Understanding and Minimising Outlier Features in Neural Network Training

Bobby He [1] **Lorenzo Noci** [1] **Daniele Paliotta** [2] **Imanol Schlag** [1 3] **Thomas Hofmann** [1]

Abstract

Outlier Features (OF) are neurons whose activation magnitudes significantly exceed the average over a neural network's (NN) width. They are well known to emerge during standard transformer training and have the undesirable effect of hindering quantisation in afflicted models. Despite their practical importance, little is known behind *why OFs emerge during training*, nor *how one can minimise them*. Our work focuses on the above questions, first identifying several quantitative metrics, such as the kurtosis over neuron activation norms, to measure OFs. With these metrics, we study how architectural and optimisation choices influence OFs, and provide practical insights to minimise OFs during training. As highlights, we emphasise the importance of controlling signal propagation throughout training and propose the *Outlier Protected* transformer block, which removes standard Pre-Norm layers to mitigate OFs, without loss of convergence speed or training stability. Overall, our findings shed new light on our understanding of, our ability to prevent, and the complexity of this important facet in NN training dynamics.

1. Introduction

Despite their widespread use, our understanding of deep neural networks (NNs) and their training dynamics is very much incomplete. This, in part, reflects the complexity of traversing high-dimensional non-convex loss landscapes but is also symptomatic of the myriad design choices, such as NN architecture and optimiser hyperparameters, that a practitioner must take before training. While standard choices of architecture and optimiser exist, it is often unclear how these choices affect performance or the emergence of various empirically observed phenomena during NN training.

Outlier Features (OF) are one such NN training phenomenon. Intuitively, OFs are neurons whose activation magnitudes are significantly larger than the average in the same NN layer, i.e. across NN width (Kovaleva et al., 2021; Timkey & van Schijndel, 2021; Bondarenko et al., 2021). They have been widely observed in pre-trained transformer models (Devlin et al., 2018; Radford et al., 2019; Zhang et al., 2022b), as we verify in Fig 1, and are of practical interest because their existence hinders quantisation (Bondarenko et al., 2021; Wei et al., 2022; Dettmers et al., 2022; Zeng et al., 2023; Wortsman et al., 2023; Ashkboos et al., 2024; Nrusimha et al., 2024). In particular, OFs cause large dynamic ranges in activations across NN width, which lead to high quantisation errors in low-precision matrix multiplications. As a result, Outlier Feature Emergence (OFE) hinders low-precision training and inference, and minimising OFE could yield significant potential efficiency gains.

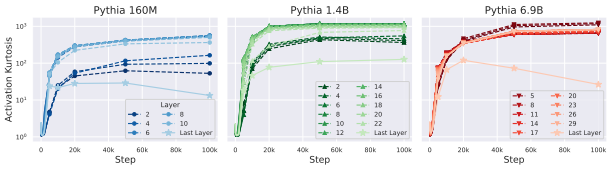


Figure 1. Outlier Features emerge in open-source transformers (Biderman et al., 2023) during training, measured by our kurtosis, Eq (1). Our work studies factors that influence OF emergence.

In this paper, we tackle OFE from two related angles: by (1) proposing interventions to minimise OFE without affecting model convergence or training stability, using insights motivated through (2) enhancing our understanding of why OFs appear during training. We argue that it is important to first understand why OFs appear during standard NN training dynamics in order to identify which design choices influence OFE, and how. Though progress has been made (Kovaleva et al., 2021; Puccetti et al., 2022; Wortsman et al., 2023; Bondarenko et al., 2023), the mechanisms behind OFE remain largely unknown.

Our contributions Overall, we show that OFE can be mitigated relative to standard practices, and highlight key design choices to do so. In Sec 3, we study the role of normalisation layers for OFE, and find that existing hypotheses do not fully capture the OF phenomenon. We proceed to show that removing normalisation through our *Outlier Protected* transformer block minimises OFs, without loss of convergence speed. In Sec 4, we consolidate our findings by identifying signal propagation as a key object that predicts OFs during training, and that choices that improve signal propagation during training also minimise OFE. In interest of space, in Apps A and E we discuss additional related

[1]Department of Computer Science, ETH Zürich [2]Machine Learning Group, University of Geneva [3]ETH AI Center. Correspondence to: Bobby He <bobby.he@inf.ethz.ch>.

Work presented at TF2M workshop at ICML 2024, Vienna, Austria. PMLR 235, 2024.

work and optimisation choices respectively.

2. Problem Setting

Consider an activation matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ obtained from some neural network layer, where n is the number of batch inputs/sequence positions, and d is the number of neurons across NN width. In a typical NN layer, we matrix multiply $\mathbf{X}$ by a weight matrix $\mathbf{W} \in \mathbb{R}^{d \times d}$ to give $\mathbf{XW} \in \mathbb{R}^{n \times d}$, with $(\alpha, j)^{\text{th}}$ element: $\sum_{k=1}^{d} \mathbf{X}_{\alpha,k}\mathbf{W}_{k,j}$. This fundamental operation is central to NN computation and can be seen as a sum over d terms, one for each neuron.

Several works have established that if the magnitudes of the summands $\{\mathbf{X}_{\alpha,k}\mathbf{W}_{k,j}\}_{k=1}^{d}$ have large variance, it becomes difficult to compute their sum in low precision, thereby precluding potential efficiency gains from "vector-wise" quantised training or inference (though significant progress has been made on the latter, (Dettmers et al., 2022; Xiao et al., 2023; Ashkboos et al., 2024)). These works have shown that trained transformer (Vaswani et al., 2017) models possess such a deficiency, which is attributed to the existence of *Outlier Features* (OFs) whose activations are much larger in magnitude compared to the other neurons.

Measuring OFs We use two metrics to measure OFs in $\mathbf{X}$:

1. **Kurtosis of neuron activation RMS**: Let $\mathbf{s} \in \mathbb{R}^d$, such that $\mathbf{s}_j = \sqrt{\frac{1}{n}\sum_{\alpha=1}^{n} \mathbf{X}_{\alpha,j}^2}$, be the vector of root mean-squared activations across inputs.[1] Then, let $\text{Kurt}(\mathbf{X})$ be the ratio of the fourth moment m_4 to the squared second moment m_2 over $\mathbf{s}$:

$$\text{Kurt}(\mathbf{X}) = \frac{m_4(\mathbf{X})}{m_2(\mathbf{X})^2} \overset{\text{def}}{=} \frac{\frac{1}{d}\sum_{j=1}^{d} \mathbf{s}_j^4}{\left(\frac{1}{d}\sum_{j=1}^{d} \mathbf{s}_j^2\right)^2} \tag{1}$$

   We see that $\min(\text{Kurt}(\mathbf{s})) = 1$ when all $\mathbf{s}_j$ are equal and no outlier features exist, and $\max(\text{Kurt}(\mathbf{X})) = d$, which is the limit when $d-1$ neurons have activation magnitudes dominated by a single outlier feature.

2. **Max-Median Ratio** (across neurons): A metric for OFs more aligned with the original motivation of studying variation in summand magnitudes, described in App B for space considerations.

Bondarenko et al. (2023) show that activation kurtosis is a suitable metric for OFs, but define a different form of kurtosis. We aggregate over inputs first in Eq (1), which allows us to link OFs and signal propagation in Sec 4.

