 40% - 50% of the columns in their embedding matrices have been masked. After that point, the performance of all models collapses "catastrophically"

Model	**GPT-2 small**	**GPT-2 medium**	**GPT-2 large**	**GPT-2 XL**
# of parameters	124M	355M	774M	1.5B
# of masked attention heads	12	92	388	1080
% of masked attention heads	9	24	54	90
ACC	0.83	0.83	0.81	0.81
AUC	0.86	0.85	0.80	0.82

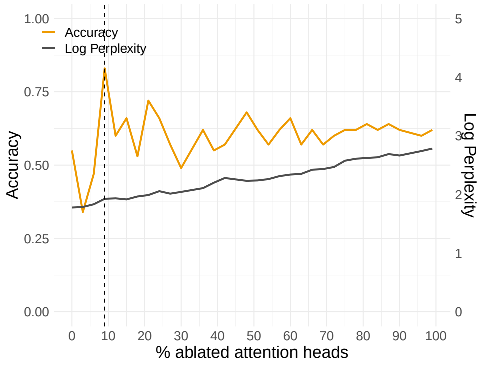
Table 2: Classification performance of the paired-perplexity approach based on pre-trained and masked GPT-2 models on the ADReSS test set.

cognitive impairment from text as it probably does not result in a good dementia-like model for the paired-perplexity classification task.

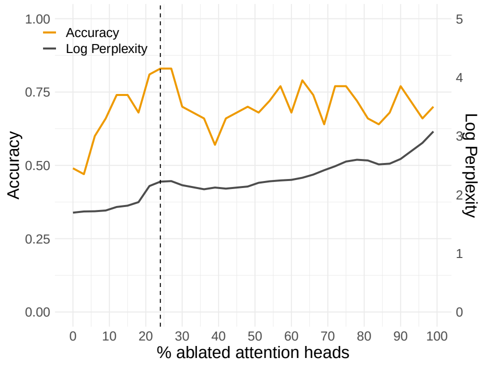
5 Discussion

The results of experiments presented in this paper suggest that the notion of cognitive reserve in the brain may have an analogue in transformer-based ANNs that is localized to the attention mechanism. Recent neuroscientific evidence shows that NLMs' PPL is predictive of human behavioral responses and and neural responses in functional MRI studies (Schrimpf et al., 2021; Hosseini et al., 2024). Based on this evidence, we interpret our findings of the differences in log PPL changes as a result of masking attention heads in NLMs of variable size as at least suggestive that the resilience to damage is non-linear to the number of attention heads in NLMs. In other words, it takes disproportionately more masking to damage larger NLMs to elicit the same level of degradation in performance, as compared with smaller NLMs. Furthermore, the dissociation in performance as a result of damaging attention heads vs. the token embedding weights suggests that the NLM's artificial neural reserve effects are localized to the attention mechanism.

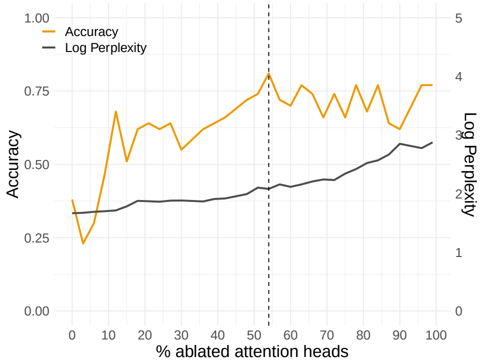
Our results also suggest that masking attention heads within the paired-perplexity paradigm using the ratio of unmasked ("control") and masked ("dementia") pre-trained GPT-2 models results in good classification performance *without* requiring a corresponding large dataset produced by demen-



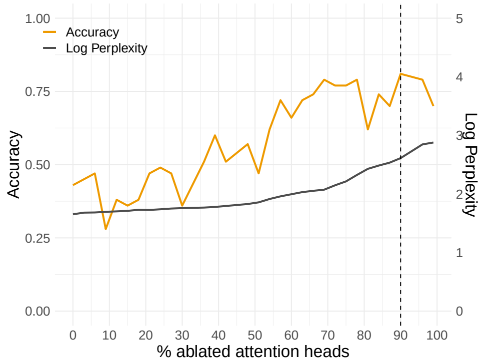
(a) GPT-2 small



(b) GPT-2 medium



(c) GPT-2 large



(d) GPT-2 XL

Figure 4: Comparison of GPT-2 models with masked attention heads on paired-perplexity classification performance. The left y-axis denotes classification performance using both masked and unmasked GPT-2 models on the ADReSS test set. The right y-axis indicates log PPL estimated from transcripts of WLS healthy individuals. The x-axis represents the percentage of attention heads getting masked. The vertical dashed line indicates the best-performing masking pattern, achieving the highest ACC.

tia patients and extensive parameter tuning. This can be achieved with as little as masking only 9% of attention heads of a pre-trained GPT-2 small model.

