Debiasing Transformer Models through Weight Masking: Addressing Gender Confounding Shift in Dementia Detection

Anonymous ACL submission

Abstract

Deep language models are often described as "black-box" systems due to their opaque inference procedures. This presents a challenge in understanding the information they capture, and how it is encoded within transformer networks, raising the possibility that encoded biases may remain undetected. This work addresses confounding bias learned during model fine-tuning, when a pretrained language model is adapted to downstream domains and tasks. Building on previous methodologies, we extend them by proposing the Extended Confounding Filter and the Dual Filter. These methods aim to isolate and address weights within the transformer network that are associated with confounding variables through distinct training phases. We evaluate these methods on the *DementiaBank* dataset, a first-person narrative dataset that contains language of patients with cognitive impairment and healthy controls. We aim to demonstrate the applicability of the proposed methods in the domain of dementia detection as a means to correct for gender-related disparities in class distribution at training time. Our results show that transformer models can overfit to the subpopulation distribution in the training data. By disrupting the weights associated with known confounders, we show that fairer models can be achieved with reduced prediction bias towards specific subgroups. Moreover, our findings highlight resilience of the model against weights deletion and show a trade-off between model performance in dementia detection and the reduction of disparities across gender groups.[1]

1 Introduction

Transformer-based models (Vaswani et al., 2017) have achieved significant success across various language and vision tasks, leading to numerous applications. In particular, bidirectional encoder models based on the self-attention mechanism, such as BERT (Devlin et al., 2019) and its variants (Liu et al., 2019; Sanh et al., 2020; Lee et al., 2020; Qian et al., 2022), have demonstrated impressive performance gains on NLP benchmarks and domain-specific tasks due to their ability to learn rich dense representations from text. As the popularity of these models increases, it is important to ensure that their outputs are not biased towards (or against) certain groups at the point of deployment. However, in practice, most transformer models are optimized for and evaluated on a task of interest, without considering biases inherent in the data that may be embedded into the model (Baldini et al., 2022; Bolukbasi et al., 2016; Hutchinson et al., 2020; Webster et al., 2021; de Vassimon Manela et al., 2021). If left unaddressed, such biases can propagate and cause the model to learn spurious correlations in downstream tasks. Efforts have been made to mitigate these data-derived biases. One approach involves task-agnostic methods that enforce the learning of fair representations (Kaneko and Bollegala, 2021; Cheng et al., 2021; Guo et al., 2022), while another focuses on reducing discrimination in specific tasks using annotated data (Shen et al., 2021; Ravfogel et al., 2022; Gira et al., 2022; Zhu et al., 2023).

Confounding bias is a particular type of bias that arises when the relationship between the anticipated signal and the outcome is distorted by the presence of extraneous factors, known as confounders. This results in discrepancies in model performance across different confounder strata. In the context of text classification, a confounder can be considered as an extraneous variable that influences both the language provided to a classifier, and the distribution of the class labels of interest (Landeiro and Culotta, 2018). In this study, we focus on task-specific methods to mitigate the effects of confounding bias. Specifically, we investigate confounding shift in binary classification, where

[1] Code repo for reproducing all the experiment results will be published upon acceptance.

positive examples are unevenly distributed across different subgroups, and these subgroup-specific class distributions differ at the point of evaluation or deployment. This can lead to errors when a model uses language indicating a subgroup, rather than the outcome of interest, as a basis for prediction. Inspired by the Confounding Filter (Wang et al., 2019), we propose two novel techniques: the Extended Confounding Filter and Dual Filter, and evaluate them on the *DementiaBank* dataset, a first-person narrative dataset collected from cognitive impairment assessments, widely used to study the effects of Alzheimer's disease dementia on language.

Our main contributions in this paper are as follows:

- We identified gender confounding bias in *DementiaBank*, which had not been reported previously for dementia detection in any picture description dataset.
- We extended the Confounding Filter method which targets specific layers in a neural network to the Transformer architecture and demonstrated improvements in task performance.
- We introduced the Dual Filter as a novel weight masking algorithm, that identifies and ablates parameters associated with the confounding bias in the entire model's network (vs. individual layers).

2 Related Work

Our work focuses on bias mitigation through weight masking, which requires finding and isolating the influence of model weights that represent information about a confounding variable. As such, our work relates to prior efforts to access information encoded within transformer networks. Meng et al. (2023) analyze the factual information stored in GPT2 (Radford et al., 2019) and develop a causal intervention on neuron activations to trace the information flow that determines the model's predictions. A causal intervention modifies certain weights inside the network and evaluates the altered model outcome. Other work has also used causal interventions to probe the behavior of language models (Vig et al., 2020; Elazar et al., 2021). To locate the neurons associated with specific information or functionality within the network, Liu et al. (2024) propose a gradient integration method to pinpoint neurons that cause gaps in output logits distribution among demographic groups. There are also other scoring metrics used for pruning neural networks (Lee et al., 2019; Sun et al., 2024) that can be used for locating associated weights. They either track neuron activation or loss output by masking certain weights within a layer and assign an importance score to each entry given a calibrated dataset. Compared to these prior efforts, the method we employ for identifying the weights complies with the training procedure and requires no granular weight inspection yet still yield desired debiasing effects.

3 Methods

3.1 Confounding Filter

Deep learning models often recognize false signals from confounding factors, leading to sub-optimal performance in many real-world cases (Szegedy et al., 2013; Nguyen et al., 2015; Wang et al., 2017b,a). To address this issue, the Confounding Filter (Wang et al., 2019) was proposed to address confounding biases in models trained on electroencephologram and medical imaging data. The Confounding Filter method is straightforward and model-agnostic, designed to mitigate the impact of confounding factors.