Exp details Our smaller scale setting uses 130M autoregressive transformers trained on CodeParrot,[2] with a similar setup to He & Hofmann (2024). Our larger transformer experiments are on the Languini dataset (Stanić et al., 2023). Further details and results in Apps G and H respectively.

[1] We do not centre $\mathbf{X}$ in s_j for ease of exposition. Fig 13 shows that centring does not make a qualitative difference for OFE.

[2] `https://huggingface.co/datasets/transformersbook/codeparrot-train`.

3. Normalisation Layers and Outlier Features

Several works have highlighted *Layer Normalisation* (LN) (Ba et al., 2016) as a cause of OFE (Kovaleva et al., 2021; Wei et al., 2022; Bondarenko et al., 2023). LN belongs to a family of normalisation (Norm) layers commonly used in sequence models, which normalise a representation vector $\boldsymbol{x} \in \mathbb{R}^d$ across the width dimension independently for different sequence positions. In general, for a centring scalar $c \in \{0, 1\}$, a Norm layer maps $\boldsymbol{x}$ to:

$$\text{Norm}(\boldsymbol{x}) = \frac{\boldsymbol{x} - c\mu(\boldsymbol{x})}{\sigma(\boldsymbol{x})} \odot \boldsymbol{\gamma} + \boldsymbol{\beta}, \quad \text{where:} \tag{2}$$

$$\mu(\boldsymbol{x}) = \frac{1}{d}\sum_{i=1}^{d} \boldsymbol{x}_i, \quad \sigma(\boldsymbol{x})^2 = \frac{1}{d}\sum_{i=1}^{d} (\boldsymbol{x}_i - c\mu(\boldsymbol{x}))^2 \tag{3}$$

LN is when $c = 1$, with a trainable scale $\boldsymbol{\gamma}$ and bias $\boldsymbol{\beta}$ vectors initialised to all 1s and 0s respectively.

Previous works have attributed OFE in standard architectures to the $\boldsymbol{\gamma}, \boldsymbol{\beta}$ parameters of LN incurring outliers during training (Kovaleva et al., 2021; Wei et al., 2022). It is therefore natural to ask if simpler Norms with different formulations of Eq (2) remove OFE. In particular, *Root Mean Square Normalisation* (RMSNorm) (Zhang & Sennrich, 2019) is a commonly used Norm known to be as performant as LN in Transformer training (Rae et al., 2021; Touvron et al., 2023). Compared to LN, RMSNorm fixes the bias $\boldsymbol{\beta} = 0$ and removes the centring by setting $c = 0$. One step further would be to remove trainable parameters entirely by fixing $\boldsymbol{\gamma} = 1$, thus simply projecting $\boldsymbol{x}$ to the hypersphere of norm $\sqrt{d}$. This is dubbed *Simple RMSNorm* (SRMSNorm) by Qin et al. (2023), who find that SRMSNorm has minimal performance degradation but is more computationally efficient than LN and RMSNorm.

Fig 2 shows that Transformers trained with RMSNorm and SRMSNorm, alongside LN, incur OFE: peak kurtosis during training across Norms is over 4 orders of magnitude larger than initialisation. In fact, the Pre-SRMSNorm model has the highest Kurtosis, despite its lack of trainable Norm weights.

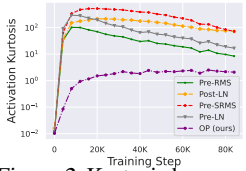


Figure 2. Kurtosis becomes large (i.e. OFE) when training with different Norms at 130M scale. We plot the residual stream entering the 2nd of 6 blocks. Other layers in Fig 12.

This result demonstrates that the previous explanations for OFE relating to trainable scales and biases in Norms cannot fully explain why OFs emerge during training. Furthermore, we show OFE in both Pre-Norm (Baevski & Auli, 2018; Child et al., 2019) and Post-Norm (Vaswani et al., 2017) architectures, which are the two most popular ways to place Norm layers relative to residual connections (Xiong et al., 2020). This further highlights that OFE occurs independent of the standard choices of Norm location.

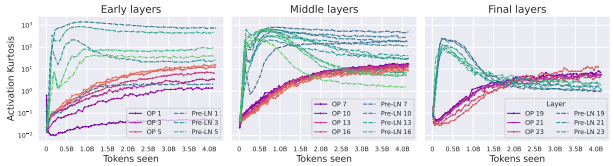


Figure 3. Our OP block mitigates OFE. We plot activation kurtosis of the inputs before Query/Key/Value weights in a layer. Experiments are at 1.2B scale using a max learning rate of 0.001. The OP model removes the final LN before unembedding; the effect of the final LN on OFE is shown in Fig 17.

Having established that removing trainable weights in Norms still results in OFE, the next question we ask is: *how does removing standard Norms entirely influence Outlier Feature emergence*?

Recovering training benefits in unnormalised Transformers This is a challenging question, not least because it is not fair to compare OFE in architectures that converge at different speeds: Norms are well known to be an important component in most NN architectures, providing various benefits for initialisation, convergence speed, and training stability. Thus, to answer the above question, in App C we review different hypotheses for the benefits of Norms in Transformer training dynamics in order to motivate a novel Transformer Block in Fig 5, which we call the Outlier Protected (OP) block, that matches the Pre-Norm block in convergence speed, while eschewing standard Norm layers.

Removing Norms mitigates Outlier Features In Fig 2 we see that the Outlier Protected (OP) Block greatly reduces OFE compared to standard blocks. Fig 3 presents the corresponding plots in our 1.2B parameter experiments using our kurtosis metric, across layers. We draw several consistent insights: 1) the peak kurtosis across the course of training is consistently higher in Pre-LN, sometimes by over 2 orders of magnitude, across different layers; 2) the kurtosis across training is usually higher in Pre-LN (up to 4 orders of magnitude here), especially at early training times and in earlier layers; 3) OFE need not be monotonic in training time, at least when measured by our proposed metrics. Tab 3 ablates the effect of Norm positioning on OFE.

Nevertheless, we observe in Fig 3 that kurtosis still slightly increases in our OP blocks (to relatively modest values; around 20), usually monotonically throughout training. Moreover, the question of why normalisation layers cause outlier features is still unanswered despite the clear evidence that removing them mitigates OF prevalence.

Sec 3 key takeaways: normalisation layers and OFE.

- OFE still occurs for weight-less or uncentred Norms, & both Pre/Post-Norm (Figs 2, 12 and 15).
- The OP Block (Fig 5) matches Pre-LN training speed/stability (Tabs 1 and 2), without standard Norms.
- The OP Block greatly reduces OFE compared to standard blocks (Figs 2, 3 and 11).

4. Signal Propagation and Outlier Features

To better understand why OFs appear (albeit greatly reduced) in the OP block, and why Norms cause OFs, we examine *Signal Propagation* behaviour during training and its effect on OFE. This will also clarify why modifications that improve Signal Propagation reduce OFE (Wortsman et al., 2023). Signal Propagation (Poole et al., 2016; Schoenholz et al., 2017; Hayou et al., 2019; 2021; Martens et al., 2021; Noci et al., 2022; He et al., 2023) studies the *input-wise* Gram matrix $\Sigma_{\text{I}} = \mathbf{X}\mathbf{X}^\top \in \mathbb{R}^{n\times n}$, & how Σ_{I} evolves in deep NNs for different layer features $\mathbf{X} \in \mathbb{R}^{n\times d}$.