In contrast to previous studies, which typically involved purging attention heads determined to be the *least* important, our bidirectional masking method adds supporting evidence of content units (Yorkston and Beukelman, 1980; Berube et al., 2019), suggesting the importance of these contextual features in addition to the predominant emphasis on linguistic structure and representation modeling in previous research (e.g., Orimaye et al. (2014), Fraser et al. (2016)). The results of bidirectional masking also offers an interpretable explanation for transfer learning's remarkable performance using pre-trained NLMs. It suggests that during fine-tuning, pre-trained NLMs use a combination of both task-specific (the *most* important) and task-agnostic (the *least* important) heads to achieve remarkable performance on various downstream tasks. Those task-agnostic attention heads may play an important role in transfer learning. This also may explain why distilled NLMs in which the "nonvolitional cues" that fall outside of common NLP benchmarks are purged during the distillation, generalize less-than-ideally to other types of data produced by individuals with dementia (Li et al., 2022). With larger models, there are considerably more attention heads that can serve as those "nonvolitional cues," helping a larger NLM perform better (Agbavor and Liang, 2022).

As the columns of the token embedding matrix in a pre-trained NLMs represent the global semantics of tokens in the vocabulary, the observations that the best-performing masking pattern appears in the later stage of the token embedding matrix masking are consistent with previously published findings that semantic impairment often occurs in the later stage (i.e., moderate) of the disease (Huff et al., 1986; Giffard et al., 2001; Hodges and Patterson, 1995). As illustrated in Figure 5, when masking the later 66% columns (675 out of 1024) of the word embedding matrix, the paired unmasked and masked GPT-2 medium achieves an ACC of 0.85 on the ADReSS test set. This finding is consistent with a previous work (Hewitt and Manning, 2019), suggesting that some syntactic information is embedded implicitly in the word embedding matrix. This also provide an empirical support of our findings that masking word embedding matrix of a pre-trained NLM can provide some degree of discriminating effect on this downstream task. However, masking the word embedding matrix is far less effective than masking attention heads to simulate dementia-related cognitive impairment.

Our results suggest that similar mechanisms of resilience may exist in both human cognition and computational models. This could lead to more nuanced strategies in response to develop early screening tools for the delayed onset of cognitive impairment in the population with high risk. Our study holds promise for deploying such early-screening method in resource-constrained clinical settings to improve early intervention and patient management for AD.

6 Conclusion

We presented experimental findings suggesting the presence of artificial neural reserve in transformer-based NLMs, analogous to the concepts of brain/cognitive reserve in studies of human cognition. In addition, we introduced a novel bidirectional attention head ablation method that enables using unmasked and masked GPT-2 models in the paired-perplexity paradigm for detecting linguistic anomalies with significantly less parameter masking or fitting.

Acknowledgement

Research reported in this publication was supported by the National Library of Medicine of the National Institutes of Health under award number R01 LM014056-02S1, and the National Institute on Aging of the National institutes of Health under award number AG069792.

Limitations

The work presented here has several limitations. First, the size of datasets used in this study is relatively small compared to datasets typically analyzed in the open-domain NLP tasks, therefore the results may not be readily generalizable. Second, all datasets used in our study are in American English, and many participants of these two studies are representative of White, non-Hispanic American men and women located at the north part of the United States, which certainly limit their applicability to other languages. Third, while we propose that the findings presented in this paper may be interpreted as an analogue of the notions of cognitive or brain reserve, we do not suggest that GPT-2 models are accurate models of the human brain. Rather, our interpretation of these findings is that experimenting with masking of attention heads in models of various sizes and architectures may be useful in helping us understand cognitive processes that take place in the human brain. The observed effects of attention masking on the model's performance and behavior, while suggestive of an analog to cognitive reserve in the human brain, should not establish a direct causal link to human cognitive processes. Additionally, the ranking of attention heads by their relative importance is specific to the ADReSS dataset as it was derived in the training portion of the dataset and may not readily generalize to other datasets and types of data. Lastly, in this paper we did not address the distinction between the notions of cognitive and brain reserves. It would be important to investigate in future work if NLMs of the same size and architecture but different quantities and quality of the training data (i.e., as a simulation of educational attainment) exhibit differential resilience to damage independently of the effects observed in models of variable size.