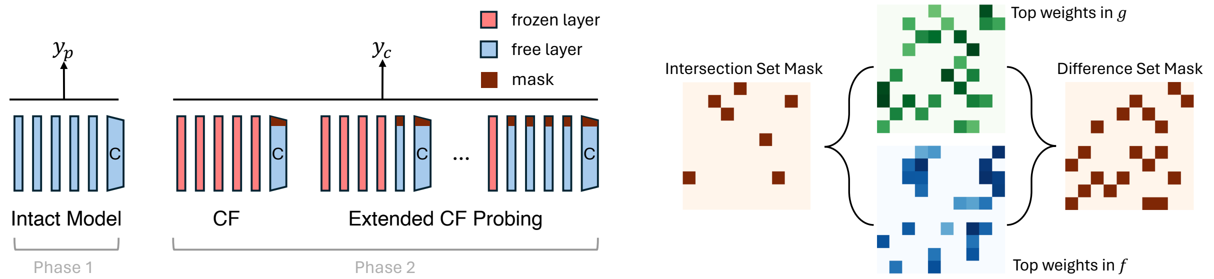
In this approach, a deep learning model is denoted as having two components: $g(\cdot;\theta)$, a representation learning network, and $f(\cdot;\phi)$, a classification network. The algorithm first optimizes the entire network by solving the following objective:

$$\hat{\theta}, \hat{\phi} = \arg\min_{\theta,\phi} \mathcal{L}(y, f(g(X);\theta);\phi)$$

where $\mathcal{L}$ denotes the loss function to be minimized.

In the second phase, assuming we have access to the confounder label m in the dataset, the algorithm localizes weights that are reactive to the confounding variable. This is achieved through tuning $f(\cdot;\phi)$ towards M while keeping $g(\cdot;\theta)$ fixed. During the second phase, updates in $\hat{\phi}$ are tracked and normalized after each batch. The sum of normalized updates is denoted as $\pi = \frac{1}{b}\sum_{i=1}^{b}|\Delta\phi_i|$ where b is the number of total batches in the second phase of training. The importance of each element in π is determined by their magnitude. A threshold function is then employed to get the mask:

$$M_i = \begin{cases} 0 & \text{if } \pi_i > \tau \\ 1 & \text{otherwise} \end{cases}$$



(a) (b)

Figure 1: **a**: Illustration of the Extended Confounding Filter (ECF) Probing framework for weights identification. **b**: Illustration of the Dual Filter (DF) procedure to find weights to mask.

Here, τ is the k^{th} percentile in π, where k is a hyperparameter. The element-wise product $\hat{\phi}' = \hat{\phi} \otimes M$ results in the confounder-mitigated network $f(g(X); \hat{\theta}); \hat{\phi}')$.

3.2 Extended Confounding Filter

While the original Confounding Filter algorithm has shown improvements over the baseline in some neural network architectures (Wang et al., 2019), its adaptation to transformer networks remains unexplored. Transformer-based language models learn to generate distributional semantic representations (Vaswani et al., 2017) through the attention mechanism and positional encoding. By fine-tuning a pretrained language model, semantic information pertinent to a task of interest is dynamically stored across the transformer network layers.

Our hypothesis is that fixing $g(\cdot; \theta)$ when training for the confounder variable may not effectively capture the most confounder-associated weights within the transformer network. To test this hypothesis, we sequentially unfroze each layer in the transformer network, starting from the top layer down to the embedding layer and observed its impact on the outcome. This is different from the original Confounder Filtering method, where only the classification head is trainable in the encoder model.

As shown in Figure 2, the matrices $W_Q, W_K, W_V, W_O, W_1, W_2$ are tracked in a single transformer block, while W_{emb} and W_{cls} represent the token embedding matrix and classification weight matrix in a sequence classification model, respectively. Similarly to the Confounding Filter, we start by training a classification model towards the primary outcome Y_p (Phase 1) and then continue training the model towards classifying the confounder Y_c (Phase 2).

By sequentially unfreezing different numbers of

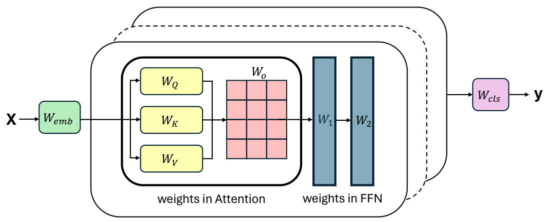


Figure 2: Tracked weight matrices in the transformer network

layers, we allow varying amounts of the model's parameter spaces to react to the information introduced during Phase 2. (Figure. 1a) This sequential probing scheme follows the idea from Confounding Filter but provides more flexibility as you can partition the classification network $f(x)$ and representation learning network $g(x)$ at different points. The change in parameter $\Delta\phi_i$ is normalized within the matrix and recorded after each training batch. Following the Confounding Filter methodology, we restrict $\Delta\phi_i$ to each W in this probing procedure, and the threshold τ is calculated for each individual weight matrix. The probing step size is by layer. Masking matrices, derived from the threshold function, are applied to the tracked weight matrix from Phase 1 fine-tuning. We later evaluate the effectiveness of this method in mitigating confounding bias against the probing depth on a real world dataset.

3.3 Dual Filter

Next, we further lift the restriction on Phase 2 training from the ECF method that the masking be performed locally, ignoring the dynamics and interaction the language model might have during finetuning. We propose Dual Filter, a method that tracks the weight change from two separate models starting from the same checkpoint, one for the primary target and the other for the confounder. After obtaining change matrices π from both models, we utilize set operations to isolate weights that are most reactive to the confounder label during finetuning. Specifically, we chose top $k\%$ most changed weights from the primary model f and the confounder model g, and take the intersection or the difference from these two weight sets to generate the mask matrices (Figure 1b). One could apply either the intersection set mask, the difference set mask, or the joint set of the two masks (which is equivalent to the top $k\%$ most changed weights from the confounder model), depending on the dataset or tasks. We formally describe the proposed algorithm in Algorithm 1.

Algorithm 1 Dual Filter for weights masking

Input: pretrained language model: $f_0(x)$, $g_0(x)$; dataset: $\mathcal{D}(x, y_p, y_c)$; threshold: k

Output: Confounder-adjusted model $f(x;\theta')$

1: Train $f_0(x;\theta) \mapsto y_p$, obtain weights change Δ_p and finetuned model $f(x;\hat{\theta})$.

2: Train $g_0(x;\phi) \mapsto y_c$, obtain weights change Δ_c and finetuned model $g(x;\hat{\phi})$.