On the other hand, our kurtosis metric, Eq (1), is related to the *feature-wise* Gram $\Sigma_{\text{F}} \overset{\text{def}}{=} \mathbf{X}^\top\mathbf{X} \in \mathbb{R}^{d\times d}$. Recall our kurtosis is the 4$^{\text{th}}$ moment of $\mathbf{X} \in \mathbb{R}$ normalised by the square second moment $m_2(\mathbf{X}) = \frac{1}{nd}\sum_{\alpha\leq n, j\leq d}\mathbf{X}^2_{\alpha,j} = \frac{1}{nd}\|\mathbf{X}\|^2_F$. As kurtosis is scale-invariant we can consider the setting where $m_2(\mathbf{X}) = 1$ and the average squared activation is 1 without loss of generality[3]. In this case, $\text{Tr}(\Sigma_{\text{I}}) = \text{Tr}(\Sigma_{\text{F}}) = nd$ by the cyclic trace property.

Then, $\text{Kurt}(\mathbf{X}) = \frac{1}{d}\sum_j^d\left(\frac{1}{n}\sum_\alpha^n \mathbf{X}^2_{\alpha,j}\right)^2 = \frac{1}{d}\sum_{j=1}^d(\frac{1}{n}\Sigma_{\text{F}})^2_{j,j}$, which is simply a second moment (or average of squares) of diagonal entries of the feature-wise Gram Σ_{F}. At the same time, again by the cyclic property of the trace, we have:

$$\text{Tr}(\Sigma_{\text{F}}\Sigma_{\text{F}}) = \text{Tr}(\mathbf{X}^\top\mathbf{X}\mathbf{X}^\top\mathbf{X}) = \text{Tr}(\mathbf{X}\mathbf{X}^\top\mathbf{X}\mathbf{X}^\top) = \text{Tr}(\Sigma_I\Sigma_I)$$

$$\implies n^2 d\cdot\text{Kurt}(\mathbf{X}) + \sum_{i\neq j}^{d}\left(\Sigma_{\text{F}}\right)^2_{i,j} = \sum_{\alpha,\beta\leq n}\left(\Sigma_{\text{I}}\right)^2_{\alpha,\beta} \qquad (4)$$

In words, Eq (4) tells us that the sum of squared elements of Σ_{F} is equal to the sum of squared elements of Σ_{I}. On the left of Eq (4) we decompose Eq (4) into our feature-wise kurtosis (Eq (1), of interest for OFE), plus the sum of squared off-diagonal elements of Σ_{F}, equal to the sum of squared elements of Σ_{I} on the right. Hence, it is clear that Signal Propagation is relevant for OFE. Contrary to most existing works in Signal Propagaton, Eq (4) is true throughout training, not only at initialisation.

In particular, we see that the right-hand side of Eq (4) captures both the (normalised) activation norms across inputs $\sum_{\alpha\leq n}\left(\Sigma_{\text{I}}\right)^2_{\alpha,\alpha}$ from the diagonal terms, and inner products between inputs $\sum_{\alpha,\beta\leq n;\alpha\neq\beta}\left(\Sigma_{\text{I}}\right)^2_{\alpha,\beta}$ in the off-diagonals. If $\mathbf{X}$ is the output of a Norm layer, then $\frac{1}{d}\Sigma_{\text{I}}$ becomes a cosine similarity matrix with diagonals equal to 1. Deep NNs, and Transformers in particular, are well known to be susceptible to a particular Signal Prop defect called *rank collapse* (Dong et al., 2021; Martens et al., 2021) where this cosine similarity matrix $\frac{1}{d}\Sigma_{\text{I}}$ degenerates to the all ones ma-

[3] In all experiments concerning signal propagation (i.e. input-wise correlations or equivalently, the elements of Σ_{I}), we first scale $\mathbf{X}$ down by $\sqrt{m_2(\mathbf{X})}$ to give $m_2(\mathbf{X}) = 1$ and make $\mathbf{X}$ scale invariant.

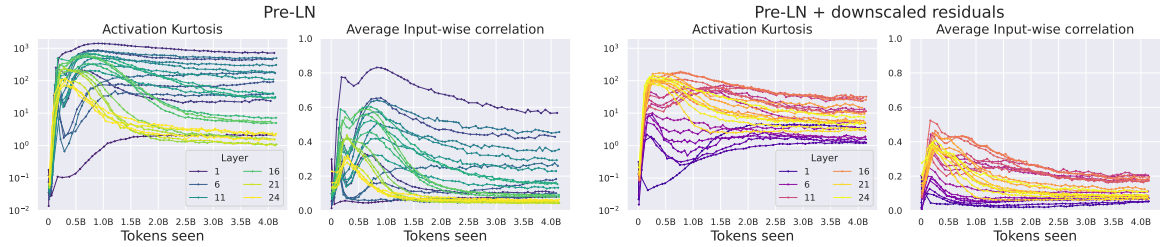


Figure 4. Pre-LN layers at 1.2B scale with extreme OFE (left) are those with bad Signal Prop close to rank collapse during training (centre left). (**Right vs. left two plots**) Downweighting residual branches improves signal propagation during training and results in less extreme OFE, particularly in early layers. Respective plots for OP (with & without final LN before unembedding) in Fig 17.

trix and all inputs look identical to a deep layer. Noci et al. (2022) and He et al. (2023) demonstrate that, at least at initialisation, the off-diagonals of Σ_{I} are positive and increase monotonically with depth in deep Transformers towards rank collapse, even with Signal Prop inspired modifications that ensure a non-degenerate deep limit exists.

Bad Signal Prop encourages OFE For OFE, the upshot of these observations is that poor Signal Propagation (in terms of large off-diagonal values of Σ_{I}, close to rank collapse) will make the right-hand side of Eq (4) large (the rank collapsed limit has RHS n^2d^2, compared to nd^2 when the inputs are orthogonal and Σ_{I} is diagonal). In turn, this puts pressure on the LHS, which contains the feature kurtosis, to be large, hence OFE. This argument is not fully rigorous because the off-diagonals $\sum_{i,j\leq d, i\neq j}\left(\Sigma_{\text{F}}\right)^2_{i,j}$, which captures correlations between different neuron features, could increase on the LHS to allow the kurtosis to remain low. Having said that, we formalise the intuition of bad Signal Prop leading to larger feature kurtosis in the context of Gaussian features in Prop J.1.

In any case, we can empirically study the link between bad signal propagation and OFEs, which we do in Figs 4 and 17 for Pre-LN and OP blocks at 1.2B scale. For each layer in different architectures, we plot both the evolution of the kurtosis on the left and the average off-diagonal entry of $\frac{1}{d}\Sigma_{\text{I}} = \frac{1}{d}\mathbf{X}\mathbf{X}^\top$ (i.e. the average input-wise correlation) on the right, normalised so that $m_2(\mathbf{X}) = 1$.

As implied by Eq (4), we see a strong association between kurtosis and Signal Propagation: the layers with larger kurtosis tend to be the ones with larger input correlations, and vice versa. In particular, in Fig 4, we see that the Pre-LN layer (2 in this case) with the most extreme OFE (kurtosis peaking over 1000) is precisely the one with the worst Signal Propagation closest to rank collapse (average input correlation peaking over 0.8) during training. Moreover, the trajectory of kurtosis closely tracks the trajectory of input correlations throughout training, with their peaks appearing at similar training steps, across layers.