References

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

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.

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.

Shauna Berube, Jodi Nonnemacher, Cornelia Demsky, Shenly Glenn, Sadhvi Saxena, Amy Wright, Donna C Tippett, and Argye E Hillis. 2019. Stealing cookies in the twenty-first century: Measures of spoken narrative in healthy versus speakers with aphasia. *American journal of speech-language pathology*, 28(1S):321–329.

Andrea Bradford, Mark E Kunik, Paul Schulz, Susan P Williams, and Hardeep Singh. 2009. Missed and delayed diagnosis of dementia in primary care: prevalence and contributing factors. *Alzheimer disease and associated disorders*, 23(4):306.

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. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1946–1957, Online. Association for Computational Linguistics.

MM Esiri, F Matthews, C Brayne, PG Ince, FE Matthews, JH Xuereb, JC Broome, J McKenzie, M Rossi, IG McKeith, et al. 2001. Pathological correlates of late-onset dementia in a multicentre, community-based population in england and wales. *Lancet*.

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

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.

Bénédicte Giffard, Béatrice Desgranges, Florence Nore-Mary, Catherine Lalevée, Vincent de la Sayette, Florence Pasquier, and Francis Eustache. 2001. The nature of semantic memory deficits in Alzheimer's disease: New insights from hyperpriming effects. *Brain*, 124(8):1522–1532.

Elaine Giles, Karalyn Patterson, and John R Hodges. 1996. Performance on the boston cookie theft picture description task in patients with early dementia of the alzheimer's type: Missing information. *Aphasiology*, 10(4):395–408.

Harold Goodglass and Edith Kaplan. 1983. *Boston diagnostic aphasia examination booklet*. Lea & Febiger.

Sarah A Graham, Ellen E Lee, Dilip V Jeste, Ryan Van Patten, Elizabeth W Twamley, Camille Nebeker, Yasunori Yamada, Ho-Cheol Kim, and Colin A Depp. 2020. Artificial intelligence approaches to predicting and detecting cognitive decline in older adults: A conceptual review. *Psychiatry research*, 284:112732.

Pamela Herd, Deborah Carr, and Carol Roan. 2014. Cohort profile: Wisconsin longitudinal study (wls). *International journal of epidemiology*, 43(1):34–41.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

John R. Hodges and Karalyn Patterson. 1995. Is semantic memory consistently impaired early in the course of alzheimer's disease? neuroanatomical and diagnostic implications. *Neuropsychologia*, 33(4):441–459.

Eghbal A Hosseini, Martin Schrimpf, Yian Zhang, Samuel R Bowman, Noga Zaslavsky, and Evelina Fedorenko. 2024. Artificial neural network language models predict human brain responses to language even after a developmentally realistic amount of training. *Neurobiology of Language*, pages 1–50.

F.Jacob Huff, Suzanne Corkin, and John H. Growdon. 1986. Semantic impairment and anomia in alzheimer's disease. *Brain and Language*, 28(2):235–249.

Maryam Iraniparast, Yidan Shi, Ying Wu, Leilei Zeng, Colleen J Maxwell, Richard J Kryscio, Philip D St John, Karen S SantaCruz, and Suzanne L Tyas. 2022. Cognitive reserve and mild cognitive impairment: predictors and rates of reversion to intact cognition vs progression to dementia. *Neurology*, 98(11):e1114–e1123.

Pentti Kanerva. 2009. Hyperdimensional computing: An introduction to computing in distributed representation with high-dimensional random vectors. *Cognitive computation*, 1:139–159.

Robert Katzman. 1993. Education and the prevalence of dementia and alzheimer's disease. *Neurology*, 43(1_part_1):13–13.

