EFFICACY OF DATA-FREE METRICS: ROBUST AND CRITICAL EVIDENCE FROM ROBUST AND CRITICAL LAYERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Data-free methods for analysing and understanding the layers of neural networks have offered many metrics for quantifying notions of "strong" versus "weak" layers, with the promise of increased interpretability. We examine the robustness and predictive power of data-free metrics under randomised control conditions across a wide range of models, datasets and architectures. Contrary to some of the literature, we find strong evidence *against* the efficacy of data-free methods. We show that they are not reparametrisation-invariant even for *robust* layers, that is to say layers that can be reparametrised by re-initialisation or re-randomisation without affecting the accuracy of the model. Moreover, we also show that data-free metrics cannot be used for the arguably simpler tasks of (i) distinguishing between robust layers and critical layers, i.e. layers that cannot be reparametrised without significantly degrading the accuracy of the model, or (ii) predicting if there will be a performance difference between re-initialisation and re-randomisation. Thus, we argue that to understand neural networks, and in particular the difference between 'strong" versus "weak" layers, we must adopt mechanistic and functional approaches, contrary to the traditional Random Matrix Theory perspective.

1 INTRODUCTION

Understanding and interpreting deep learning models is a critical area of research, especially as the prevalence of these models increases in real-world applications. The holy grail of neural network interpretability lies in identifying computationally cheap metrics that can provide insights into the effectiveness of neural networks and their components. Data-free methods typify this endeavour by analysing the properties of the neural network parameters without regard for the data. A key example of data-free methods is Martin & Mahoney (2021), which claims to be able to predict the performance of a neural network without the requirement of test data through the use of Random Matrix Theory to analyse the layer weight matrices. In contrast, data-dependent layer analysis via mechanistic interpretability or functional analysis attempts to quantify how inputs interact at specific layers and use comparative analysis to understand the interaction between model parameters and data (Olah et al., 2020; Klabunde et al., 2025; Nanda et al., 2023).

Zhang et al. (2022) identified an interesting and unexpected phenomenon in neural network layers: some layers within a network are robust, while others are critical. A critical layer is a layer that cannot be re-initialisatised or re-randomised without dramatically affecting the performance of the network. In contrast, a robust layer can be either re-initialisatised or re-randomised without any noticeable effect on performance. Re-initialisation sets the layer back to its initial parameters before training, whilst re-randomisation sets the parameters of the layer to random values by re-sampling from the same distribution used for initialisation. It was observed that in some cases, re-initialisation and re-randomisation can result in significant performance differences for a given layer, with re-initialisation maintaining performance but re-randomisation significantly degrading it (Zhang et al., 2022). In other cases, re-initialisation and re-randomisation of a layer lead to a negligible difference in performance.

We follow in the footsteps of Dinh et al. (2017) which studied another type of metrics, namely metrics for minima flatness and how they (fail to) relate to generalisation. In particular, the authors make the following point: "*Since we are interested in finding a prediction function in a given family*

of functions, no reparametrisation of this family should influence generalisation of any of these functions". In the same spirit, the robustness phenomenon suggests that certain reparametrisations of this family – re-initialisations and re-randomisations of robust layers – should not influence the functional behaviour of any of these functions (as quantified by accuracy). We believe that this provides a strong basis for asking:

- Are data-free metrics reparameterisation-invariant, particularly under re-initialisations or re-randomisations of robust layers?
- Can data-free metrics distinguish robust layers from critical layers?
- Can data-free metrics predict performance difference between re-initialisation and re-randomisation of a layer?

In contrast to Dinh et al. (2017), our approach is purely empirical and our experiments show that across data modalities, architectures and datasets:

- Data-free metrics are not invariant under reparametrisation, even when we restrict our attention to re-initialisations and re-randomisations of robust layers,
- Current data-free analyses have no predictive capacity to identify robust and critical layers in small scale experiments, nor to predict performance difference between re-initialisation and re-randomisation,
- For Large Language Models (LLMs) and ImageNet vision models norm-based metrics have no predictive capacity over the change in performance under re-randomisation.

Consequently, we argue that improving the interpretability of neural networks requires moving beyond the weak predictive power of metrics defined by data-free metrics and power norms.

2 BACKGROUND

Data-free methods of interpretability aim to understand the inner workings of neural networks by studying the properties of the network parameters. Data-free approaches often focus on the matrix norm properties of layer weight matrices to understand learning or improve the performance of neural networks (Yunis et al., 2024; Martin et al., 2021; Sanyal et al., 2020; Feng et al., 2022; Salimans & Kingma, 2016; Bartlett et al., 2017). However, Zhang et al. (2022) showed in their work that matrix norms, such as the Frobenius norm, are too coarse to understand the generalisation properties of neural networks. Martin & Mahoney (2021) use Random Matrix Theory (RMT) to analyse the weight matrices (excluding biases) of neural network layers through training to create a theory of heavy-tailed self-regularisation. With this theory, they construct a set of predominately power-norm metrics related to generalisation that is applied after training to assess layer performance: alpha (α), alpha-weighted ($\hat{\alpha}$), log alpha norm, and MP soft rank Martin & Mahoney (2021). In this work, they identified a value of α between 2 and 6 as a property of a good, well-trained layer, whereas $\alpha > 6$ indicates that a layer is underfitted and $\alpha < 2$ indicates that it is overfitted. Martin et al., Martin et al. (2021) showed a correlation between these metrics and the generalisation performance of pre-trained models in language and computer vision tasks.

The theory of heavy-tailed self-regularisation (Martin & Mahoney, 2021) has been used to provide justification for layer-wise pruning ratios in large language models (Lu et al., 2024), additionally it has been used to explain and understand stages of the grokking phenomenon (Power et al., 2022) namely, pre-grokking, grokking and anti-grokking (Prakash & Martin, 2025). The promise of understanding both redundancy in neural networks and transitions from memorisation to generalisation means that RMT promises a lot with respect to improving the interpretability of deep neural networks, a topic of great importance given their increasing adoption across a range of different data domains.

Since Zhang et al. (2022) established that norm-based methods are ineffective, our work will focus on alpha, alpha-weighted, log alpha norm, MP Soft Rank and Generalized von-Neumann Matrix Entropy (Martin & Mahoney, 2021) as well as Frobenius Norm, Spectral Norm and Stable Rank (Rudelson & Vershynin, 2007) within the critical and robust layer phenomenon to establish whether these metrics are invariant under re-initialisation and re-randomisation of robust layers and if they can disambiguate between (i) robust and critical layers, and (ii) the performance difference between re-initialisation and re-randomisation of a layer.

3 EXPERIMENTAL SETUP

Zhang et al. (2022), showed the robust and critical layer phenomenon across a range of trained architectures, MLPs, VGGs (Simonyan & Zisserman, 2015), ResNets (He et al., 2016), Transformers (Vaswani et al., 2017), Vision Transformers (Dosovitskiy et al., 2021a), MLPMixers (Tolstikhin et al., 2021) across datasets MNIST (LeCun et al., 1998), CIFAR10 (Krizhevsky & Hinton, 2009), ImageNet (Deng et al., 2009) and LM1B(Chelba et al., 2014). We first choose the simplest model (ReLU FCN 5x512), Figure 1, and dataset (MNIST (LeCun et al., 1998)) identified by Zhang et al. (2022) that demonstrates this phenomenon to systematically explore the behaviour of data-free metrics under layer re-initialisations and re-randomisations of a large number of trained models.

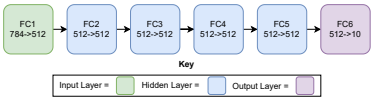


Figure 1: ReLU FCN 5x512 Model Architecture.