Given that Signal Propagation characteristics during training depict how a model creates structure (through increasing or decreasing the inner product for different inputs) in its layer representations to best learn the task at hand, our results suggest that OFs occur partly due to the inherent nature of the task that the model is trained on, particularly in architectures that are less prone to OFs, such as our OP block. In Transformers, this is most apparent in the inputs to the final unembedding layer, which are linearly projected to the predictions: they tend to have similar kurtosis levels in both OP and Pre-Norm blocks, and the most extreme OFE rarely occurs in the final layers, (Figs 1, 3, 12 and 16). We hypothesise this is because extreme OFE in late layers would imply high kurtosis which would imply representations close to rank collapse by Eq (4), from which it would be hard to make useful linear predictions.

The correlation between OFE and bad Signal Propagation also allows us to observe that interventions that worsen Signal Propagation *during training* cause increased OFE. Likewise, methods improving Signal Propagation throughout training help to mitigate OFE, as seen for downscaled residuals in Fig 4. We explore this link further in App D, and discuss the effect of optimiser choices on OFE in App E.

Sec 4 key takeaways: Signal Propagation and OFE.

- Signal Propagation is fundamentally connected to OFE: worse Signal Prop generally implies higher kurtosis and vice versa, throughout training (Eq (4), Prop J.1, and Figs 4, 17 and 16).
- The OP block's mild OFs can be traced to increasing input correlations while training (Fig 17).
- Choices that improve Signal Prop during training (e.g. scaled residuals) also reduce OFs (Fig 4).
- Removing standard Norms can improve Signal Prop, & OFE, during training (Figs 4 and 17).

5. Discussion

The goal of this work was to better understand the emergence of Outlier Features during standard NN training, and propose architectural and optimisation (App E) interventions that minimise their prevalence. Our main contributions include identifying signal propagation as a key quantity to predict OFE during training, and the Outlier Protected block to reduce OFE. Future work is discussed in App F. Overall, our results complement and further existing works on OFs, & our understanding of NN training dynamics in general.

References

Allen-Zhu, Z. and Li, Y. Towards understanding ensemble, knowledge distillation and self-distillation in deep learning. *arXiv preprint arXiv:2012.09816*, 2020.

Ashkboos, S., Mohtashami, A., Croci, M. L., Li, B., Jaggi, M., Alistarh, D., Hoefler, T., and Hensman, J. Quarot: Outlier-free 4-bit inference in rotated llms. *arXiv preprint arXiv:2404.00456*, 2024.

Ba, J. L., Kiros, J. R., and Hinton, G. E. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Bachlechner, T., Majumder, B. P., Mao, H., Cottrell, G., and McAuley, J. Rezero is all you need: Fast convergence at large depth. In *Uncertainty in Artificial Intelligence*, pp. 1352–1361. PMLR, 2021.

Baevski, A. and Auli, M. Adaptive input representations for neural language modeling. In *International Conference on Learning Representations*, 2018.

Baratin, A., George, T., Laurent, C., Devon Hjelm, R., Lajoie, G., Vincent, P., and Lacoste-Julien, S. Implicit regularization via neural feature alignment. In Banerjee, A. and Fukumizu, K. (eds.), *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 2269–2277. PMLR, 13–15 Apr 2021. URL https://proceedings.mlr.press/v130/baratin21a.html.

Biderman, S., Schoelkopf, H., Anthony, Q. G., Bradley, H., O'Brien, K., Hallahan, E., Khan, M. A., Purohit, S., Prashanth, U. S., Raff, E., et al. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pp. 2397–2430. PMLR, 2023.

Bondarenko, Y., Nagel, M., and Blankevoort, T. Understanding and overcoming the challenges of efficient transformer quantization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7947–7969, 2021.

Bondarenko, Y., Nagel, M., and Blankevoort, T. Quantizable transformers: Removing outliers by helping attention heads do nothing. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Brock, A., De, S., Smith, S. L., and Simonyan, K. High-performance large-scale image recognition without normalization. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 1059–1071. PMLR, 18–24 Jul 2021.

Child, R., Gray, S., Radford, A., and Sutskever, I. Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*, 2019.

Chizat, L., Oyallon, E., and Bach, F. On lazy training in differentiable programming. *Advances in neural information processing systems*, 32, 2019.

Dao, T., Fu, D., Ermon, S., Rudra, A., and Ré, C. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in Neural Information Processing Systems*, 35:16344–16359, 2022.

De, S. and Smith, S. Batch normalization biases residual blocks towards the identity function in deep networks. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 19964–19975. Curran Associates, Inc., 2020.

Dehghani, M., Djolonga, J., Mustafa, B., Padlewski, P., Heek, J., Gilmer, J., Steiner, A. P., Caron, M., Geirhos, R., Alabdulmohsin, I., et al. Scaling vision transformers to 22 billion parameters. In *International Conference on Machine Learning*, pp. 7480–7512. PMLR, 2023.

Dettmers, T., Lewis, M., Belkada, Y., and Zettlemoyer, L. Gpt3. int8 (): 8-bit matrix multiplication for transformers at scale. *Advances in Neural Information Processing Systems*, 35:30318–30332, 2022.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Dong, Y., Cordonnier, J.-B., and Loukas, A. Attention is not all you need: Pure attention loses rank doubly exponentially with depth. In *International Conference on Machine Learning*, pp. 2793–2803. PMLR, 2021.

Gao, J., He, D., Tan, X., Qin, T., Wang, L., and Liu, T.-Y. Representation degeneration problem in training natural language generation models, 2019.

Gu, A. and Dao, T. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

Hayou, S., Doucet, A., and Rousseau, J. On the impact of the activation function on deep neural networks training. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2672–2680. PMLR, 09–15 Jun 2019.

Hayou, S., Clerico, E., He, B., Deligiannidis, G., Doucet, A., and Rousseau, J. Stable resnet. In *International*

Conference on Artificial Intelligence and Statistics, pp. 1324–1332. PMLR, 2021.

He, B. and Hofmann, T. Simplifying transformer blocks. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=RtDok9eS3s.

He, B. and Ozay, M. Feature kernel distillation. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=tBIQEvApZK5.

He, B., Martens, J., Zhang, G., Botev, A., Brock, A., Smith, S. L., and Teh, Y. W. Deep transformers without shortcuts: Modifying self-attention for faithful signal propagation. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=NPrsUQgMjKK.

Henry, A., Dachapally, P. R., Pawar, S. S., and Chen, Y. Query-key normalization for transformers. In Cohn, T., He, Y., and Liu, Y. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 4246–4253, Online, November 2020a. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.379. URL https://aclanthology.org/2020.findings-emnlp.379.

Henry, A., Dachapally, P. R., Pawar, S. S., and Chen, Y. Query-key normalization for transformers. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 4246–4253, 2020b.

Huang, X. S., Perez, F., Ba, J., and Volkovs, M. Improving transformer optimization through better initialization. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 4475–4483. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/huang20f.html.

Ioffe, S. and Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pp. 448–456. pmlr, 2015.