Robert Katzman, Robert Terry, Richard DeTeresa, Theodore Brown, Peter Davies, Paula Fuld, Xiong Renbing, and Arthur Peck. 1988. Clinical, pathological, and neurochemical changes in dementia: a subgroup with preserved mental status and numerous neocortical plaques. *Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society*, 23(2):138–144.

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

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. In

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1866–1877, Dublin, Ireland. Association for Computational Linguistics.

Changye Li, Weizhe Xu, Trevor Cohen, Martin Michalowski, and Serguei Pakhomov. 2023. Trestle: Toolkit for reproducible execution of speech, text and language experiments. *AMIA Summits on Translational Science Proceedings*, 2023:360.

Alexander V Lukashin and Apostolos P Georgopoulos. 1994. A neural network for coding of trajectories by time series of neuronal population vectors. *Neural Computation*, 6(1):19–28.

Alexander V Lukashin, George L Wilcox, and Apostolos P Georgopoulos. 1994. Overlapping neural networks for multiple motor engrams. *Proceedings of the National Academy of Sciences*, 91(18):8651–8654.

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 *Proc. Interspeech 2020*, pages 2172–2176.

Paul Michel, Omer Levy, and Graham Neubig. 2019. Are sixteen heads really better than one? In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Sylvester Olubolu Orimaye, Jojo Sze-Meng Wong, and Karen Jennifer Golden. 2014. Learning predictive linguistic features for Alzheimer's disease and related dementias using verbal utterances. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 78–87, Baltimore, Maryland, USA. Association for Computational Linguistics.

Sai Prasanna, Anna Rogers, and Anna Rumshisky. 2020. When BERT Plays the Lottery, All Tickets Are Winning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3208–3229, Online. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Victor Sanh, Thomas Wolf, and Alexander Rush. 2020. Movement pruning: Adaptive sparsity by fine-tuning. In *Advances in Neural Information Processing Systems*, volume 33, pages 20378–20389. Curran Associates, Inc.

Paul Satz. 1993. Brain reserve capacity on symptom onset after brain injury: a formulation and review of evidence for threshold theory. *Neuropsychology*, 7(3):273.

Nikolaos Scarmeas and Yaakov Stern. 2003. Cognitive reserve and lifestyle. *Journal of clinical and experimental neuropsychology*, 25(5):625–633.

Nikolaos Scarmeas and Yaakov Stern. 2004. Cognitive reserve: implications for diagnosis and prevention of alzheimer's disease. *Current neurology and neuroscience reports*, 4:374–380.

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.

Mengke Shi, Gary Cheung, and Seyed Reza Shahamiri. 2023. Speech and language processing with deep learning for dementia diagnosis: A systematic review. *Psychiatry Research*, 329:115538.

David A Snowdon, Susan J Kemper, James A Mortimer, Lydia H Greiner, David R Wekstein, and William R Markesbery. 1996. Linguistic ability in early life and cognitive function and alzheimer's disease in late life: Findings from the nun study. *Jama*, 275(7):528–532.

Jason Steffener and Yaakov Stern. 2012. Exploring the neural basis of cognitive reserve in aging. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*, 1822(3):467–473.

Yaakov Stern. 2002. What is cognitive reserve? theory and research application of the reserve concept. *Journal of the international neuropsychological society*, 8(3):448–460.

Yaakov Stern. 2009. Cognitive reserve. *Neuropsychologia*, 47(10):2015–2028.

Yaakov Stern. 2012. Cognitive reserve in ageing and alzheimer's disease. *The Lancet Neurology*, 11(11):1006–1012.

Laura Stokes, Helen Combes, and Graham Stokes. 2015. The dementia diagnosis: a literature review of information, understanding, and attributions. *Psychogeriatrics*, 15(3):218–225.