An added benefit of the ReLU FCN 5x512 model is that it also offers a clear performance contrast between re-initialisation and re-randomisation. Zhang et al. (2022) showed that residual blocks are robust to re-randomisation and attributed this to the residual layer potentially playing a lesser role in the network and thus having smaller activations than the skip connection. To analyse how effective the data-free metrics are at disambiguating between re-initialisation and re-randomisation, we analyse the correlation between data-free metrics and test accuracy using the Spearman correlation coefficient ρ, the root mean square error (RMSE) of the linear regression and Kendall's tau measure (K-τ). Where ρ and K-τ score of -1 indicates a very strong negative correlation, 0 indicates no correlation, and 1 indicates a very strong positive correlation. We use RMSE and Kendall's Tau measure for this study as they are two of the correlation metrics used in Martin et al. (2021) to highlight the predictive capacity of the data-free metrics.

We trained 100 ReLU FCN 5x512 models, creating 100 initialisations and 100 trained models, to obtain a representative sample of possible initialisations and trained models. The model weights and biases are initialised and re-randomised from the same distribution $\mathcal{U}(-\sqrt{k}, \sqrt{k})$ where k is $\frac{1}{\text{in\_features}}$, e.g. FC1 has $k = \frac{1}{784}$. We record the data-free metric properties of these trained models' layers when they undergo re-initialisation and re-randomisation and observe whether these metrics:

(a) are invariant, particularly for robust layers,

(b) can distinguish critical and robust layers,

(c) can predict the performance difference between re-initialisation and re-randomisation.

This exploration also demonstrates the overall predictive capacity of data-free metrics after training.

Data-Free Metrics. Power Norm based data-free methods analyse a layer weight matrix, W, excluding the bias. A variety of data-free metrics have been developed in the literature to quantify the importance of a layer, we focus on the following metric (Martin & Mahoney, 2021):

- **Alpha (α):** The fitted power law exponent, α, for the empirical spectral density of the correlation matrix $X = W^T W$, such that $p_{emp}(\lambda) \sim \lambda^{-\alpha}$, with λ the eigenvalues of X.

In section 4.2 we additionally explore the following metrics:

- **Alpha Weighted ($\hat{\alpha}$):** $\alpha \log(\lambda_{max})$, where λ_{max} is the max eigenvalue from X Martin & Mahoney (2021).
- **Log Alpha Norm:** $\log(||X||_\alpha^\alpha)$, where $||X||_\alpha^\alpha = \sum_i^M \lambda_i^\alpha$, where M is the rank of W Martin & Mahoney (2021).
- **MP Soft Rank:** is the ratio between the bulk edge of the $p_{emp}(\lambda)$, λ^+, and the max eigenvalue, λ_{max}, $\frac{\lambda^+}{\lambda_{max}}$ Martin & Mahoney (2021).

- **Spectral Norm**: The max singular value of W denoted as $||W||_\infty$.
- **Stable Rank:** The ratio of the squared Frobinues Norm and the squared Spectral Norm, denoted as $\frac{||W||_F^2}{||W||_2^2}$ Rudelson & Vershynin (2007).
- **Generalized von-Neumann Matrix Entropy:** $\frac{-1}{log(M)} \sum_i p_i \log p_i$, where M is the rank of matrix W and p_i is $\frac{\sigma_i^2}{\sum_i(\sigma_i^2)}$ where σ is the singular values of W Martin & Mahoney (2021).

The metrics are collected using the weightwatcher[1] tool.

We extend our analysis of questions (a) and (b) to large-scale pre-trained computer vision models (ResNet34 [2] (Wightman et al., 2021; Wightman, 2019; He et al., 2016) and ViT[3] (Steiner et al., 2021; Dosovitskiy et al., 2021b; Wightman, 2019), both trained on ImageNet(Russakovsky et al., 2015)) and pre-trained large language models (GPT2[4] and GPT2-Large[5] (Radford et al., 2019) evaluated on WikiText103 (Merity et al., 2016)), in Section 5.

4 Results and Discussion For Small Scale Reparametrisations

For clarity and succinctness, we primarily present our results for alpha (α) of Martin & Mahoney (2021) in the body of the paper, however in Section 4.2 we show that these result generalise to additional metrics. In Appendix Section A we present the analysis of the Frobenius Norm.

4.1 Analysis of Alpha Under Re-Initialisation and Re-Randomisation

Table 1: Alpha (α) of the layers in ReLU FCN 5x512 and test accuracy of the model. Mean and $\pm$ 1 SEM (Belia et al., 2005) (Standard Error from the Mean) derived from 100 trained models on MNIST.

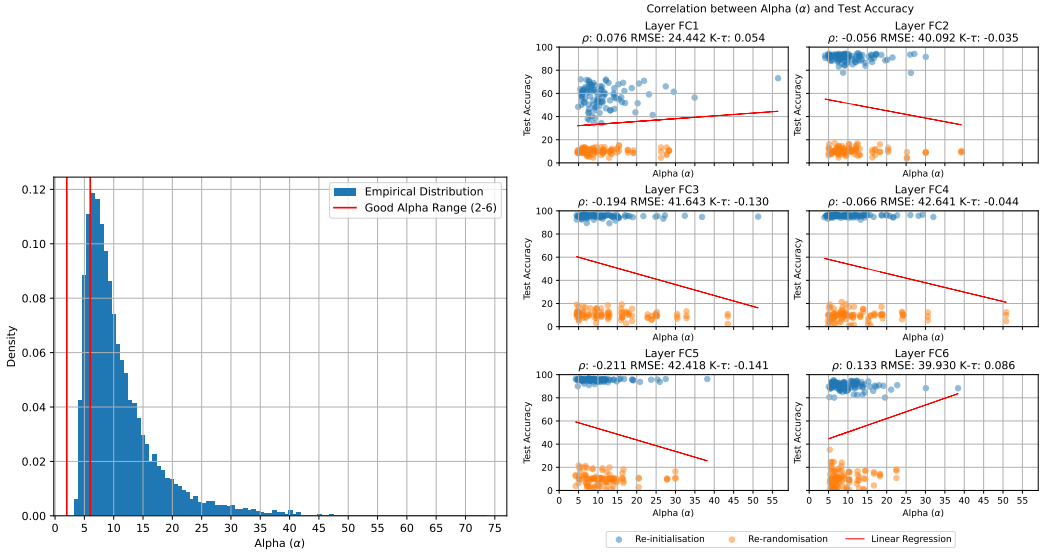
Metric	Layer						Test Accuracy
	FC1	FC2	FC3	FC4	FC5	FC6	
Alpha (α)	4.82 $\pm$ 0.025	4.205 $\pm$ 0.039	4.126 $\pm$ 0.038	4.135 $\pm$ 0.035	4.193 $\pm$ 0.034	3.793 $\pm$ 0.805	96.822 $\pm$ 0.057

Quality of training vs alpha. Data-free metrics such as α aim to identify well-trained layers, with a well-trained layer having a value of α between 2 and 6 Martin & Mahoney (2021). This seems to be borne out by training 100 ReLU FCN 5x512 models to good values of test accuracy, achieving $\alpha \in [2, 6]$ for every layer (see Table 1). Whilst well-trained may imply $\alpha \in [2, 6]$ (but more on this below), a simple experiment shows that the converse does not hold. We plot the empirical distribution of α for a 512x512 fully connected layer in Figure 2 (left), sampled from 10,000 potential initialisations. The resulting distribution shows that an initialised, *untrained* layer of this network, can fall, with a small but non-negligeable probability, within the optimal α value range of 2 and 6.