Jacot, A., Gabriel, F., and Hongler, C. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

Joudaki, A., Daneshmand, H., and Bach, F. On the impact of activation and normalization in obtaining isometric embeddings at initialization. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Kovaleva, O., Kulshreshtha, S., Rogers, A., and Rumshisky, A. Bert busters: Outlier dimensions that disrupt transformers. *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2021.

Kunstner, F., Yadav, R., Milligan, A., Schmidt, M., and Bietti, A. Heavy-tailed class imbalance and why adam outperforms gradient descent on language models. *arXiv preprint arXiv:2402.19449*, 2024.

Lee, J., Sohl-dickstein, J., Pennington, J., Novak, R., Schoenholz, S., and Bahri, Y. Deep Neural Networks as Gaussian Processes. In *International Conference on Learning Representations*, 2018.

Li, M. B., Nica, M., and Roy, D. M. The neural covariance sde: Shaped infinite depth-and-width networks at initialization. *arXiv preprint arXiv:2206.02768*, 2022.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Lou, Y., Mingard, C. E., and Hayou, S. Feature learning and signal propagation in deep neural networks. In *International Conference on Machine Learning*, pp. 14248–14282. PMLR, 2022.

Martens, J., Ballard, A., Desjardins, G., Swirszcz, G., Dalibard, V., Sohl-Dickstein, J., and Schoenholz, S. S. Rapid training of deep neural networks without skip connections or normalization layers using deep kernel shaping. *arXiv preprint arXiv:2110.01765*, 2021.

Matthews, A. G. d. G., Rowland, M., Hron, J., Turner, R. E., and Ghahramani, Z. Gaussian Process Behaviour in Wide Deep Neural Networks. In *International Conference on Learning Representations*, volume 4, 2018.

Meterez, A., Joudaki, A., Orabona, F., Immer, A., Rätsch, G., and Daneshmand, H. Towards training without depth limits: Batch normalization without gradient explosion. *arXiv preprint arXiv:2310.02012*, 2023.

Noci, L., Bachmann, G., Roth, K., Nowozin, S., and Hofmann, T. Precise characterization of the prior predictive distribution of deep relu networks. *Advances in Neural Information Processing Systems*, 34:20851–20862, 2021.

Noci, L., Anagnostidis, S., Biggio, L., Orvieto, A., Singh, S. P., and Lucchi, A. Signal propagation in transformers: Theoretical perspectives and the role of rank collapse. *arXiv preprint arXiv:2206.03126*, 2022.

Noci, L., Li, C., Li, M. B., He, B., Hofmann, T., Maddison, C., and Roy, D. M. The shaped transformer: Attention models in the infinite depth-and-width limit. *arXiv preprint arXiv:2306.17759*, 2023.

Nrusimha, A., Mishra, M., Wang, N., Alistarh, D., Panda, R., and Kim, Y. Mitigating the impact of outlier channels for language model quantization with activation regularization. *arXiv preprint arXiv:2404.03605*, 2024.

Oono, K. and Suzuki, T. Graph neural networks exponentially lose expressive power for node classification. In *International Conference on Learning Representations*, 2019.

Pennington, J., Schoenholz, S., and Ganguli, S. Resurrecting the sigmoid in deep learning through dynamical isometry: theory and practice. *Advances in neural information processing systems*, 30, 2017.

Poole, B., Lahiri, S., Raghu, M., Sohl-Dickstein, J., and Ganguli, S. Exponential expressivity in deep neural networks through transient chaos. In Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

Puccetti, G., Rogers, A., Drozd, A., and Dell'Orletta, F. Outliers dimensions that disrupt transformers are driven by frequency. In *Findings of EMNLP 2022*. Association for Computational Linguistics, 2022.

Qin, Z., Li, D., Sun, W., Sun, W., Shen, X., Han, X., Wei, Y., Lv, B., Yuan, F., Luo, X., et al. Scaling transnormer to 175 billion parameters. *arXiv preprint arXiv:2307.14995*, 2023.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. Language models are unsupervised multitask learners. 2019.

Rae, J. W., Borgeaud, S., Cai, T., Millican, K., Hoffmann, J., Song, F., Aslanides, J., Henderson, S., Ring, R., Young, S., et al. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*, 2021.

Rosca, M. C. On discretisation drift and smoothness regularisation in neural network training. *arXiv preprint arXiv:2310.14036*, 2023.

Schoenholz, S. S., Gilmer, J., Ganguli, S., and Sohl-Dickstein, J. Deep information propagation. In *International Conference on Learning Representations*, 2017.

Smith, S. L., Brock, A., Berrada, L., and De, S. Convnets match vision transformers at scale. *arXiv preprint arXiv:2310.16764*, 2023.

Stanić, A., Ashley, D., Serikov, O., Kirsch, L., Faccio, F., Schmidhuber, J., Hofmann, T., and Schlag, I. The languini kitchen: Enabling language modelling research at different scales of compute. *arXiv preprint arXiv:2309.11197*, 2023.

Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., and Liu, Y. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Sun, M., Chen, X., Kolter, J. Z., and Liu, Z. Massive activations in large language models. *arXiv preprint arXiv:2402.17762*, 2024.

Team, G., Mesnard, T., Hardin, C., Dadashi, R., Bhupatiraju, S., Pathak, S., Sifre, L., Rivière, M., Kale, M. S., Love, J., et al. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024.

Tian, Y., Wang, Y., Zhang, Z., Chen, B., and Du, S. S. JoMA: Demystifying multilayer transformers via joint dynamics of MLP and attention. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=LbJqRGNYCf.

Tieleman, T. and Hinton, G. Lecture 6.5-rmsprop, coursera: Neural networks for machine learning. *University of Toronto, Technical Report*, 6, 2012.

Timkey, W. and van Schijndel, M. All bark and no bite: Rogue dimensions in transformer language models obscure representational quality. In Moens, M.-F., Huang, X., Specia, L., and Yih, S. W.-t. (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 4527–4546, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.372. URL https://aclanthology.org/2021.emnlp-main.372.

Touvron, H., Cord, M., Sablayrolles, A., Synnaeve, G., and Jégou, H. Going deeper with image transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 32–42, 2021.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Vladimirova, M., Verbeek, J., Mesejo, P., and Arbel, J. Understanding priors in bayesian neural networks at the unit level. In *International Conference on Machine Learning*, pp. 6458–6467. PMLR, 2019.

Wei, X., Zhang, Y., Zhang, X., Gong, R., Zhang, S., Zhang, Q., Yu, F., and Liu, X. Outlier suppression: Pushing the limit of low-bit transformer language models. *Advances*

in Neural Information Processing Systems, 35:17402–17414, 2022.

Wortsman, M., Dettmers, T., Zettlemoyer, L., Morcos, A., Farhadi, A., and Schmidt, L. Stable and low-precision training for large-scale vision-language models. *Advances in Neural Information Processing Systems*, 36: 10271–10298, 2023.

Xiao, G., Lin, J., Seznec, M., Wu, H., Demouth, J., and Han, S. Smoothquant: Accurate and efficient post-training quantization for large language models. In *International Conference on Machine Learning*, pp. 38087–38099. PMLR, 2023.

Xiao, L., Bahri, Y., Sohl-Dickstein, J., Schoenholz, S., and Pennington, J. Dynamical isometry and a mean field theory of cnns: How to train 10,000-layer vanilla convolutional neural networks. In *International Conference on Machine Learning*, pp. 5393–5402. PMLR, 2018.