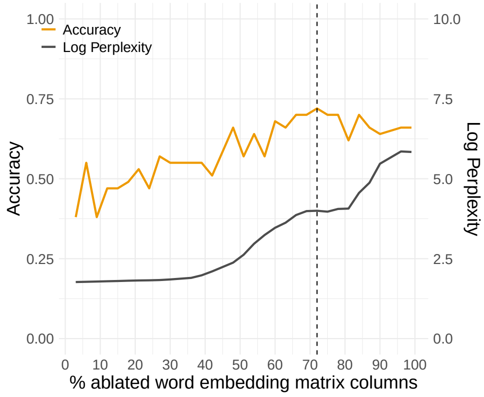
Behrad TaghiBeyglou and Frank Rudzicz. 2024. Context is not key: Detecting alzheimer's disease with both classical and transformer-based neural language models. *Natural Language Processing Journal*, 6:100046.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

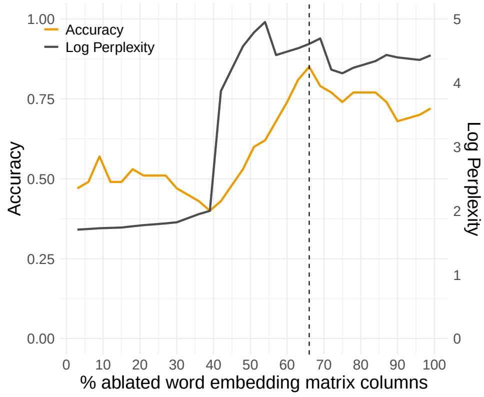
Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander 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.

Kathryn M. Yorkston and David R. Beukelman. 1980. An analysis of connected speech samples of aphasic and normal speakers. *Journal of Speech and Hearing Disorders*, 45(1):27–36.

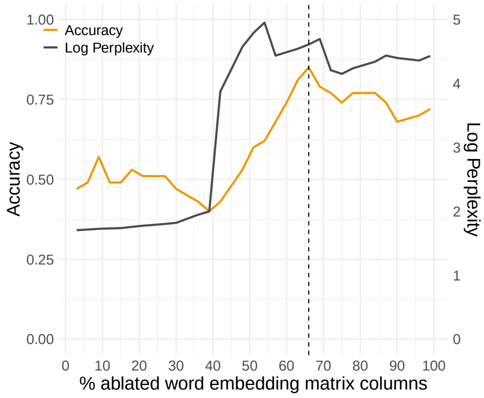
A Appendix



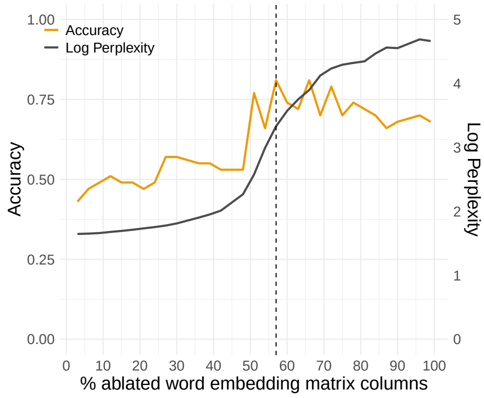
(a) GPT-2 small



(b) GPT-2 medium



(c) GPT-2 large



(d) GPT-2 XL

Figure 5: Comparison of GPT-2 models with masked columns of word embedding matrix on classification performance and cognitive reserve manifestation. The left y-axis denotes classification performance using both masked and unmasked GPT-2 models on the ADReSS test set. The right y-axis indicates log PPL estimated from transcripts of WLS healthy individuals. The x-axis represents the percentage of attention heads getting masked. The vertical dashed line indicates the best-performing masking pattern, achieving the highest ACC.

0	1	2	3	4	5	6	7	8	9	10	11
121	5	7	65	19	46	58	56	9	1	0	60
16	34	94	71	13	90	42	4	113	86	47	14
74	106	107	32	89	18	17	87	75	11	8	128
48	15	82	78	93	68	129	79	77	72	54	23
40	67	22	115	36	131	100	83	140	61	141	135
41	29	134	137	119	12	92	21	31	3	69	28
76	95	101	125	91	130	37	99	80	98	10	122
117	45	124	81	116	49	25	26	62	97	143	136
64	123	30	43	88	38	27	55	73	114	118	142
111	53	102	70	50	57	105	84	120	138	139	20
132	110	66	103	44	52	126	108	109	59	96	85
6	24	127	39	33	133	104	51	2	112	63	35

Table 3: The rank of importance for each attention head in the GPT-2 small model. The rows represent the layer of attention blocks in the model whereas the columns represent attention heads per layer.