Next, we perform independent re-initialisations (**blue** in Figure 2) and re-randomisations (**orange**) of each layer for 100 trained ReLU FCN 5x512 networks and record the impact on the networks' test accuracy and alpha values. Figure 2 shows that a non-negligeable proportion of networks whose performance is severely degraded by re-randomisation maintain a good value of α around 5. Conversely, many networks maintain good performance after re-initialisations (particularly of layers FC3-FC5) but with α values significantly outside of $[2, 6]$. Good models can have bad alphas, bad models can have good alphas.

Reparametrisation-invariance of alpha. As is clear from 2, whilst all values of α start in the "good" range $[2, 6]$, reparametrisation in the form of re-initialisation or re-randomisation scatters these values over a very wide range (typically deep into "underfitted" territory)), irrespective of the change

[1] https://weightwatcher.ai
[2] https://huggingface.co/timm/resnet34.tv_in1k
[3] https://huggingface.co/timm/vit_base_patch16_224.augreg_in1k
[4] https://huggingface.co/openai-community/gpt2
[5] https://huggingface.co/openai-community/gpt2-large



(a) (b)

Figure 2: Empirical distribution of α values on a 512x512 fully connected layer, sampled from 10,000 initialisations (**Left**). Layer re-initialisation (**blue**) and re-randomisation (**orange**) test accuracy vs α, ρ is the Spearman correlation coefficient, $RMSE$ is the root mean square error of the linear regression (red line), and K-τ is the Kendall's tau measure, all with respect to the relationship between test accuracy and α values (**Right**).

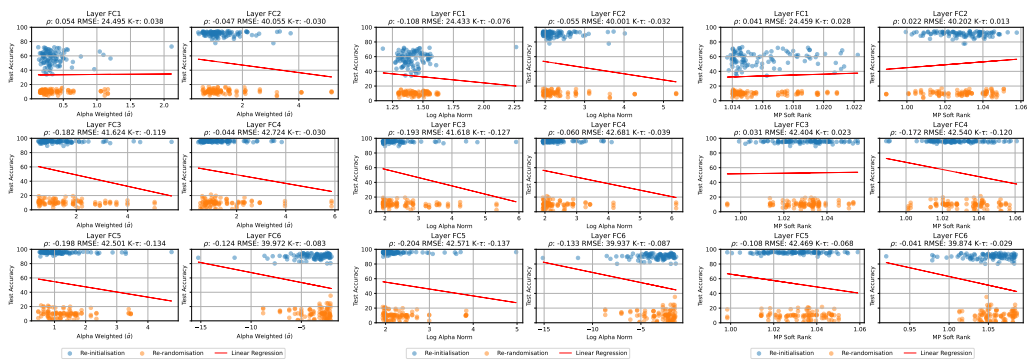
in accuracy incurred by this reparametrisation. Alpha is not invariant under these reparametrisations, even when accuracy is (i.e. on robust layers).

The robust vs critical phenomenon and alpha Figure 2 shows the stark contrast in how layers respond to re-initialisation and re-randomisation. For example, we only see a large drop in test accuracy when applying re-initialisation to the Layer FC1, re-initialising other layers leaves performance almost unchanged. From this perspective, FC1 is a critical layer whilst FC2-FC6 are robust (to re-initialisation). In this experiment α cannot distinguish between these behaviours, the range of values taken by α shows no noticeable difference between critical and robust layers.

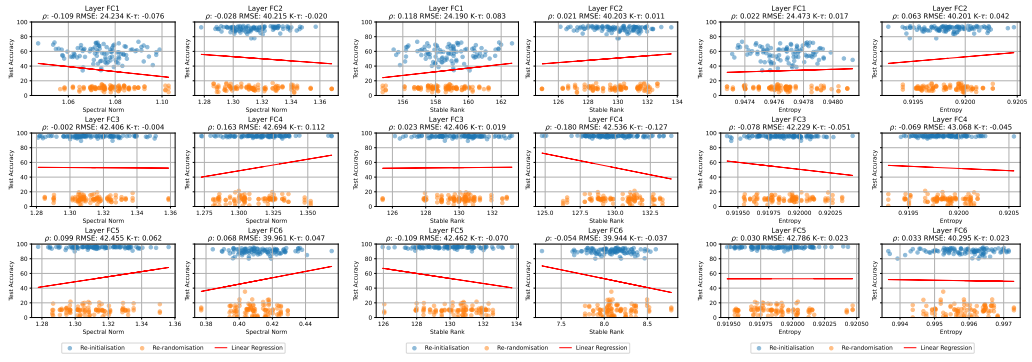
The re-initialisation vs re-randomisation phenomenon and alpha. We observe different results when re-randomisation is applied. We find that re-randomising any layer degrades accuracy to circa random accuracy on the test set, in other words none of the layers are robust to re-randomisation. Surprisingly, this is not reflected in the corresponding α values of these two conditions, as the distribution of α values is relatively similar for each layer and each condition. If α had predictive power we would expect to observe a negative correlation between re-initialisation (corresponding to "good" alphas) to re-randomisation (corresponding to "bad" alphas). However, there is almost no difference between the α values of re-initialisation and re-randomisation, with a mean Spearman correlation coefficient and Kendall's tau measure across layers of -0.053 and -0.035, respectively. Thus α does not distinguish between these behaviours either. These findings extend to other data-free metrics in Appendix Section A and Section 4.2.

4.2 Generalisation of Reparametrisation Invariance to Data Free Metrics

We extend our analysis to other data-free metrics that have been defined in literature, we argue that given the lack of predictive power of the α metric in the previous section, it is important to verify across other representative measures that data-free metrics broadly cannot disambiguate between critical and robust layers, between re-initialisation and re-randomisation. When we conduct a similar analysis across six other metrics, Alpha Weighted, Log Alpha Norm, MP Soft Rank, Spectral Norm,



(a) Alpha Weighted (b) Log Alpha Norm (c) MP Soft Rank



(d) Spectral Norm (e) Stable Rank (f) Generalized von-Neumann Matrix Entropy

Figure 3: Layer re-initialisation (**blue**) and re-randomisation (**orange**) test accuracy vs the respective metric. ρ is the Spearman correlation coefficient, $RSME$ is the root mean square error of the linear regression (red line), and K-τ is the Kendall's tau measure, all with respect to the relationship between test accuracy and the metric.

Stable Rank and Generalized von-Neumann Matrix Entropy we find, as shown in Figure 3, that we can draw exactly the same conclusion as with α.

4.3 Comparing To Data-Based Metrics

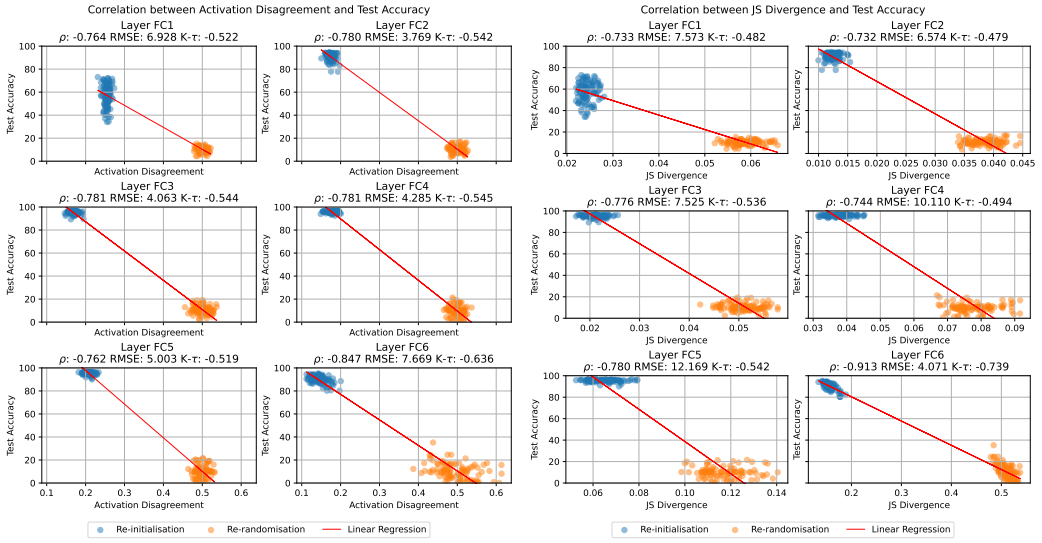
To compare data-free and data-based metrics we explore if two widely used data-based metrics can distinguish the output of a re-initialised layer from that of a re-randomised layer. The outputs of the original and modified layers are represented as $\mathcal{O}_l$ and $\mathcal{M}_l$, respectively. The model layer outputs are collected by passing through the test dataset, $\mathcal{D}_{test}$. The metrics explored are defined as follows:

- **Activation Disagreement:** The mean percentage of times that the neurons from $\mathcal{O}_l$ and $\mathcal{M}_l$ fail to agree to activate on $\mathcal{D}_{test}$.
- **Jensen–Shannon (JS) Divergence:** The mean weighted average of the Kullback–Leibler (KL) Divergence of the softmax($\mathcal{O}_l$ on $\mathcal{D}_{test}$) compared to the softmax($\mathcal{M}_l$ on $\mathcal{D}_{test}$) (Lin, 1991).

When considering the results of the activation disagreement and JS divergence between the original model layer and each model with re-initialised and re-randomised layers in Figure 4, it becomes clear that there is a stark difference in similarity between the re-initialised and re-randomised layers compared to their original layer. The figure demonstrates that the difference between re-initialisation and re-randomisation can be explained by the amount the layer disagrees on activation compared to the original non-modified layer. It is evident that when there is less disagreement (re-initialisation), the model can maintain accuracy, while a lot of disagreement (re-randomisation) leads to a drop

in accuracy. This observation is further supported by a mean Spearman correlation coefficient and Kendall's tau measure across layers of -0.776 and -0.5375, respectively for Activation Disagreement, indicating a negative relationship between activation disagreement and test accuracy.

We have produced these results to show that there exist metrics that can disambiguate between re-initialisation and re-randomisation behaviours which we believe deserve increased focus over data-free metrics (whether they can distinguish critical from robust is, in this instance, less clear).



(a) Activation Disagreement (b) JS Divergence

Figure 4: Layer re-initialisation (**blue**) and re-randomisation (**orange**) test accuracy vs the respective metric. ρ is the Spearman correlation coefficient, $RSME$ is the root mean square error of the linear regression (red line), and K-τ is the Kendall's tau measure, all with respect to the relationship between test accuracy and the metric.

5 Results and Discussion For Large Scale Reparametrisations

We further explore how predictive the α metric can be when predicting performance of layers in large scale models that have been trained for high-complexity tasks using the re-randomisation setup. Overall we find, consistent with our small scale experiments, that the α metric has no predictive capacity under this condition, therefore questioning its utility in predicting the performance of layer criticality and "trainedness". For these results we employ pre-trained open source models, therefore, we do not have access to the starting initialisation and, as a result, we only focus on the case of re-randomisation to test the predictive capacity of the α metric at this scale. We show the relationship reported in this section is consistent with other data-free metrics in this experimental set up in Appendix Sections A.2, A.3, A.4 and A.5.

5.1 ImageNet Scale

We explore the relationship of a layer's α value with the corresponding performance of the model on ImageNet with pre-trained ResNet34 and ViT models. The respective performance and model details are shown in Table 2; from this table it can be observed that these models are well trained for the task of ImageNet. In line with the Random Matrix Theory perspective, the competitive performance of these models is aligned with a mean α value across layers within the optimal range, representing two well-fitted networks.

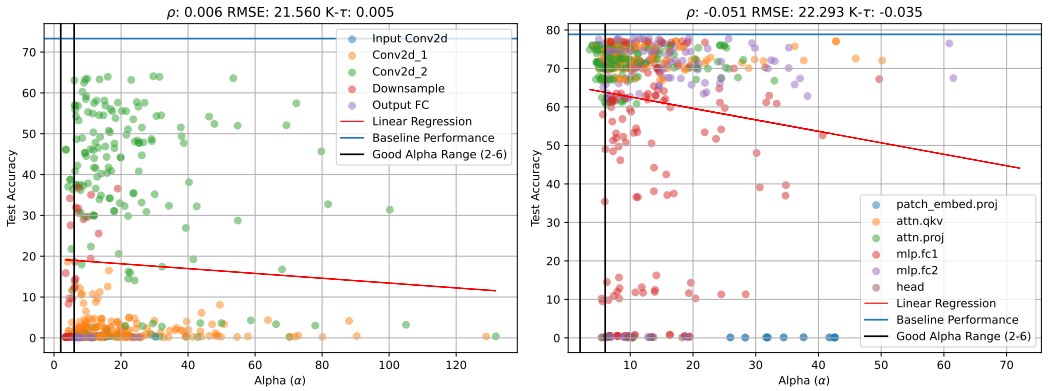
We employ both the pre-trained ResNet34 and ViT as base models for our re-randomisation experiment. To conduct this experiment we randomise a layer independently and then record the corresponding test accuracy and α value post re-randomisation. Given that different implicit biases

respond uniquely to re-randomisation we believe that our selection of models is representative for the core architectural differences, allowing a robust analysis of the impacts of re-randomisation and α metrics at scale. We repeat our re-randomisation process 10 times per layer to capture stochastic variation in randomisations. In both ResNet34 and ViT, Figure 5, we find no relationship between α and test accuracy, as found in the small scale experiment. The mean Spearman correlation coefficients are 0.006 and -0.051, and Kendall's tau values are 0.005 and -0.035, respectively.

Table 2: ResNet34 and ViT performance on ImageNet and mean layer Alpha (α) value $\pm$ 1 SEM (Belia et al., 2005) (standard error from the mean).

Model	Number of Parameters	Loss	Accuracy	Mean Layer Alpha Value
ResNet34	21.8M	1.071	73.302	3.380 ± 0.195
ViT	86.6M	0.825	78.852	4.711 ± 0.271

As previously reported in the small-scale experiments, layers in vision models can be re-randomised to α values outside of the desired range but continue to retain accuracy. While in the case of the ResNet34 we see that the model is not so robust to re-randomisation, having a lower accuracy than the baseline, there is a large proportion of layers with optimal α values that result in a model with essentially random accuracy. The ViT is more robust to layer re-randomisation and show the inverse case where many layers are significantly outside of the optimal α range but display a high preservation of the baseline accuracy. As a result, for large scale vision models these results confirm that α is not invariant under reparametrisation by re-randomisation, even on robust layers, and has very limited predictive capacity over post re-randomisation performance.



(a) ResNet34 (b) ViT

Figure 5: ResNet34 and ViT layer re-randomisation Test Accuracy vs Alpha (α) on ImageNet. ρ is the Spearman correlation coefficient, $RMSE$ is the root mean square error of the linear regression (red line), and K-τ is the Kendall's tau measure, all with respect to the relationship between the performance and α values.