Xiao, L., Pennington, J., and Schoenholz, S. Disentangling trainability and generalization in deep neural networks. In *International Conference on Machine Learning*, pp. 10462–10472. PMLR, 2020.

Xiong, R., Yang, Y., He, D., Zheng, K., Zheng, S., Xing, C., Zhang, H., Lan, Y., Wang, L., and Liu, T. On layer normalization in the transformer architecture. In *International Conference on Machine Learning*, pp. 10524–10533. PMLR, 2020.

Yang, G. Wide feedforward or recurrent neural networks of any architecture are gaussian processes. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Yang, G. and Hu, E. J. Feature learning in infinite-width neural networks. *arXiv preprint arXiv:2011.14522*, 2020.

Yang, G., Pennington, J., Rao, V., Sohl-Dickstein, J., and Schoenholz, S. S. A mean field theory of batch normalization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=SyMDXnCcF7.

Zeng, A., Liu, X., Du, Z., Wang, Z., Lai, H., Ding, M., Yang, Z., Xu, Y., Zheng, W., Xia, X., Tam, W. L., Ma, Z., Xue, Y., Zhai, J., Chen, W., Liu, Z., Zhang, P., Dong, Y., and Tang, J. GLM-130b: An open bilingual pre-trained model. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=-Aw0rrrPUF.

Zhai, S., Likhomanenko, T., Littwin, E., Busbridge, D., Ramapuram, J., Zhang, Y., Gu, J., and Susskind, J. M. Stabilizing transformer training by preventing attention entropy collapse. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 40770–40803. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/zhai23a.html.

Zhang, B. and Sennrich, R. Root mean square layer normalization. *Advances in Neural Information Processing Systems*, 32, 2019.

Zhang, G., Botev, A., and Martens, J. Deep learning without shortcuts: Shaping the kernel with tailored rectifiers. In *International Conference on Learning Representations*, 2022a.

Zhang, H., Dauphin, Y. N., and Ma, T. Fixup initialization: Residual learning without normalization. In *International Conference on Learning Representations*, 2018.

Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X. V., et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022b.

A. Additional Related Work

Understanding Outlier Features Kovaleva et al. (2021); Timkey & van Schijndel (2021) first identified Outlier Features in trained Transformers and demonstrated that OFs are critical for representational quality and performance. Puccetti et al. (2022) highlight the importance of token frequency (Kunstner et al., 2024) for OFs in transformers trained on language data, which is related to the representation degeneration phenomenon of Gao et al. (2019), and certain "vertical" structures appearing in attention matrices during training. Bondarenko et al. (2023) term this vertical structure "no-op" behaviour, where uninformative tokens are given high attention weights, and show that modifying attention to encourage no-op behaviour can mitigate OFs. Dettmers et al. (2022) show that the effect of OFs is more pronounced at larger parameter scales, and Wortsman et al. (2023) suggest that OFs are related to increasing activation scales during training, motivating their use of downweighted residuals. Kovaleva et al. (2021); Wei et al. (2022) attribute OFs to the trainable parameters in Layer Normalisation. Nrusimha et al. (2024) show that OFs occur early in training, and are stronger in residual stream layers. Sun et al. (2024) demonstrate the existence of "massive activations" and show they act as bias terms in transformers. Allen-Zhu & Li (2020); He & Ozay (2022) study a theoretical framework where sparse activations naturally appear and grow with gradient descent, owing to certain "lucky" neurons being correlated with useful features at initialisation, in order to study ensembling and knowledge distillation in two-layer convolutional NNs.

Outlier Features and Quantisation Wei et al. (2022); Bondarenko et al. (2021) identified Outlier Features as an issue for quantised NNs. Most work in this area has focused on (weight) quantisation of already trained transformers (Dettmers et al., 2022; Xiao et al., 2023; Ashkboos et al., 2024), for efficiency gains at inference time. Dettmers et al. (2022) keep outlier features in full precision to avoid their quantisation errors, while Xiao et al. (2023) propose to migrate the quantisation difficulty of outlier features to their corresponding weights using some scaling factors. Ashkboos et al. (2024) apply the elegant idea of rotating the feature vectors with a random orthogonal matrix, which removes OFs in the rotated features. In terms of quantised training, Bondarenko et al. (2023) show that encouraging "no-op" behaviour can mitigate OFs and enable low-precision training, while Wortsman et al. (2023) employ downweighted residuals (among other techniques) for quantised CLIP training. We discuss how our findings relate and extend these insights in the main text. Nrusimha et al. (2024) propose to regularise the kurtosis of the outputs of a linear layer for low-precision training, which the authors argue prevents migrating quantisation difficulty to the weights. We employ kurtosis to measure OFs, but focus on the kurtosis of the inputs to a linear layer.

Normalisation Layers Normalisation Layers have been near ever-present in NNs since their introduction (Ioffe & Szegedy, 2015; Ba et al., 2016), owing to their training benefits. Many works since have considered removing Normalisation layers, by finding alternative mechanisms that keep their benefits. De & Smith (2020) identify an implicit effect of Normalisation layers in Pre-Norm is to downweight residual branches, which enables training deep NNs without Normalisation. Hayou et al. (2021) show this theoretically using Signal Propagation theory, and propose downweighting residuals with a scale factor $O(1/\sqrt{\text{depth}})$ to do so, which Noci et al. (2022) corroborate in the transformer setting. Martens et al. (2021); Zhang et al. (2022a) demonstrate how to remove residual connections alongside normalisation layers in convolutional NNs using "transformed" activations, which He et al. (2023) extend to the Transformer architecture by making attention more identity-like (see also "shaped" attention, Noci et al. (2023)). Brock et al. (2021); Smith et al. (2023) propose NFNets, and achieve state of the art performance on the ImageNet benchmark in an unnormalised residual convolution architecture, highlighting that Normalisation layers are not necessary for best performance in convolutional models. NFNets employ downweighted residual branches to fix Signal Propagation at initialisation, among other techniques including adaptive gradient clipping. However, He et al. (2023); He & Hofmann (2024) find that removing Normalisation Layers, even with Signal Propagation modifications like downweighting residuals, leads to a loss of performance in simplified Transformer blocks, implying that transformer training has different instabilities to convolutional models, and Normalisation layers have other training benefits in transformers.

Entropy Collapse Zhai et al. (2023) identify entropy collapse as a key training instability in transformers, where attention logits grow large during training. This causes the rows of the post-softmax attention matrix to become one-hot vectors and the attention weights are non-zero on only a single sequence position. To remedy entropy collapse, it is important to control the logits entering softmax from growing too large, and Zhai et al. (2023) propose σReparam which regularises the spectrum of Query-Key weights in order to do so. As an alternative, Query-Key Normalisation (Henry et al., 2020a), where the Queries and Keys are normalised using e.g. LayerNorm or RMSNorm after the Query/Key weight matrix (c.f. Post-QK Norm in Fig 40) has seen growing popularity, particularly in ViT-22B (Dehghani et al., 2023) where it was crucial

for stable training. Other "entropy regulating" mechanisms include tanh thresholding (Grok-1) and clamping attention logits (DBRX). The training stability benefits of controlling attention entropy through QK-Norm were shown at smaller scales in Wortsman et al. (2023), who argue that the quadratic dependence in the attention logits (on the queries and keys), causes large attention logits to appear during training, hence entropy collapse. This is as opposed to convolutional/MLP models which depend linearly on their inputs. Tian et al. (2024) propose joint MLP/Attention dynamics to predict attention entropy during training. We note that the "vertical" or "no-op" attention structures discussed in previous OF works (Puccetti et al., 2022; Bondarenko et al., 2023) have collapsed attention entropy, and can be thus be seen as undesirable from the perspective of other existing works.