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
139	115	36	376	354	132	355	357	109	70	148	84	217	156	309	363
77	5	28	126	41	95	266	57	286	316	61	258	59	13	138	202
268	131	21	209	367	46	86	228	45	29	22	63	48	76	155	18
170	222	346	235	38	12	257	301	172	112	193	121	188	26	238	103
294	47	198	102	315	64	291	40	246	375	66	151	292	293	264	165
199	272	250	364	285	226	192	343	17	114	119	260	237	166	62	68
50	16	360	185	20	368	359	130	350	72	8	289	267	251	310	7
279	88	216	79	141	256	85	83	101	82	204	232	243	142	107	55
53	284	195	129	253	133	97	154	137	262	281	60	203	177	31	248
58	261	366	25	135	227	320	333	197	94	242	125	65	15	273	117
255	271	160	54	49	269	303	140	176	296	167	311	110	92	290	239
143	241	158	325	74	299	56	342	287	225	214	6	372	98	150	100
183	182	52	370	19	11	186	113	240	353	184	34	312	297	179	259
32	210	358	212	331	67	371	230	330	116	304	263	159	278	163	149
87	324	208	334	356	220	319	162	136	236	69	108	327	190	337	27
383	14	4	106	300	275	96	71	207	75	352	10	221	178	307	180
339	174	336	340	44	191	99	317	39	244	361	249	274	73	206	377
362	341	201	345	231	219	35	111	205	80	93	23	347	168	229	30
378	146	171	145	3	181	321	196	124	164	152	120	224	276	382	24
33	302	128	78	252	298	189	144	344	42	105	349	329	288	104	283
369	37	90	247	295	305	81	314	318	215	173	123	313	254	365	306
277	43	381	118	373	280	233	380	234	270	374	211	335	282	323	332
157	322	2	51	326	161	147	200	218	338	348	213	379	89	153	122
187	194	265	175	134	1	0	351	169	127	328	91	308	223	245	9

Table 4: The rank of importance for each attention head in the GPT-2 medium model. The rows represent the layer of attention blocks in the model whereas the columns represent attention heads per layer.

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
39	79	318	53	587	121	273	390	84	125	433	510	653	548	379	255	193	132	555	622
567	285	553	403	481	670	375	398	399	354	298	424	502	624	665	544	709	473	620	268
219	181	585	683	621	463	326	506	434	615	339	651	100	312	497	332	412	428	14	634
12	22	633	391	436	658	94	563	614	386	355	104	174	40	90	172	78	407	328	673
149	302	105	322	474	264	469	184	35	319	177	185	295	203	296	438	666	489	543	87
606	323	209	195	579	204	51	584	267	569	590	662	395	55	93	290	552	159	508	186
163	425	225	25	448	107	531	325	58	27	503	612	92	559	287	118	352	275	251	546
681	592	523	77	494	56	549	342	368	611	528	320	558	47	308	575	248	383	671	314
134	250	340	272	712	396	171	137	610	430	238	269	26	604	211	630	261	133	532	111
34	240	388	408	337	44	568	247	359	411	672	215	258	198	410	547	472	525	581	702
617	500	566	468	365	311	692	533	422	699	660	645	265	131	643	526	657	36	189	632
116	146	81	413	625	284	161	62	460	588	627	545	330	88	293	527	650	71	303	245
372	299	357	524	640	685	11	31	674	346	565	599	194	564	439	145	196	97	260	167
688	175	224	263	41	239	114	294	331	609	202	249	373	120	550	119	454	246	98	220
173	310	179	381	377	103	479	135	516	475	148	205	401	443	169	560	218	556	654	600
637	432	301	394	65	156	123	329	66	70	214	574	2	10	307	153	46	112	613	237
166	542	50	155	217	117	435	305	417	43	138	102	367	327	253	570	414	343	143	207
243	647	7	191	698	109	141	423	598	126	182	0	421	17	24	52	110	80	59	91
234	488	351	256	649	113	551	668	164	400	501	128	83	282	210	317	45	511	222	68
513	274	199	594	519	348	140	168	151	157	664	188	244	449	144	347	515	324	522	619
23	122	82	656	447	358	642	1	16	42	589	646	233	466	304	361	37	32	96	30
28	216	397	364	562	29	101	356	471	695	270	276	162	283	170	306	277	5	573	486
300	493	74	517	678	402	130	576	165	459	418	4	426	442	583	190	6	291	208	38
221	69	370	85	60	178	641	491	577	682	154	106	703	392	362	54	603	99	315	371
13	349	415	313	124	427	281	338	580	420	416	380	499	644	20	384	530	183	192	72
297	286	440	648	75	201	57	561	136	152	33	242	95	108	369	697	716	229	266	534
206	477	687	538	707	127	706	280	64	409	669	19	496	200	376	490	150	498	514	616
482	158	487	444	9	160	197	231	406	76	21	187	180	437	693	257	139	467	321	504
223	462	485	419	572	591	142	288	363	350	719	536	704	718	677	334	675	680	445	623
686	405	701	539	8	596	278	582	230	456	341	529	710	241	344	86	458	652	635	541
63	89	366	382	476	232	638	271	453	235	227	404	262	73	67	717	374	228	512	212
605	601	15	484	389	3	176	483	509	667	495	345	393	289	309	259	115	387	446	129
48	465	557	385	18	535	628	705	554	602	478	492	316	360	236	521	254	684	252	586
689	571	464	607	691	292	450	696	49	520	333	626	505	713	700	631	452	593	694	335
636	578	147	61	714	279	455	708	378	655	595	480	461	679	518	597	451	659	661	507
608	540	226	639	431	336	676	213	470	715	618	441	711	690	353	457	629	537	429	663