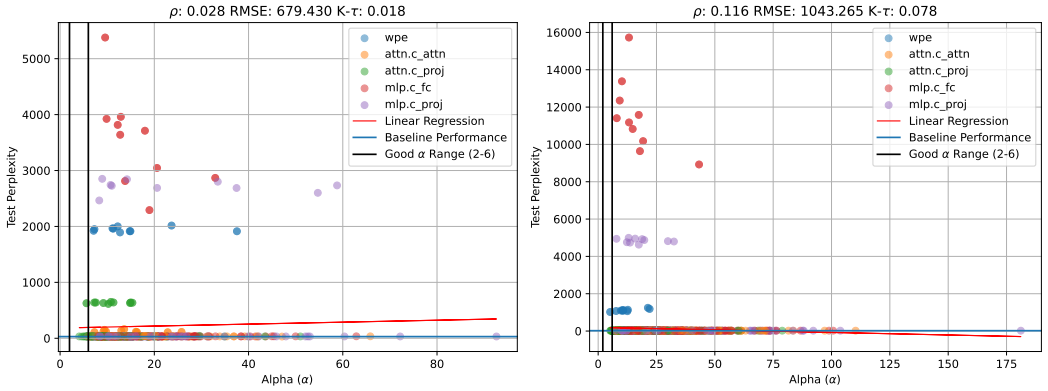
5.2 Large Language Models

Given our negative results for the predictive capacity of the α metric in the vision domain under re-randomisation, it is important to verify the generality of our findings under a different modality. Therefore, we have extended the exploration of the relationship of a layers α value and the corresponding performance of the model to language models on WikiText103 with pre-trained GPT2 and GPT2-Large, the respective performance and model details is shown in Table 3. These models represent an increased scale of our experiments as they have hundreds of millions of parameters. Both models have competitive performance with low loss and perplexity on the WikiText-103 datasets and, in-line with Random Matrix Theory have mean layer α values that are within the optimal range.

Table 3: GPT and GPT-Large Test performance on WikiText-103 and mean layer Alpha (α) value $\pm$ 1 SEM (Belia et al., 2005) (standard error from the mean).

Model	Number of Parameters	Loss	Perplexity	Mean Layer Alpha Value
GPT	127M	3.399	29.941	3.865 ± 0.136
GPT-Large	774M	2.967	19.436	4.013 ± 0.091

We use the pre-trained models as the base model and then we randomise a layer and record its α value and the corresponding test perplexity and loss of the model. We do this 10 times for each layer to obtain a representative set of random layers. For both GPT2 and GPT2-Large, Figure 6 we observe that their is no relationship between the performance of the model and a layer's α value, with a mean Spearman correlation coefficient of 0.028, 0.116 and Kendall's tau measure of 0.018 and 0.078 for GPT2 and GPT2-Large receptively. We also find that randomising layers in the GPT2 and GPT2-Large model often has a negligible affect on the performance of the model irrespective of the α value associated with the layer. Furthermore, when re-randomising a layer and achieving negligible performance degradation we can observe that the bulk of these layers are far outside the optimal range provided by the α metric. Again, we see that α is not invariant under reparametrisation by re-randomisation, even on robust layers, and has no predictive power over performance.



(a) GPT2 (b) GPT2-Large

Figure 6: GPT2 amd GPT2-Large layer re-randomisation Test Perplexity vs Alpha (α) on WikiText103. ρ is the Spearman correlation coefficient, $RMSE$ is the root mean square error of the linear regression (red line), and K-τ is the Kendall's tau measure, all with respect to the relationship between the performance and α values.

6 Conclusion

In this work we concretely showed that data-free metrics cannot explain the robust vs critical layer phenomenon, nor the re-initialisation vs re-randomisation phenomenon. Based on experiments covering a wide range of these metrics and a wide range of models, we argue that they have little to no predictive capacity over these important performance-related layer properties. We highlighted how data-free metrics can be described as non-reparameterisation invariant even for robust layers for which accuracy is (to some degree) reparameterisation invariant. Our results scaling from MNIST to ImageNet and Large Language Models confirm the generality of the robustness of our findings.

As a consequence, our results advocate for a reappraisal of the way that we approach interpretability of neural networks. Instead of using metrics which lack predictive capacity, we argue that there is a requirement for an in-depth exploration of data-free methods that can suitably disambiguate between the robust and critical layer phenomena. Or rather, a focus on methods that consider more than just weight distributions of models, which we show can be arbitrarily reparametrised without impacting the performance, and instead seek metrics which further understand the more nuanced interplay between between weights and data.

REFERENCES

Peter L Bartlett, Dylan J Foster, and Matus J Telgarsky. Spectrally-normalized margin bounds for neural networks. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://papers.nips.cc/paper_files/paper/2017/hash/b22b257ad0519d4500539da3c8bcf4dd-Abstract.html.

Sarah Belia, Fiona Fidler, Jennifer Williams, and Geoff Cumming. Researchers misunderstand confidence intervals and standard error bars. *Psychological methods*, 10(4):389, 2005. URL https://psycnet.apa.org/record/2005-16136-002.

Ciprian Chelba, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, Phillipp Koehn, and Tony Robinson. One billion word benchmark for measuring progress in statistical language modeling, 2014. URL https://arxiv.org/abs/1312.3005.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009. URL https://ieeexplore.ieee.org/document/5206848.

Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp minima can generalize for deep nets. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1019–1028. PMLR, 2017. URL https://proceedings.mlr.press/v70/dinh17b.html.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021a. URL https://openreview.net/forum?id=YicbFdNTTy.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021b.

Ruili Feng, Kecheng Zheng, Yukun Huang, Deli Zhao, Michael Jordan, and Zheng-Jun Zha. Rank diminishing in deep neural networks. In *Advances in Neural Information Processing Systems*, volume 35, pp. 33054–33065, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/d5cd70b708f726737e2ebace18c3f71b-Abstract-Conference.html.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. URL https://www.cv-foundation.org/openaccess/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html.

Max Klabunde, Tobias Schumacher, Markus Strohmaier, and Florian Lemmerich. Similarity of neural network models: A survey of functional and representational measures. *ACM Computing Surveys*, 57(9):1–52, 2025. URL https://doi.org/10.1145/3728458.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. 2009. URL http://www.cs.utoronto.ca/~kriz/learning-features-2009-TR.pdf.

Yann LeCun, Corinna Cortes, and CJ Burges. Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2, 1998.

J. Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, 1991. doi: 10.1109/18.61115.

Haiquan Lu, Yefan Zhou, Shiwei Liu, Zhangyang Wang, Michael W. Mahoney, and Yaoqing Yang. Alphapruning: Using heavy-tailed self regularization theory for improved layer-wise pruning of large language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=fHq4x2YXVv.

Charles H. Martin and Michael W. Mahoney. Implicit self-regularization in deep neural networks: Evidence from random matrix theory and implications for learning. *Journal of Machine Learning Research*, 22(165):1–73, 2021. URL http://jmlr.org/papers/v22/20-410.html.

Charles H Martin, Tongsu Peng, and Michael W Mahoney. Predicting trends in the quality of state-of-the-art neural networks without access to training or testing data. *Nature Communications*, 12(1): 4122, 2021. URL https://www.nature.com/articles/s41467-021-24025-8.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models, 2016.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=9XFSbDPmdW.

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 2020. URL https://distill.pub/2020/circuits/zoom-in.

Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. Grokking: Generalization beyond overfitting on small algorithmic datasets, 2022. URL https://arxiv.org/abs/2201.02177.

Hari Kishan Prakash and Charles H Martin. Grokking and generalization collapse: Insights from htsr theory. In *High-dimensional Learning Dynamics 2025*, 2025.

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

Mark Rudelson and Roman Vershynin. Sampling from large matrices: An approach through geometric functional analysis. *J. ACM*, 54(4):21–es, 2007. ISSN 0004-5411. URL https://doi.org/10.1145/1255443.1255449.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115 (3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

Tim Salimans and Durk P Kingma. Weight normalization: A simple reparameterization to accelerate training of deep neural networks. In *Advances in Neural Information Processing Systems*, volume 29, 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/hash/ed265bc903a5a097f61d3ec064d96d2e-Abstract.html.

Amartya Sanyal, Philip H. Torr, and Puneet K. Dokania. Stable rank normalization for improved generalization in neural networks and gans. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=H1enKkrFDB.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2015. URL https://arxiv.org/abs/1409.1556.