3:

$$\Delta_{p,k} = \underset{p \subseteq \Delta_p, |p|=k}{\arg\max} \sum_{p_i \in \Delta_p} p_i$$

$$\Delta_{c,k} = \underset{c \subseteq \Delta_c, |c|=k}{\arg\max} \sum_{c_i \in \Delta_c} c_i$$

4: $M_I \leftarrow \Delta_{p,k} \cap \Delta_{c,k}$, $M_D \leftarrow \Delta_{c,k} \setminus \Delta_{p,k}$

5: Pick mask $M \in \{M_I, M_D, M_I \cup M_D\}$

6:

$$\theta' \leftarrow \hat{\theta}_i = 0 \quad \forall i \in M$$

4 Evaluations

Confounding Shift One fundamental assumption in machine learning is the test dataset and training dataset are from the same distribution. However this assumption is often violated in real world applications resulting in distribution shifts. One specific form of distribution shift is sub-population shift (Cao et al., 2019; Cai et al., 2021). A model optimized on a distribution shifted training set tends to learn spurious correlations with the majority class and may lead to poor performance on data with a different from the training data class distribution (Yang et al., 2023).

While the sub-population shifts are determined by the product of group attributes and the label, and the group attributes are not independent of the label, it is a special type of dataset shift referred to as *Confounding Shift* (Landeiro and Culotta, 2018). Formally, confounding shift exists when two conditions are met: (i) a confounding variable Y_c exists that impacts both X and Y_p through distributions $P(X|Y_c)$ and $P(Y_p|Y_c)$ through the backdoor path in a causal graph (Pearl, 2009); (ii) a subpopulation distribution $P_{train}(Y_p|Y_c)$ is different from $P_{test}(Y_p|Y_c)$ (Landeiro and Culotta, 2018).

To quantitatively assess the degree of confounding shift, we use a framework proposed by Ding et al. (2024) in our experiments. This allows us to perturb the target variable and confounding variable distributions in both training and test splits to different degrees through sampling from the original dataset. Under this framework, we consider a dataset with a binary target and binary confounder, the joint distribution $P(Y_p, Y_c)$ governed by the following quantity: $P(Y_c = 1), P(Y_p = 1), P(Y_p = 1|Y_c = 1), P(Y_p = 1|Y_c = 0)$. Next Ding et al. (2024) introduced an positive auxiliary variable $\alpha = \frac{P(Y_p=1|Y_c=1)}{P(Y_p=1|Y_c=0)}$, which serves as a knob for controlling the degree of subpopulation shift. By setting different α values, we control the source of the positive examples. If we hold $P(Y_c = 1)$ and $P(Y_p = 1)$ constant, we can vary α_{train} and α_{test} to create a mixture of datasets with various degrees of shift for model evaluation. Details are described in Section 5.2.

Fairness The concept of fairness in machine learning addresses the goal of ensuring that models operate without bias and equitably across different demographic groups. A widely accepted notion of group fairness, which focuses on equity at the population level, is statistical parity (Dwork et al., 2011). In problems with a binary outcome Y and a binary group variable G, statistical parity is defined as the absolute difference or ratio between $P(\hat{y} = 1|g = 1)$ and $P(\hat{y} = 1|g = 0)$. Smaller values of statistical parity indicate greater equality in the model's outputs across the two groups. In addition to statistical parity, other fairness metrics consider ground truth labels and compare the true positive rates between groups (Romano et al., 2020; Hardt et al., 2016). These metrics assess the model's ability to make accurate predictions without discriminating against any group.

In our context, the test set attributes vary due to different data distributions associated with parameter α, rendering comparisons of statistical parity across different α values infeasible. Therefore, we evaluate $P(\hat{y} = 1|G, y = 0)$, which describes the predicted probability for dementia among healthy participants, and is equivalent to the false positive rate (FPR). We calculate the absolute difference of FPR between the subgroups. This metric helps us assess fairness by examining the model's behavior across different α values.

5 Dementia Detection Case Study

In recent years, transformer models have demonstrated promising performance in dementia detection using *Cookie Theft* picture description data (Figure S2) (Hernandez-Dominguez et al., 2018; Cohen and Pakhomov, 2020; Luz et al., 2020; Guo et al., 2021; Li et al., 2022), a clinical test widely adopted for assessing cognitive impairment. However, these models are susceptible to bias due to the small size of publicly available datasets utilized in most studies. Within this context, confounding by gender is an unexplored potential source of bias that could lead to erroneous predictions if the confounding effects are not addressed. The underlying hypothesis is that the language used by male and female participants in response to the picture description task may vary, and the model might learn these differences to make dementia predictions, regardless of the participants' true cognitive status.

5.1 DementiaBank

The benchmark dataset used for our experiments is the Pittsburgh Corpus from DementiaBank (Becker et al., 1994; MacWhinney, 2007) This corpus is a widely used resource in the fields of computational linguistics and dementia studies. It provides detailed speech and language data from elderly participants with dementia as well as healthy controls. Notably, the Pittsburgh Corpus includes responses to the Cookie Theft picture description task from the Boston Diagnostic Aphasia Examination (Goodglass and Kaplan, 1983). The dataset comprises 548 examples collected from longitudinal records of 290 participants. To ensure that the transcripts accurately reflect the diagnosis label, we selected the last transcript for each patient as input for our model.

5.2 Experiments

We start by examining whether a text classification model will recognize gender confounding bias from such picture description data. We trained a BERT-base model (Devlin et al., 2019) on the full dataset and evaluated the model's performance on the task of recognizing each gender [2]. We ran the experiments using 5-fold cross validation with 3 repeats on both the original dataset, and a perfectly balanced dataset created by down-sampling the more prevalent category. The result is shown in Figure 3 - performance discrepancies were observed among male and female examples across multiple runs. These findings hold for some other encoder models as well (Figure S1). This result shows that there exists confounding by gender in the dementia detection task which is independent of the gender distribution in the dataset. It provides insights that the gender of the speaker influences the language they use to complete the Cookie Theft picture description task, and confound the dementia signals during model fine-tuning. Hereby, we further investigate this confounding by gender effects in dementia detection and evaluate our proposed deconfounding methods.