Table 5: The rank of importance for each attention head in the GPT-2 large model. The rows represent the layer of attention blocks in the model whereas the columns represent attention heads per layer.

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
131	329	685	483	1,163	980	687	210	590	35	636	1,138	84	203	800	611	993	710	1,135	822	158	828	122	869	1,052
1,134	580	98	849	915	900	468	564	1,061	141	367	529	725	489	961	352	823	260	1,027	841	522	547	209	521	226
1,086	699	850	735	268	840	647	1,019	985	1,068	757	138	1,028	690	1,033	762	708	499	811	635	193	256	441	720	820
294	414	753	261	743	30	244	1,032	598	388	691	1,063	62	574	628	325	1,037	548	785	538	599	763	803	380	798
867	799	524	304	223	858	85	814	937	911	646	838	614	106	61	66	133	172	836	173	427	262	144	1,085	749
931	342	276	1,056	1,048	686	531	101	904	116	1,073	923	295	550	880	376	473	1,148	429	501	410	422	313	587	895
802	862	257	853	1,067	1,174	280	661	563	208	1,003	946	950	26	702	629	813	440	252	361	975	493	660	1,112	790
568	818	609	321	772	1,065	196	1,087	847	1,092	1,004	945	165	758	246	949	553	1,108	387	963	582	285	572	876	514
341	891	123	509	195	269	697	404	534	666	825	731	1,115	1,079	199	463	701	674	546	642	667	936	736	1,069	781
567	1,013	274	600	353	864	496	637	729	922	604	952	976	1,066	474	754	791	874	999	318	243	1,055	909	832	328
556	623	1,006	767	1,076	594	998	833	616	643	222	1,075	776	227	1,043	535	877	902	349	752	533	494	436	815	301
1,024	296	288	603	898	613	872	1,090	334	482	1,040	419	537	1,094	464	1,042	982	1,050	719	507	236	644	459	395	504
1,143	467	391	694	1,136	848	924	656	372	442	300	444	1,016	715	184	491	997	392	751	1,147	801	174	511	704	330
323	851	189	204	322	968	78	525	714	27	544	416	439	1,054	552	727	1,078	559	804	761	506	1,118	215	390	358
816	1,038	118	162	38	381	485	201	309	368	933	95	284	183	760	396	889	755	82	649	111	282	446	455	221
1,164	335	806	454	912	247	943	59	449	756	526	48	402	1,121	919	212	412	837	684	277	497	156	819	831	437
103	782	86	3	586	665	827	595	765	695	431	766	187	100	348	167	344	709	645	1,074	251	71	640	479	558
224	592	216	1,002	161	658	308	887	555	420	317	258	995	528	339	46	153	302	1,011	265	255	272	241	99	75
530	676	428	516	808	305	712	64	705	155	225	393	76	596	287	397	69	92	137	220	634	266	360	430	248
927	659	333	458	857	0	235	242	978	424	117	389	185	120	14	351	44	878	996	152	830	487	654	905	128
942	168	477	232	698	855	602	178	495	386	327	177	51	2	1,045	331	88	519	403	413	81	132	608	369	470
1,035	575	110	213	157	39	707	1,039	139	1,126	466	411	211	121	638	316	45	159	1,072	515	1,071	633	175	539	49
956	401	96	597	721	965	663	160	217	36	768	143	472	194	320	56	170	297	25	113	696	984	566	23	681
283	462	53	40	517	127	626	125	12	291	83	1	1,020	55	33	443	310	286	500	958	150	65	197	108	839