Andreas Steiner, Alexander Kolesnikov, , Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. How to train your vit? data, augmentation, and regularization in vision transformers. *arXiv preprint arXiv:2106.10270*, 2021.

Ilya O Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, Mario Lucic, and Alexey Dosovitskiy. Mlp-mixer: An all-mlp architecture for vision. In *Advances in Neural Information Processing Systems*, volume 34, pp. 24261–24272, 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/cba0a4ee5ccd02fda0fe3f9a3e7b89fe-Paper.pdf.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Ross Wightman. Pytorch image models. https://github.com/huggingface/pytorch-image-models, 2019.

Ross Wightman, Hugo Touvron, and Herve Jegou. Resnet strikes back: An improved training procedure in timm. In *NeurIPS 2021 Workshop on ImageNet: Past, Present, and Future*, 2021.

David Yunis, Kumar Kshitij Patel, Samuel Wheeler, Pedro Savarese, Gal Vardi, Karen Livescu, Michael Maire, and Matthew R. Walter. Approaching deep learning through the spectral dynamics of weights, 2024. URL https://arxiv.org/abs/2408.11804.

Chiyuan Zhang, Samy Bengio, and Yoram Singer. Are all layers created equal? *Journal of Machine Learning Research*, 23(67):1–28, 2022. URL http://jmlr.org/papers/v23/20-069.html.

A Further Analysis on Data-Free Metrics

In this section we extend our analysis to the following data free metrics in our existing experimental setup. W represents the weight matrix of the layer and X is for the empirical spectral density of the correlation matrix, $X = W^T W$, such that $p_{emp}(\lambda) \sim \lambda^{-\alpha}$, where λ are the eigenvalues of X.

- **Alpha Weighted ($\hat{\alpha}$):** $\alpha \log(\lambda_{max})$, where λ_{max} is the max eigenvalue from X Martin & Mahoney (2021).
- **Log Alpha Norm:** $\log(||X||_\alpha^\alpha)$, where $||X||_\alpha^\alpha = \sum_i^M \lambda_i^\alpha$, where M is the rank of W Martin & Mahoney (2021).
- **MP Soft Rank:** is the ratio between the bulk edge of the $p_{emp}(\lambda)$, λ^+, and the max eigenvalue, λ_{max}, $\frac{\lambda^+}{\lambda_{max}}$ Martin & Mahoney (2021).
- **Frobenius Norm**: The sum of the singular values of W denoted as $||W||_F$.
- **Spectral Norm**: The max singular value of W denoted as $||W||_\infty$.
- **Stable Rank:** The ratio of the squared Frobinues Norm and the squared Spectral Norm, denoted as $\frac{||W||_F^2}{||W||_2^2}$ Rudelson & Vershynin (2007).
- **Generalized von-Neumann Matrix Entropy:** $\frac{-1}{log(M)} \sum_i p_i \log p_i$, where M is the rank of matrix W and p_i is $\frac{\sigma_i^2}{\sum_i(\sigma_i^2)}$ where σ is the singular values of W Martin & Mahoney (2021).

Each metric can be found below with the appropriate subsection that corresponds to our analysis of these data-free metrics.

The correlation between the Frobenius Norm and the associated test accuracy when layers undergo re-initialisation (blue) or re-randomisation is shown in Figure 7. The Frobenius Norm observes approximately zero correlation between the metric values and the test accuracy, highlighting the same findings as in the body of the paper.

A.1 Frobenius Norm

Norm-based metrics were originally shown to be too coarse a metric to measure the generalisability of the neural networks in Zhang et al. (2022). Figure 7 strengthens these findings, highlighting that there is essentially no correlation between the Frobenius Norm of a layer and the test accuracy of a model.

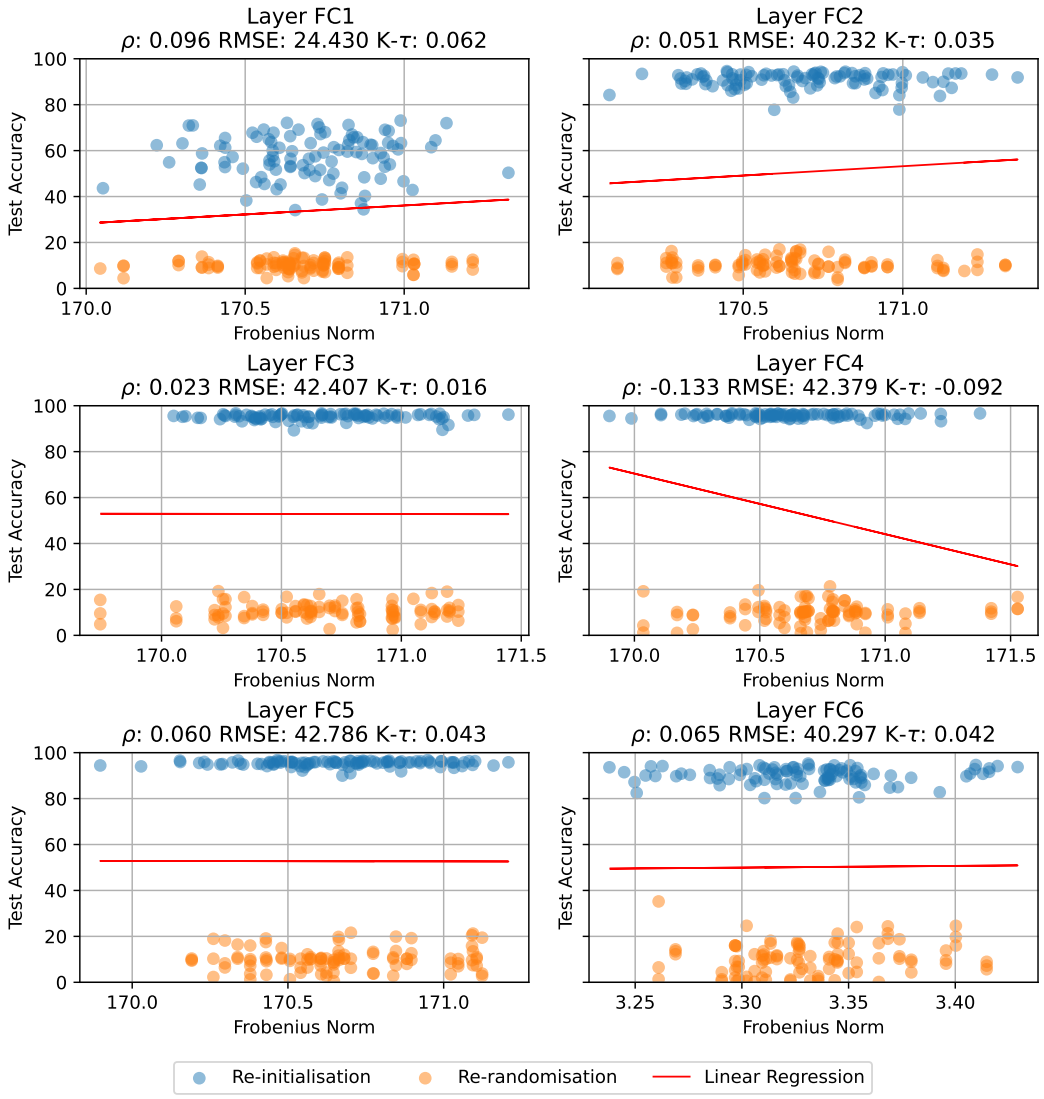
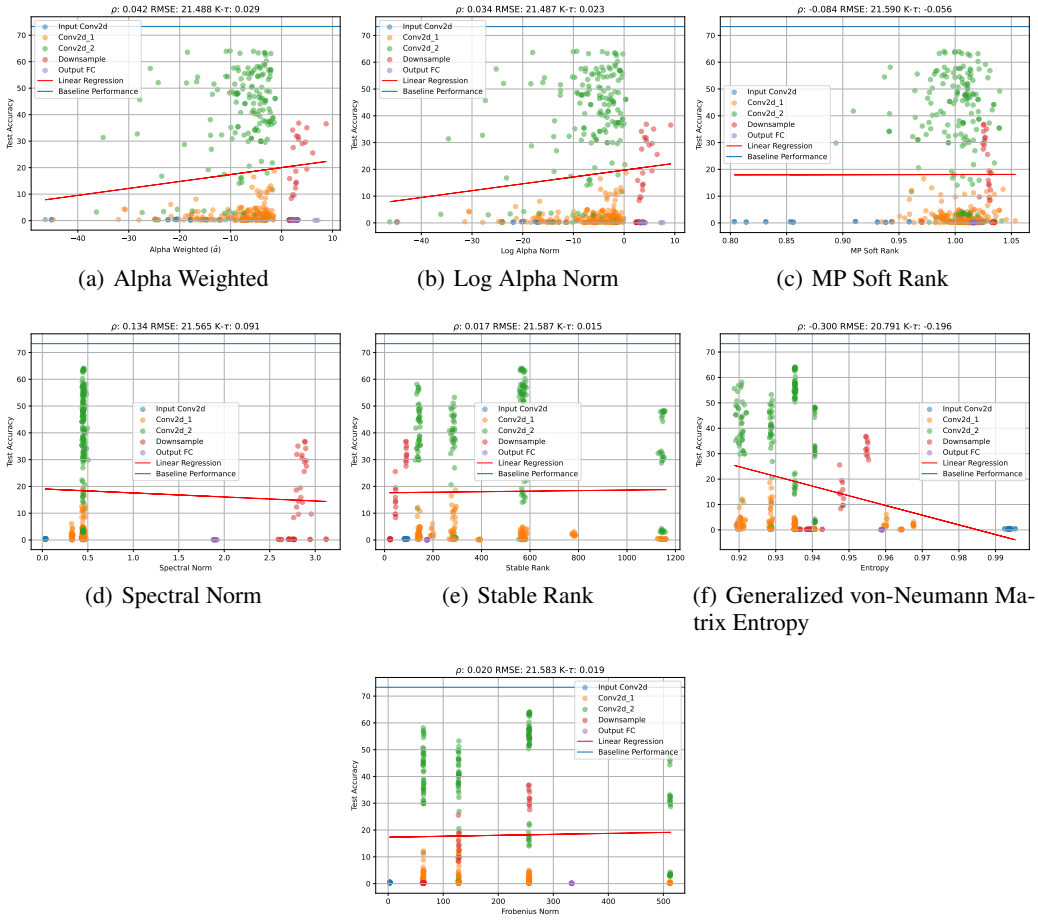