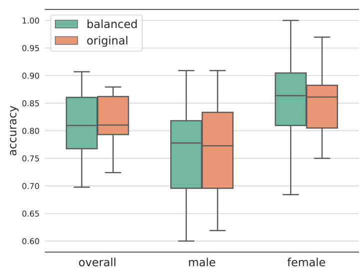


Figure 3: Performance discrepancies in dementia detection when trained with a BERT-base model

Dataset Perturbation As described in Section 4 , we manipulated the conditional distribution of dementia by gender in our dataset through random sampling, creating a series of datasets with varying levels of confounding shift. In our experiments, dementia cases and female cases are coded as 1, respectively. We fixed $P(\text{gender} = 1) = 0.5$ and $P(\text{dementia} = 1) = 0.5$ in both the training and test sets to ensure fair comparisons across different configurations. This way, the dataset is balanced with respect to both dementia and gender. Then we adjusted the value of $\alpha = \frac{P(\text{dementia}|\text{female})}{P(\text{dementia}|\text{male})}$ to create an imbalance in the source of dementia cases (subpopulation shift). If $\alpha > 1$, more dementia cases are drawn from females, while $\alpha < 1$ indicates the opposite. The further α is from 1, the more severe the imbalance. To evaluate the model's robustness to confounding shifts, the model is trained

[2]This dataset provides labels only for two genders: male and female

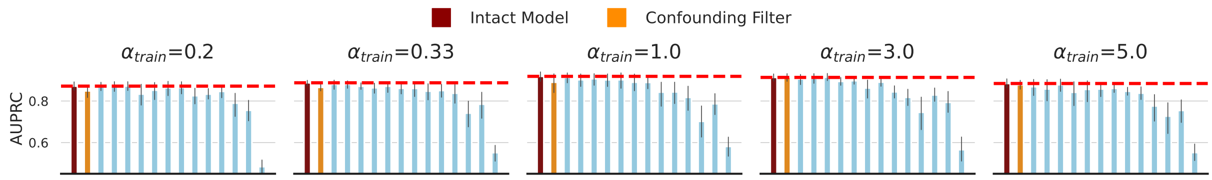


Figure 4: Extended Confounding Filter with 15% masking ratio at each tracked weight matrix

on one α_{train} value and tested on its reciprocal value $\alpha_{\text{test}} = \frac{1}{\alpha_{\text{train}}}$, simulating an extreme shift in the test set compared to the distribution the model was exposed to during training. Models are trained for 20 epochs on 600 training examples and evaluated on 150 examples for each configuration. The best checkpoint is selected based on AUPRC for each training process.

Extended Confounding Filter On creating a series of datasets with different α_{train} values, we probed each layer in the model in response to the shift. The encoder model we used for dementia detection is BERT-base, with 12 encoder layers and 12 attention heads in each layer. Once we obtain the dementia finetuned model $f(x)$ in the first Phase, we take a snapshot of the parameters and only make some parts of it trainable towards the gender label in the second Phase. The trainable layer starts from $\{cls\}$, and 1 layer is added to the trainable set each time sequentially. Eventually the trainable set becomes $\{cls, layer12, layer11, ..., layer1, emb\}$ and spans the whole network. Then for each trainable set, f_d is trained towards gender prediction. We ranked weights that changed in each layer and picked top 15% of the weights that changed the most in each layer to mask (Figure 1a). Then we evaluated the masked models.

Dual Filter In the Dual Filter approach, we track the global weights change throughout the model's architecture. The classification head is exempt from tracking as it is training towards two different tasks and the weights in the classification head are assumed to have the most significant change compared to the rest of network. We first obtain two lists of weights change matrices from $f(x)$ and $g(x)$, using the same approach as Extended CF. Then we rank and select top k% weights by their locations in the network. A sequence of k values are tested, ranging from 0 to 60 and step size of 1. Then three kinds of set $(M_I, M_D, M_I \cap M_D)$ are calculated and applied to $f(x)$ to create the masked model. Note when training toward gender in both Extended CF and Dual Filter, we select only non-dementia cases to let the model learn from texts that are representative of the gender differences. consequently, only healthy cases are used in the evaluation as well.

6 Results

6.1 Extended Confounding Filter

Figure 4 demonstrates the Extended Confounding Filter results, the red dotted line shows the performance of the intact model and models whose weights are eliminated cumulatively layer-by-layer from left to right until the embedding layer is reached (the right most bar). The orange bar represents the idea of the original Confounding Filter, where only the classification head is trained in the second phase and then masked. These results show that a model trained and tested on the same distribution reaches the highest performance among all the configurations, while the degree of confounding shift correlates with the model's performance (i.e. the model performance drops as α_{train} shifts away from 1). Another observation is that the model demonstrates some resilience in its ability to detect dementia to removing gender associated weights from upper layers in the network. No significant performance drops are observed until we start to remove weights at 5^{th} layer. The next observation we have is the large decline when top changed weights are removing from the embedding layer.

6.2 Dual Filter

In Figure 5, we visualize the dementia prediction performance change as we apply three different types of mask to the original model and gradually increase the masking ratio. The results from ECF with 15% layer-specific masking ratio have also been added for comparison. The plot shows the relation between how many weight entries are ablated within the whole network against dementia detection performance in terms of AUPRC. The rows indicate three types of masks that are generated by Dual Filter and the columns indicate the specific α_{train} configurations that control the distribution shift. The relationships between the ablation

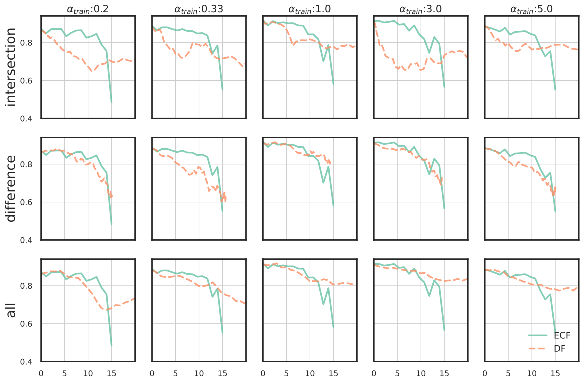


Figure 5: Side by side AUPRC comparison on ECF and DF for different α_{train} configurations

ratio of the three types of masks and the choice of k are shown in Figure 6. As we tune k to increase the coverage of active parameters in the model, the size of M_D first grows then reaches its peak at around $k = 40$ and then fall back to zero, while the size of M_I keeps increasing.