778	682	10	332	693	114	180	679	346	383	169	90	289	237	706	182	503	1,009	347	480	861	319	1,137	377	119
783	409	171	202	452	16	130	856	164	303	565	866	253	4	475	1,144	250	451	773	711	19	1,053	1,021	746	1,093
6	275	967	1,156	908	15	29	188	780	415	1,060	206	24	846	47	668	200	536	281	577	860	615	1,064	883	438
104	703	270	399	94	179	234	1,057	578	759	77	445	627	371	884	733	50	518	264	1,104	1,034	523	826	1,123	80
9	621	142	146	354	312	363	238	259	1,023	935	450	1,184	239	407	1,102	73	357	741	425	207	307	769	89	421
292	576	22	109	688	405	540	214	988	983	770	498	293	896	810	794	271	63	1,081	28	398	744	502	70	914
240	879	508	1,036	151	1,103	364	58	433	492	112	434	8	873	350	1,195	871	273	1,180	417	52	1,141	617	1,175	620
1,167	954	1,110	510	135	1,172	105	724	870	31	892	298	824	692	340	657	554	994	486	315	562	717	115	974	448
1,015	934	91	549	1,159	581	453	356	835	418	732	513	888	102	5	488	490	906	859	1,140	97	447	570	43	821
1,031	481	1,130	456	1,193	149	32	683	734	166	1,162	205	573	370	355	233	951	1,139	817	662	147	561	700	219	972
903	1,129	136	986	560	72	1,154	716	957	41	672	625	190	973	145	1,089	894	605	1,181	916	1,082	218	469	484	42
1,128	932	579	362	385	249	793	1,083	962	723	788	435	632	1,150	601	1,047	1,091	229	34	738	326	191	730	532	588
465	1,005	1,149	722	882	829	1,099	1,151	11	359	20	792	797	1,131	737	726	148	1,098	718	198	365	87	713	299	675
1,022	471	461	163	610	655	337	1,194	1,077	991	1,044	1,179	57	1,012	426	777	1,018	545	1,170	750	1,176	126	886	901	343
669	408	140	928	378	779	254	899	947	314	382	520	324	1,111	955	1,095	622	21	338	893	278	689	311	925	107
673	263	1,029	748	1,158	60	940	1,192	742	843	1,058	345	457	966	868	854	1,025	366	245	591	885	787	929	740	680
1,008	186	1,169	154	1,124	1,062	541	960	379	181	653	897	1,197	1,000	1,046	1,096	786	1,101	807	306	845	569	67	1,119	959
1,155	54	969	18	970	863	964	989	93	739	13	842	375	423	651	917	374	1,105	1,030	907	987	812	1,173	795	583
17	589	279	1,185	938	134	1,145	944	593	1,117	747	571	875	913	267	948	37	68	630	400	129	585	124	1,041	1,166
910	1,153	1,182	920	1,160	671	1,199	618	1,109	639	74	1,188	527	881	1,049	612	771	953	1,088	1,196	1,168	1,152	290	926	764
1,165	1,100	512	844	1,059	784	336	1,107	1,120	852	992	641	1,127	670	981	1,178	1,001	373	921	230	918	939	1,171	1,183	1,187
789	1,014	1,198	228	631	551	930	971	460	432	1,122	1,070	406	1,142	774	619	805	557	728	607	678	1,113	796	1,191	1,017
1,026	584	1,007	648	941	476	1,133	650	1,186	624	1,190	1,132	1,114	775	384	1,051	1,189	1,125	1,106	1,010	1,097	1,161	1,080	990	890
79	1,157	652	231	1,177	176	542	192	1,116	1,146	505	478	745	977	394	677	664	865	979	543	606	1,084	7	809	834

Table 6: The rank of importance for each attention head in the GPT-2 XL model. The rows represent the layer of attention blocks in the model whereas the columns represent attention heads per layer.