Signal Propagation Signal Propagation studies how different inputs evolve through a deep NN, and how their activation magnitudes and cosine similarities evolve with depth.

For an input activation matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ of n inputs and width d, mapped to an activation matrix $\mathbf{X}_l \in \mathbb{R}^{n \times d}$ at layer l, signal propagation theory studies the evolution of the input-wise Gram matrix $\Sigma_{\text{I}}^l = \mathbf{X}_l \mathbf{X}_l^\top \in \mathbb{R}^{n \times n}$ for increasing depths l. This is a key object in an NN, as it tracks the "geometric information" that is conveyed in a deep layer, through inner products between different inputs. The diagonal elements of Σ_{I}^l indicate the activation norms, and the off-diagonal elements indicates how similar a deep layer views two inputs to be.

At initialisation, Σ_{I}^l can be tracked through its large d limits (Lee et al., 2018; Matthews et al., 2018; Yang, 2019). By studying Σ_{I}^l, one can see several issues that will afflict badly designed NNs (Schoenholz et al., 2017; Hayou et al., 2019; Yang et al., 2019; Dong et al., 2021; Martens et al., 2021), that affect either the diagonal elements, the off-diagonal elements or both at large depths. For example, the diagonal elements of Σ_{I} could blow up, which indicates exploding activation norms. For transformers, a particular degeneracy, known as rank collapse (Dong et al., 2021), can appear where the off-diagonals of Σ_{I}^l become positive and large, and Σ_{I}^l becomes proportional to the all ones matrix if activation norms are constant. Rank collapse is also possible in MLPs/CNNs Schoenholz et al. (2017); Hayou et al. (2019); Xiao et al. (2020); Martens et al. (2021), and is equivalent to the over-smoothing phenomenon in graph NNs (Oono & Suzuki, 2019). Martens et al. (2021) argue that rank collapse will lead to vanishing gradients, which Noci et al. (2022) show specifically for query and key parameters in transformers. As a result, when we refer to bad signal propagation, we mean that the off-diagonals of Σ_{I} are large and positive, close to rank collapse. This can be either through the RMS of input correlations, $\sqrt{\frac{1}{n(n-1)} \sum_{\alpha \neq \beta}^n \left(\frac{1}{d}\Sigma_{\text{I}}\right)^2_{\alpha,\beta}}$, as we show in the appendix, or the mean, $\frac{1}{n(n-1)} \sum_{\alpha \neq \beta}^n \left(\frac{1}{d}\Sigma_{\text{I}}\right)_{\alpha,\beta}$ as we show in Figs 4 and 17.

By applying Signal Propagation theory at initialisation, it is possible to design modifications to NN architectures and initialisations that correct potential degeneracies and/or yield simpler and/or more scalable architectures (Xiao et al., 2018; Hayou et al., 2021; Martens et al., 2021; Zhang et al., 2022a; Noci et al., 2022; He et al., 2023; He & Hofmann, 2024). But the vast majority of existing works in the literature do not theoretically study training beyond initialisation, and those that do are usually restricted to the NTK (Jacot et al., 2018) regime (Hayou et al., 2021; Martens et al., 2021), which precludes feature learning, and OFs. Lou et al. (2022) suggest that the feature alignment (Baratin et al., 2021) phenomenon during training is correlated to the rate at which signal propagation converges to its limit in a deep NN. Even at initialization, the distribution of the neurons becomes more heavy-tailed with depth (Vladimirova et al., 2019), thus making outliers more likely. Noci et al. (2021) gives a precise description of the kurtosis for ReLU networks, showing that it grows exponentially with depth. Together with the results presented in this work, there is empirical and theoretical evidence that depth has the double effect of increasing both the correlations and making large activations more likely, which we observe to be detrimental to outliers. However, the theoretical treatment of the emergence of outliers during training is still an open question.

B. Max Median Ratio metric

In the interests of space, we present a second metric, the Max Median Ratio (MMR), for measuring OFs here. MMR is more aligned with the original motivation for studying OFs due to large dynamic ranges in activation vectors. Specifically, we compute:

$$\text{MMR}(\mathbf{X}) \stackrel{\text{def}}{=} \text{Aggregate}_{\alpha}\left(\frac{\max_j |\mathbf{X}_{\alpha,j}|}{\text{median}_j |\mathbf{X}_{\alpha,j}|}\right), \tag{5}$$

or in words, the max neuron divided by the median absolute neuron, aggregated in some permutation invariant way across inputs. We typically use the mean to aggregate over inputs, but could also take e.g. median or max. MMR takes a minimum value 1 when all activations are identical in magnitude, and is unbounded when a dominant outlier feature exists.

MMR is shown to be highly correlated with feature Kurtosis in Figs 9, 11 and 15. Note that $\text{MMR}(\mathbf{X})$ is invariant to normalisation layers like RMSNorm without trainable parameters (Qin et al., 2023).

C. Motivating the Outlier Protected block

Several works (Zhang et al., 2018; De & Smith, 2020; Huang et al., 2020; Brock et al., 2021; Bachlechner et al., 2021; Touvron et al., 2021; Hayou et al., 2021; Noci et al., 2022; He et al., 2023; He & Hofmann, 2024) have observed that the initialisation benefits of Pre-Norm architectures can be recovered in unnormalised residual models using downweighted residual branches, through a theory known as Signal Propagation (Signal Prop) (Poole et al., 2016; Schoenholz et al., 2017; Hayou et al., 2019). Notably, Brock et al. (2021) achieve state of the art performance on the ImageNet benchmark using unnormalised convolutional architectures. However, it has been observed that fixing Signal Prop at initialisation is not sufficient to fully capture the benefits of Norms for training dynamics in unnormalised transformers (He et al., 2023; He & Hofmann, 2024), which implies that Norms have training benefits specific to the self-attention based transformer model.

At the same time, Zhai et al. (2023) show *Entropy Collapse*, where the Stochastic attention matrix has rows with low entropy (or in words, each sequence position attends to only one position, instead of many), to be a key training instability in softmax attention. Entropy collapse occurs because large attention logits saturate the softmax, and several *Entropy Regulation* (EntReg) mechanisms have been proposed to control the attention logits and thus prevent entropy collapse. Existing entropy regulating methods include QK-Norm (Henry et al., 2020b; Dehghani et al., 2023), *tanh* thresholding (Grok-1), σReparam (Zhai et al., 2023) and clamping the QK logits (DBRX). In standard Pre/Post-Norm attention blocks, a Norm layer appears before Query and Key weights and implicitly regulates attention entropy, to an extent.