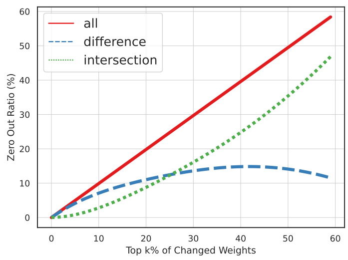


Figure 6: Ablation ratio by each masks against total masking ratio

Next, we show the absolute False Positive Rate difference (i.e. $|P(\hat{y} = 1|g = 1, y = 0) - P(\hat{y} = 1|g = 0, y = 0)|$) calculated under both Extended Confounding Filter and Dual Filter methods. The Figure 7 shows the FPR measurements change as the ablation ratios increase for all three types of masks. The mask type is indicated by row while the columns represents different α_{train} .

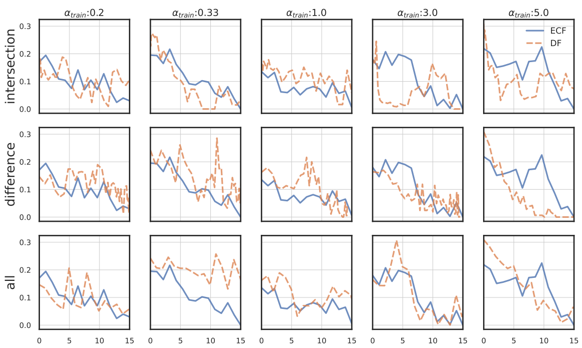


Figure 7: False Positive Rate (FPR) on ECF and DF for different α_{train} configurations

While the aim is to eliminate gender confounding effects from the model's dementia detection capability, there is a possibility that the weights associated with dementia and gender become entangled during the learning process. To investigate this, we record the change matrices for all layers in the network during the Dual Filter training process. We then conduct an analysis of the similarity between the change matrices from the fine-tuned dementia model and those from the fine-tuned gender model. For similarity measurements, we utilize the Jaccard Index to quantify the similarity between the two input matrices, which is defined as:

$$J(U, V) = \frac{|U \cap V|}{|U \cup V|}$$

To prepare the input, 85% percentile of two change matrices are calculated and then the values are used to binarize each of the matrices. Figure 8 demonstrates the barplot from six of the tracked weight matrices at each layer, with the configuration of α_{train} equals to 1. From the plot we can observe that at lower encoder layers, the similarity between dementia model and gender model concentrates on the attention block, especially W_V and W_O. As we move up to the upper layer, the FFN block starts to display more similarity and jumps up at 12^{th} layer. Similar patterns are also observed in other α_{train} configurations. This result indicates the finetuned model stores information dynamically through the whole network and shift the storage at different layers. This finding also aligns with other work (Wei et al., 2024) where weights entanglement are assessed with a larger model and different tasks.

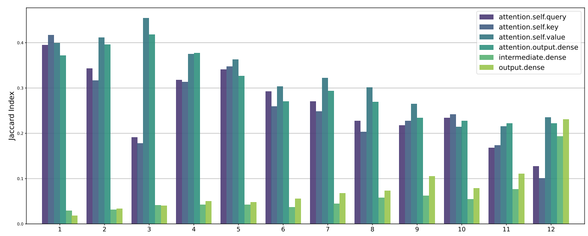


Figure 8: Jaccard Index for each of the tracked matrix in Dual Filter

7 Discussion

The ECF method probes each layer and mask associated weights in a cumulative fashion. The orange patch in Figure 4 shows simply applying Confounding Filter on the classification layer to the transformer network is not enough to detect and mitigate the confounding bias. Propagating masks layer-by-layer helps improve or retain the dementia classification performance from Confounding Filter until several layers deep into the network. We observe the resilience of weight ablation in the BERT model on dementia prediction performance which is consistent with similar resilience on capturing linguistic features reported in other work (Li et al., 2024). Model performance on dementia does not drop significantly until gender-associated weights at layer 5 or layer 6 gets masked, depending on the configuration of α. This is equivalent to weights ablation ratio around 10%-12% over the whole network. We also observe the network is slightly more robust to weight ablation when the confounding shift is more severe under this probing method. For example, with configuration of $\alpha = 1$, the drop starts before 10% of weights (around layer 5) are removed; In the meanwhile, with $\alpha_{train} = 5$ or $\alpha_{train} = 0.2$, the drop is delayed until weights at layer 2 or layer 3 gets deleted.

We also check the gender performance difference as gender-associated weights are removed in each layer. Interestingly, the masked models show a loss of gender detection ability (Figure S3), and the extent of this loss varies with the number of free layers. This demonstrates that the transformer learns information dynamically and adapts to its model capacity. From Figure S3 we also notice the effects of different level of confounding shift. While the gender performance difference by layers is not reflected balanced setting, we observe AUROC performance change gets larger by layer under confounding shifts. The embedding matrix also emerges as a critical factor in dementia detection, as deleting even a small proportion of the embedding weights causes a drastic change in the model's dementia detection ability.

The grid of Figure 5 provides an overview of the performance change in dementia against the ablation ratios in the model, for both ECF and DF. In the results we can observe the different behaviors from the three types of mask. For $M_I \cup M_D$ filter, the model illustrates resistance against the weight deletion more consistently. Comparing to ECF, masking weights inside $M_I \cup M_D$ shows less resilience at the start but the performances cross right after 10% weights are removed and then becomes more robust to weight delection compared to ECF. As for M_D, some resilience can be observed at the start of the masking but the performance degradation becomes more extensive after a certain point. Also, the resilience behavior differs across different α_{train}, with $\alpha_{train} = 1$ offering the most resilience. However, the impact of M_I mask turned out to be different. Ablating weights in M_I first results in a sharp decrease in performance and then gradually stabilizes when more weights get deleted. Those results suggest the entanglement of weights responsive for dementia detection and gender detection enrich in the intersection set from two change matrix, especially those weight entries that have changed most. Interestingly, removing all the top changed weights from the gender model side consistently exhibits more resilience than removing weights only from the difference set. We also observe the ECF method in general preserve better dementia detection capability if weights are only removed from top half layers in BERT base model (layer 6-12). By examining the between-group FPR metric in Figure 7), both methods show improvements in output equity. By align the fairness metrics with the AUPRC changes, we clearly observe the correlation between the dementia detection ability and disparity between gender group. For example in the M_I row, when dementia performance recovers, it aligns with an sudden increase in the FPR difference.