Figure 7: Layer re-initialisation (**blue**) and re-randomisation (**orange**) test accuracy vs Frobenius Norm. ρ is the Spearman correlation coefficient, $RSME$ is the root mean square error of the linear regression (red line), and K-τ is the Kendall's tau measure, all with respect to the relationship between test accuracy and alpha values.

A.2 RESNET34 ON IMAGENET



(a) Alpha Weighted (b) Log Alpha Norm (c) MP Soft Rank

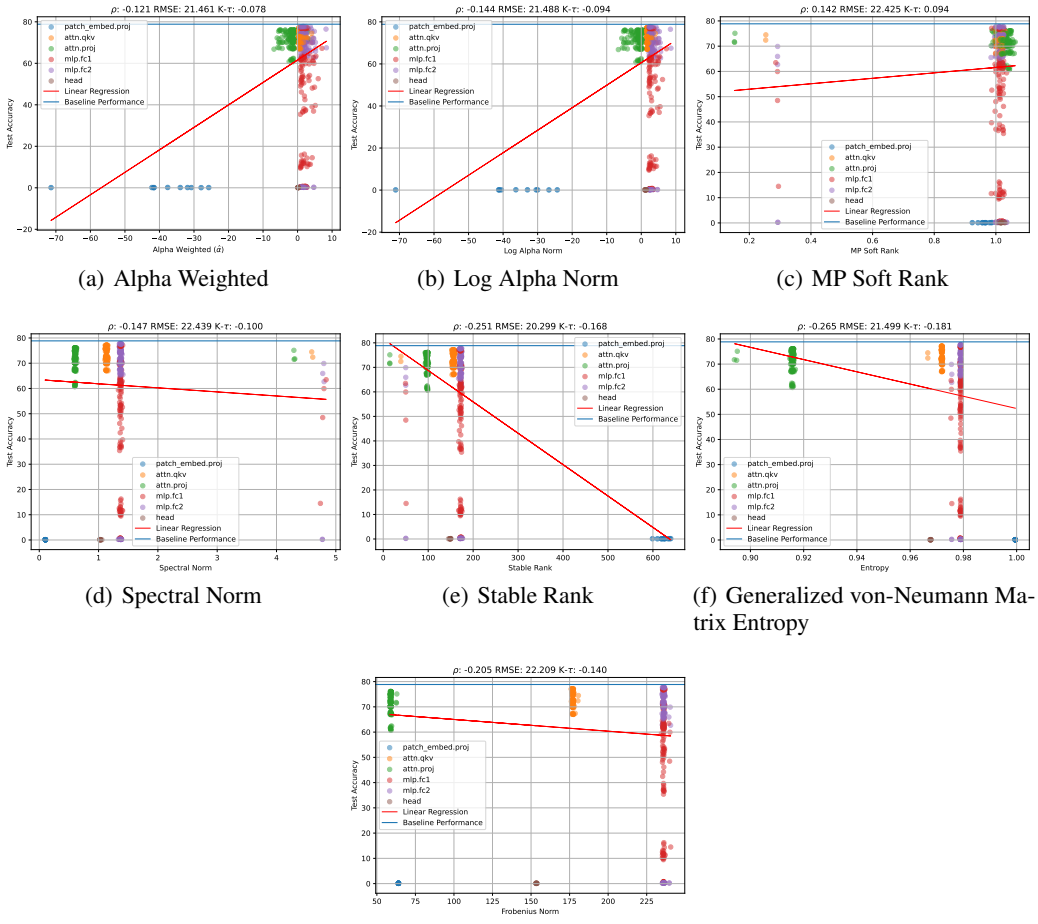
(d) Spectral Norm (e) Stable Rank (f) Generalized von-Neumann Matrix Entropy

(g) Frobenius Norm

Figure 8: ResNet 34: Layer re-randomisation test accuracy vs the respective metric. ρ is the Spearman correlation coefficient, $RSME$ is the root mean square error of the linear regression (red line), and K-τ is the Kendall's tau measure, all with respect to the relationship between test accuracy and the metric.

In Figure 8, we observe the same trend as found with α in the main body of the paper. While some correlations are higher than circa 0, i.e. (f) we can clearly observe that there is strong overlap between the values and the accuracy and that is induced by the different layer types explored.