8 Conclusion

In this paper, we address confounding bias learned during model fine-tuning and propose two model-agnostic methods for filtering confounding-associated weights in transformers. We apply these methods to a dementia detection task, demonstrating their utility in clinical practice. Our findings indicate that unaddressed confounding shifts degrade model performance even when the overall label and group distributions are balanced. Experimental results compare the identification of gender-associated weights both layer-wise and across the entire model. Both methods effectively retain performance on dementia detection while reducing gender bias. Although these results are dataset-specific, we plan to extend our approach to other benchmarks. We observe non-monotonic responses across different layers, suggesting further investigation is needed to understand the inner workings of even small transformer models. Lastly, we note that ensuring fairness and maintaining model performance often involves trade-offs, and real-world decisions should consider multiple factors, including bias tolerance and use case specifics.

9 Limitations

Dataset The experiments of our proposed methods are only conducted on a relative small dataset, generalizability to other bias-related dataset remains unknown. In addition, given the small data size, manifesting different level of confounding shift requires repetitive sampling to meet the desired subgroup distribution. Thus the resultant dataset contains significant amount of duplicates that can impact the validity of the findings.

Methods In Extended Confounding Filter methods, even though the approach we take is the most straightfoward and allows the model to absorb uni-directional effects, we ignore the possibility of other combinations of layer freezing inside the network.

Experiments While we acknowledge BERT-base as a good starting point of investigation, we did not compare other encoder model in this work. On the other hand, we briefly discussed some other weight importance measurements to isolate weights that impact certain outputs, we didn't implement and compare them with our current approach for de-confounding bias.

References

Ioana Baldini, Dennis Wei, Karthikeyan Natesan Ramamurthy, Mikhail Yurochkin, and Moninder Singh. 2022. Your fairness may vary: Pretrained language model fairness in toxic text classification. *Preprint*, arXiv:2108.01250.

JT Becker, F Boller, OL Lopez, J Saxton, and KL McGonigle. 1994. The natural history of alzheimer's disease: Description of study cohort and accuracy of diagnosis. *Archives of Neurology*, 51(6):585–594.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Preprint*, arXiv:1607.06520.

Tianle Cai, Ruiqi Gao, Jason D. Lee, and Qi Lei. 2021. A theory of label propagation for subpopulation shift. *Preprint*, arXiv:2102.11203.

Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. 2019. Learning imbalanced datasets with label-distribution-aware margin loss. In *Advances in Neural Information Processing Systems*.

Pengyu Cheng, Weituo Hao, Siyang Yuan, Shijing Si, and Lawrence Carin. 2021. Fairfil: Contrastive neural debiasing method for pretrained text encoders. In *International Conference on Learning Representations*.

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. *Preprint*, arXiv:2005.03593.

Daniel de Vassimon Manela, David Errington, Thomas Fisher, Boris van Breugel, and Pasquale Minervini. 2021. Stereotype and skew: Quantifying gender bias in pre-trained and fine-tuned language models. *Preprint*, arXiv:2101.09688.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Xiruo Ding, Zhecheng Sheng, Meliha Yetişgen, Serguei Pakhomov, and Trevor Cohen. 2024. Backdoor adjustment of confounding by provenance for robust text classification of multi-institutional clinical notes. In *AMIA ... Annual Symposium proceedings. AMIA Symposium*, pages 923–932.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Rich Zemel. 2011. Fairness through awareness. *Preprint*, arXiv:1104.3913.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. 2021. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Preprint*, arXiv:2006.00995.

Michael Gira, Ruisu Zhang, and Kangwook Lee. 2022. Debiasing pre-trained language models via efficient fine-tuning. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 59–69, Dublin, Ireland. Association for Computational Linguistics.

Harold Goodglass and Edith Kaplan. 1983. *Boston Diagnostic Aphasia Examination Booklet*. Lea & Febiger, Philadelphia.

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.

Yue Guo, Yi Yang, and Ahmed Abbasi. 2022. Autodebias: Debiasing masked language models with automated biased prompts. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL 2022)*. Association for Computational Linguistics.

Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of opportunity in supervised learning. *Preprint*, arXiv:1610.02413.

Luis Hernandez-Dominguez, Samuel Ratté, Basilio A. Sierra, and Jesus A. Roche-Berges. 2018. Computer-based evaluation of alzheimer's disease and mild cognitive impairment using lexical and syntactic information. *Journal of Alzheimer's Disease*, 63(2):709–719.

Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. 2020. Social biases in nlp models as barriers for persons with disabilities. *Preprint*, arXiv:2005.00813.

Masahiro Kaneko and Danushka Bollegala. 2021. Debiasing pre-trained contextualised embeddings. *Preprint*, arXiv:2101.09523.

Virgile Landeiro and Aron Culotta. 2018. Robust text classification under confounding shift. *J. Artif. Int. Res.*, 63(1):391–419.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Namhoon Lee, Thalaiyasingam Ajanthan, and Philip H. S. Torr. 2019. Snip: Single-shot network pruning based on connection sensitivity. *Preprint*, arXiv:1810.02340.

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, Zhecheng Sheng, Trevor Cohen, and Serguei Pakhomov. 2024. Too big to fail: Larger language models are disproportionately resilient to induction of dementia-related linguistic anomalies. *Preprint*, arXiv:2406.02830.

Yan Liu, Yu Liu, Xiaokang Chen, Pin-Yu Chen, Daoguang Zan, Min-Yen Kan, and Tsung-Yi Ho. 2024. The devil is in the neurons: Interpreting and mitigating social biases in language models. In *The Twelfth International Conference on Learning Representations*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

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 *Proceedings of INTERSPEECH 2020*, Shanghai, China.

Brian MacWhinney. 2007. The talkbank project. In *Creating and Digitizing Language Corpora*, pages 163–180. Palgrave Macmillan, London.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2023. Locating and editing factual associations in gpt. *Preprint*, arXiv:2202.05262.

Anh Nguyen, Jason Yosinski, and Jeff Clune. 2015. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Judea Pearl. 2009. *Causality*, 2nd edition. Cambridge University Press, Cambridge, UK.

Rebecca Qian, Candace Ross, Jude Fernandes, Eric Smith, Douwe Kiela, and Adina Williams. 2022. Perturbation augmentation for fairer nlp. *Preprint*, arXiv:2205.12586.

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

Shauli Ravfogel, Michael Twiton, Yoav Goldberg, and Ryan Cotterell. 2022. Linear adversarial concept erasure. *Preprint*, arXiv:2201.12091.

Yaniv Romano, Stephen Bates, and Emmanuel J. Candès. 2020. Achieving equalized odds by resampling sensitive attributes. *Preprint*, arXiv:2006.04292.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *Preprint*, arXiv:1910.01108.

Aili Shen, Xudong Han, Trevor Cohn, Timothy Baldwin, and Lea Frermann. 2021. Contrastive learning for fair representations. *Preprint*, arXiv:2109.10645.

Mingjie Sun, Zhuang Liu, Anna Bair, and J. Zico Kolter. 2024. A simple and effective pruning approach for large language models. *Preprint*, arXiv:2306.11695.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2013. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.

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

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Simas Sakenis, Jason Huang, Yaron Singer, and Stuart Shieber. 2020. Causal mediation analysis for interpreting neural nlp: The case of gender bias. *Preprint*, arXiv:2004.12265.

Haohan Wang, Akshay Meghawat, Louis-Philippe Morency, and Eric P Xing. 2017a. Select-additive learning: Improving generalization in multimodal sentiment analysis. In *IEEE International Conference on Multimedia and Expo*.

Haohan Wang, Bhiksha Raj, and Eric P Xing. 2017b. On the origin of deep learning. *arXiv preprint arXiv:1702.07800*.

Haohan Wang, Zhenglin Wu, and Eric P. Xing. 2019. Removing confounding factors associated weights in deep neural networks improves the prediction accuracy for healthcare applications. In *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, volume 24, pages 54–65.

Kellie Webster, Xuezhi Wang, Ian Tenney, Alex Beutel, Emily Pitler, Ellie Pavlick, Jilin Chen, Ed Chi, and Slav Petrov. 2021. Measuring and reducing gendered correlations in pre-trained models. *Preprint*, arXiv:2010.06032.

Boyi Wei, Kaixuan Huang, Yangsibo Huang, Tinghao Xie, Xiangyu Qi, Mengzhou Xia, Prateek Mittal, Mengdi Wang, and Peter Henderson. 2024. Assessing the brittleness of safety alignment via pruning and low-rank modifications. *Preprint*, arXiv:2402.05162.

Yuzhe Yang, Haoran Zhang, Dina Katabi, and Marzyeh Ghassemi. 2023. Change is hard: A closer look at subpopulation shift. *Preprint*, arXiv:2302.12254.

Beier Zhu, Yulei Niu, Saeil Lee, Minhoe Hur, and Hanwang Zhang. 2023. Debiased fine-tuning for vision-language models by prompt regularization. *Preprint*, arXiv:2301.12429.

A Appendix

A.1 Gender bias in dementia detection

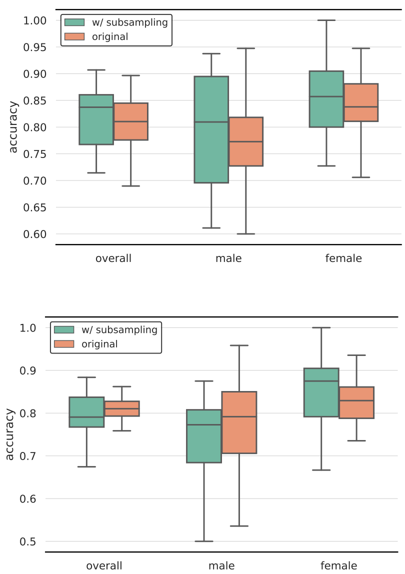


Figure S1: Performance disparency between male and female in from other encoder models. **Top**: Results from RoBERTa-base, **Bottom**: Results from FairBERTa (Qian et al., 2022)

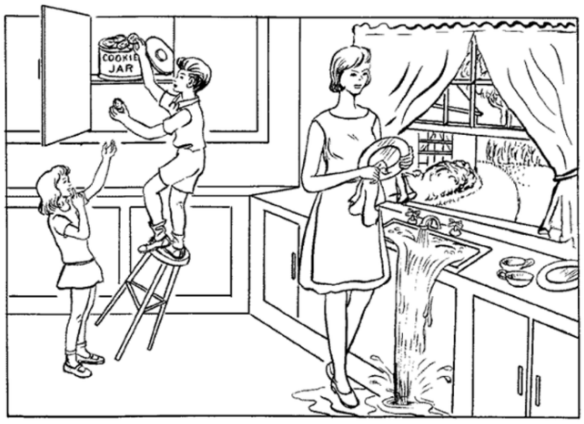


Figure S2: Cookie Theft picture for cognitive impairment assessment

A.2 Cookie Theft Picture

A.3 Gender performance change

We first evaluate model performance of AUROC on gender prediction for each layer probing process, then evaluate gender performance again after removing gender-associated weights and calculate their difference. As shown in the Figure S3 below, the performance gap between original and masked model becomes larger as more gender associated weights are removed in the confounding shift scenarios.

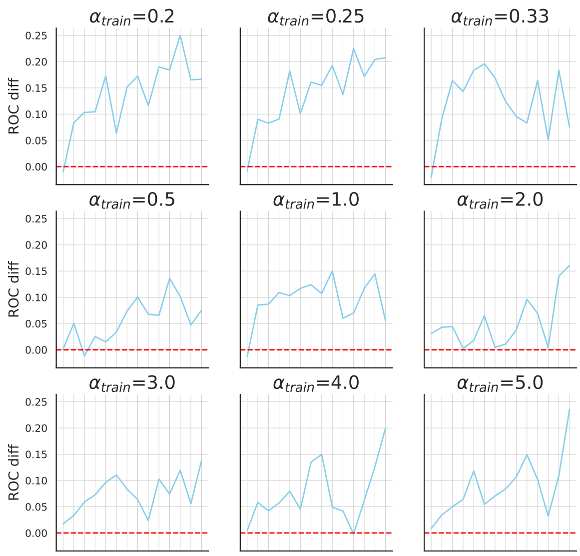


Figure S3: Performance difference between a intact and masked model for gender prediction. Left-most point means probe on classification layer and right-most means probe on the word embedding layer.