A.3 VIT ON IMAGENET



(a) Alpha Weighted (b) Log Alpha Norm (c) MP Soft Rank

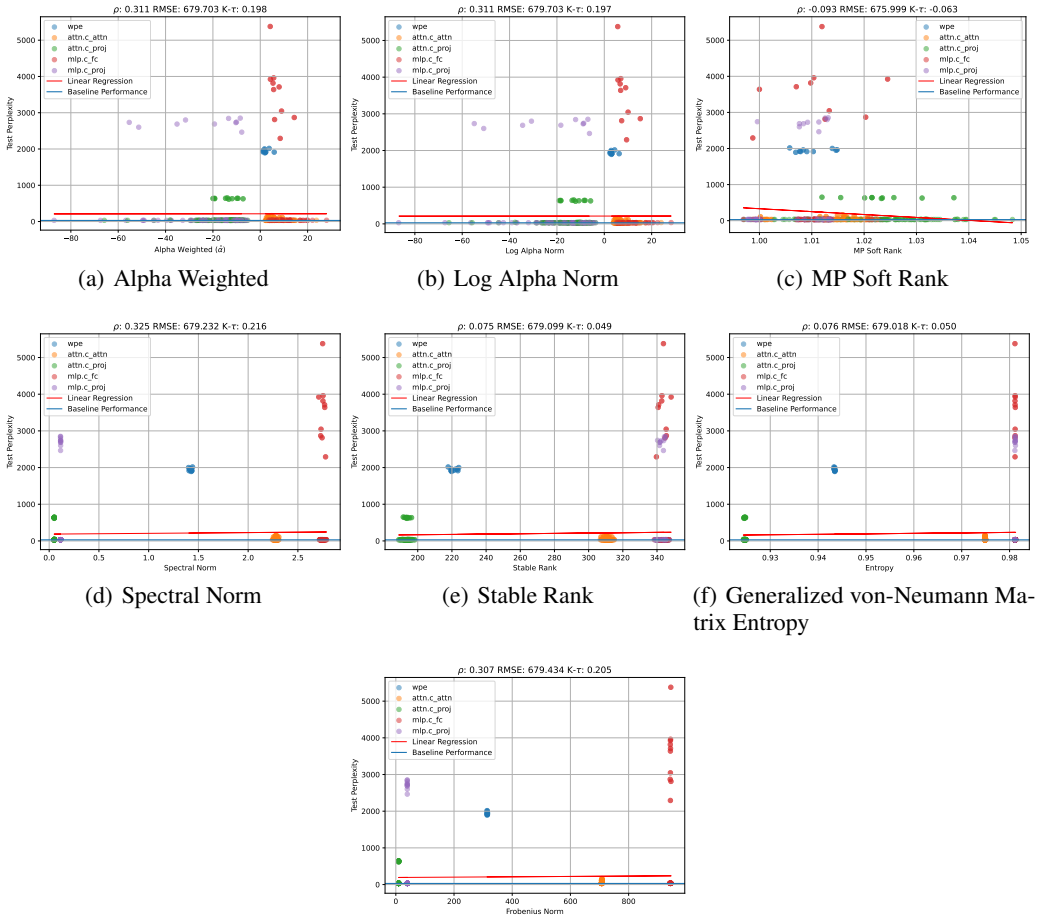
(d) Spectral Norm (e) Stable Rank (f) Generalized von-Neumann Matrix Entropy

(g) Frobenius Norm

Figure 9: ViT: Layer re-randomisation test accuracy vs the respective metric. ρ is the Spearman correlation coefficient, $RSME$ is the root mean square error of the linear regression (red line), and K-τ is the Kendall's tau measure, all with respect to the relationship between test accuracy and the metric.

In Figure 9, we observe the same trend as found with α in the main body of the paper. While some correlations are higher than circa 0, i.e. (f) we can clearly observe that there is strong overlap between the values and the accuracy and that is induced by the different layer types explored.

A.4 GPT2 ON WIKITEXT103



(a) Alpha Weighted (b) Log Alpha Norm (c) MP Soft Rank

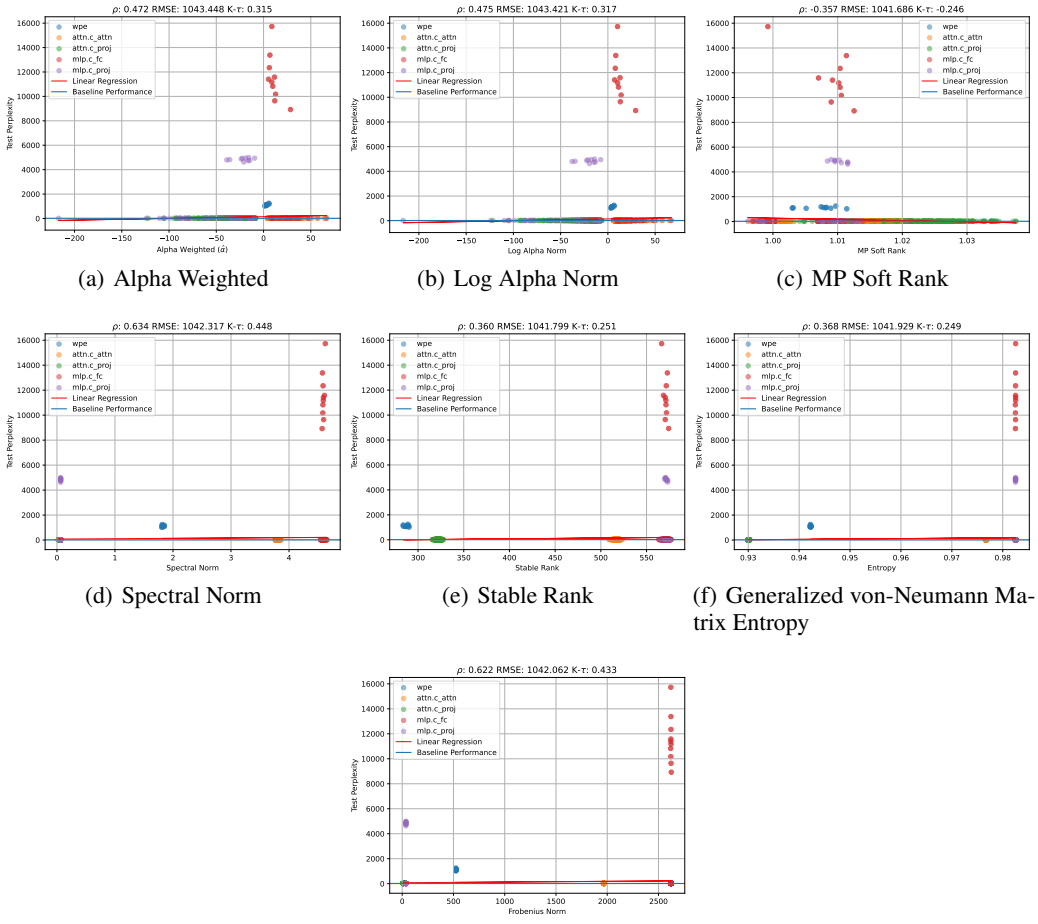
(d) Spectral Norm (e) Stable Rank (f) Generalized von-Neumann Matrix Entropy

(g) Frobenius Norm

Figure 10: GPT2: Layer re-randomisation test accuracy vs the respective metric. ρ is the Spearman correlation coefficient, $RSME$ is the root mean square error of the linear regression (red line), and K-τ is the Kendall's tau measure, all with respect to the relationship between test accuracy and the metric.

In Figure 10, we observe the same trend as found with α in the main body of the paper, across all metrics explored.

A.5 GPT2-LARGE ON WIKITEXT103



(a) Alpha Weighted (b) Log Alpha Norm (c) MP Soft Rank

(d) Spectral Norm (e) Stable Rank (f) Generalized von-Neumann Matrix Entropy

(g) Frobenius Norm

Figure 11: GPT2-Large: Layer re-randomisation test accuracy vs the respective metric. ρ is the Spearman correlation coefficient, $RSME$ is the root mean square error of the linear regression (red line), and K-τ is the Kendall's tau measure, all with respect to the relationship between test accuracy and the metric.

In Figure 11, we observe the same trend that there is little to no correlation between the metric and the test accuracy of the model as found with α in the main body of the paper, across all metrics explored.