A.4 Additional Results

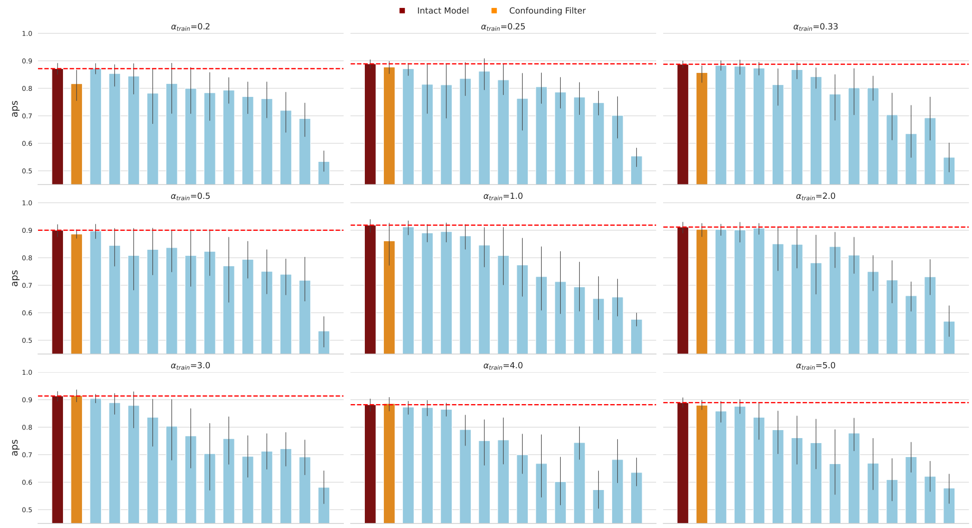


Figure S4: Selected configurations of ECF filtering with 25% masking rate at each tracked weight matrix

Table 1: Mean and Standard Deviation of APS for each experiment and α_{train} in ECF with 15% masking rate

	$\alpha_{train} =$ 0.2	$\alpha_{train} =$ 0.25	$\alpha_{train} =$ 0.33	$\alpha_{train} =$ 0.5	$\alpha_{train} =$ 1.0	$\alpha_{train} =$ 2.0	$\alpha_{train} =$ 3.0	$\alpha_{train} =$ 4.0	$\alpha_{train} =$ 5.0
Intact	0.8716 (0.0350)	0.8917 (0.0345)	0.8876 (0.0217)	0.9001 (0.0404)	**0.9186** (0.0400)	0.9120 (0.0335)	0.9135 (0.0305)	0.8818 (0.0385)	0.8842 (0.0378)
Classifier	0.8486 (0.0528)	0.8807 (0.0427)	0.8657 (0.0290)	**0.8942** (0.0308)	0.8904 (0.0778)	0.9051 (0.0406)	0.9150 (0.0328)	0.8844 (0.0391)	0.8782 (0.0363)
Layer12	0.8708 (0.0324)	0.8888 (0.0398)	0.8803 (0.0353)	0.8977 (0.0415)	**0.9132** (0.0415)	0.9049 (0.0443)	0.9059 (0.0417)	0.8694 (0.0399)	0.8686 (0.0661)
Layer11	0.8717 (0.0401)	0.8499 (0.0971)	0.8803 (0.0326)	0.9046 (0.0338)	0.9023 (0.0537)	0.9060 (0.0398)	**0.9094** (0.0388)	0.8803 (0.0543)	0.8579 (0.0873)
Layer10	0.8720 (0.0384)	0.8351 (0.1265)	0.8712 (0.0284)	0.8439 (0.1446)	0.9071 (0.0502)	0.9093 (0.0300)	**0.9145** (0.0303)	0.8801 (0.0319)	0.8781 (0.0499)
Layer9	0.8334 (0.0880)	0.8681 (0.0514)	0.8635 (0.0421)	0.8687 (0.1083)	**0.9016** (0.0507)	0.8702 (0.1044)	0.8942 (0.0318)	0.8599 (0.0478)	0.8421 (0.0977)
Layer8	0.8520 (0.0705)	0.8831 (0.0543)	0.8712 (0.0432)	0.8854 (0.0639)	0.9022 (0.0630)	**0.9076** (0.0340)	0.8977 (0.0296)	0.8644 (0.0440)	0.8559 (0.0910)
Layer7	0.8636 (0.0572)	0.8665 (0.0500)	0.8612 (0.0453)	0.8565 (0.1184)	0.8904 (0.0672)	**0.9014** (0.0306)	0.8625 (0.0776)	0.8700 (0.0352)	0.8577 (0.0564)
Layer6	0.8648 (0.0511)	0.8456 (0.0933)	0.8606 (0.0596)	0.8742 (0.0967)	0.8893 (0.0499)	0.8466 (0.1167)	**0.8909** (0.0349)	0.8497 (0.0369)	0.8607 (0.0363)
Layer5	0.8256 (0.0656)	0.8348 (0.0729)	0.8475 (0.0663)	**0.8670** (0.0480)	0.8428 (0.1006)	0.8654 (0.0659)	0.8440 (0.0513)	0.8347 (0.0482)	0.8469 (0.0335)
Layer4	0.8331 (0.0449)	0.8418 (0.0669)	0.8503 (0.0534)	**0.8887** (0.0420)	0.8437 (0.0869)	0.8282 (0.0684)	0.8178 (0.0704)	0.7441 (0.1416)	0.8380 (0.0507)
Layer3	0.8462 (0.0546)	0.7945 (0.1090)	0.8376 (0.0791)	**0.8496** (0.0761)	0.8174 (0.0967)	0.8132 (0.0784)	0.7458 (0.1423)	0.8173 (0.0604)	0.7760 (0.1046)
Layer2	0.7897 (0.0965)	0.8425 (0.0681)	0.7416 (0.1055)	**0.8435** (0.0723)	0.7021 (0.1283)	0.7535 (0.1429)	0.8289 (0.0554)	0.7165 (0.1437)	0.7274 (0.1237)
Layer1	0.7553 (0.0869)	0.7788 (0.0998)	0.7847 (0.1108)	0.7913 (0.1137)	0.7867 (0.0885)	**0.8339** (0.0683)	0.7936 (0.0832)	0.7029 (0.1452)	0.7545 (0.0940)
Emb	0.4842 (0.0549)	0.4935 (0.0511)	0.5516 (0.0642)	0.5034 (0.0703)	**0.5815** (0.0768)	0.5472 (0.0829)	0.5657 (0.0966)	0.5205 (0.1054)	0.5521 (0.0692)