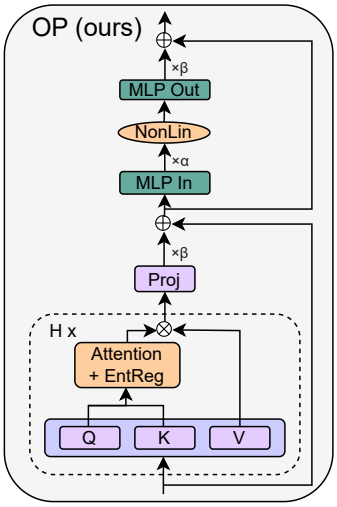


Figure 5. The Outlier Protected Transformer Block. We remove Pre-Norms and replace them with an Entropy Regulation mechanism to prevent entropy collapse, as well as downscaling residuals with $\beta < 1$.

Our key insight is to combine ideas from Signal Propagation and Entropy Collapse prevention to remove Normalisation layers while keeping their training benefits. This brings us to our *Outlier Protected* Block (OP), which replaces the Pre-Norm block by removing its normalisation layers in both Attention and MLP sub-blocks, and making three additional changes: 1) downweighting residual branches with some $\beta = O(1/\sqrt{\text{depth}}) < 1$ to recover Signal Prop benefits of Pre-Norms (De & Smith, 2020; Hayou et al., 2021; Noci et al., 2022; He & Hofmann, 2024), 2) adding an Entropy Regulation mechanism to prevent Entropy Collapse; we mainly use QK-Norm as it is relatively simple and performed well in all of our settings, but present experiments with tanh in App H.1, and 3) (optionally) scaling the inputs before the MLP nonlinearity by a scalar α to ensure the nonlinearity inputs are of order 1, as derived by Brock et al. (2021) using straightforward Signal Prop arguments.

In Tab 1, we show that our Outlier Protected block matches the standard Pre-LN block in terms of convergence speed at scales up to 1.2B parameters when trained with next token prediction on the Languini books dataset (Stanić et al., 2023) for nearly 4.5B tokens.[4] In App H.1, we ablate our OP block and show that the lack of an entropy regulation mechanism without normalisation layers causes training instabilities. This demonstrates that preventing entropy collapse is necessary to match training stability and convergence speed in unnormalised Transformers.

Table 1. OP matches Pre-LN performance at scales up to 1.2B params, on Languini Books (Stanić et al., 2023).[4]

Params	Block	Eval PPL
100M	Pre-LN	19.1
	OP	**18.9**
320M	Pre-LN	16.2
	OP	16.2
1.2B	Pre-LN	14.9
	OP	**14.6**

We note that independent of OFs, the OP block (Fig 5) is interesting in its own right because it shows that the initialisation-time Signal Prop and Entropy Collapse benefits of Norms in Transformers can be disentangled, and also reveals what was missing in previous methods that used Signal Prop arguments to correct initialisation defects in simplified unnormalised Transformers (He et al., 2023; He & Hofmann, 2024).

[4]We train for 4.2B tokens at 1.2B scale as this took 24 hours on 4 A100 80GB GPUs; we were unable to train for longer due to compute constraints. Scales under 1B were trained on a single A5000 or RTX-2080Ti GPU, taking around 2 days for 3.3B tokens (or equivalently, 50K steps at batch size 128 and sequence length 512).

D. Modifications That Affect Signal Prop During Training Affect OFE

To the best of our knowledge, in the Signal Propagation literature, most works have focused on characterising and improving Signal Propagation *at initialisation* due to analytic convenience. In particular, a practical focus of such works is to design architectural modifications that allow non-degenerate deep limits for models whose input cosine similarities can be well approximated by their large-width limits at initialisation (Pennington et al., 2017; Xiao et al., 2018; Hayou et al., 2021; Martens et al., 2021; He et al., 2023; He & Hofmann, 2024). Those considering training dynamics often reside in the kernel regime (Jacot et al., 2018) and are thus not compatible with feature learning (Chizat et al., 2019; Yang & Hu, 2020) which is necessary for OFE and Signal Prop dynamics during training. Our results connecting Signal Prop and OFE highlight the importance to the community of understanding Signal Prop dynamics during training in feature learning regimes, beyond initialisation. We note Tian et al. (2024) predict attention entropy dynamics through joint MLP/Attention. In any case, we empirically study the impact of initialisation-inspired Signal Prop architectural modifications in terms of OFE during training.

Downweighted residuals Of initialisation-inspired Signal Prop modifications, the most prevalent is downweighting residual branches $h(x) = x + \beta f(x)$ with some $\beta \ll 1$ (De & Smith, 2020; Hayou et al., 2021; Noci et al., 2022).[5] In Fig 4, we see that downweighting residuals (with a trainable scalar β initialised to 0.1) improves Signal Propagation in a 24-block 1.2B Pre-LN model, not only at initialisation but also *during training*, thereby reducing OFE (peak kurtosis is an order of magnitude lower). Having said that, Pre-LN with downscaled residuals still leads to higher kurtosis across layers than our OP block in Fig 17. Downscaling Pre-LN residuals leads to a small loss in performance of 0.2 perplexity. We show the corresponding results at 130M scale in Figs 18 to 20. Our results are consistent with previous work by Wortsman et al. (2023) who observe that downweighted residuals help for low precision training in CLIP models, motivated as a way to prevent OFs arising through increasing activations scales $\|\mathbf{X}\|_F$ during training. Given that standard models have Norm layers that are scale invariant (as are our OFE and Signal Prop metrics), we complement this argument by highlighting that the feature learning process of OFE is not only associated with increasing activation scales but also worsening Signal Propagation during training. Figs 14 and 41 show that $\|\mathbf{X}\|_F$ does not always correlate with OFs.

Normalisation layers On the other hand, for Norms, the difference between OP and standard blocks with Norms in Figs 4, 17 and 16 respectively is already clear evidence that standard Norm placements can lead to worse Signal Propagation (and OFE) during training. To the best of our knowledge, this observation has not been made previously. To test this further, we reintroduce the final LN right after the final OP block (just before the unembedding layer) into an OP model, with no Pre-Norms, in Fig 17. We see that the final LN causes some layers to see increases in both kurtosis and input correlations, and these layers correspond *precisely* to the final few blocks immediately preceding the LN. On the other hand, earlier layers further away from the final LN are largely unchanged in terms of both Signal Propagation and OFE during training. The model with a final LN performed slightly worse (0.1 perplexity difference).

Several works have discussed the effect of Norms on Signal Propagation theory at initialisation. The Deep Kernel Shaping (Martens et al., 2021) framework is compatible with LN (and also RMSNorm) layers, but makes other modifications (in weight initialisation and activation functions) that mean LN has no effect at initialisation in the wide limit. Other works show centred Norms in fact improve Signal Propagation at initialisation in MLPs by correcting imbalances in input activation norms to improve *Isometry* (Joudaki et al., 2023; Meterez et al., 2023) but consider non-standard architectures that are not residual and have Norm immediately following nonlinear activations, whereas standard Norms take the residual stream as input. Our work shows that initialisation and training can have very different Signal Prop behaviours.

Other Signal Prop modifications In Figs 21 and 23, we consider the effect of other initialisation-inspired Signal Propagation modifications in terms of OFE. In particular, we consider "transforming" activations to be more linear (Martens et al., 2021; Zhang et al., 2022a; Li et al., 2022), and "shaping" attention to be more identity-like (He et al., 2023; Noci et al., 2023; He & Hofmann, 2024). Although not predicted by initialisation theory, we find that these modifications mostly also reduce OFE and improve Signal Prop during training as well as initialisation. The latter finding is related to the work of Bondarenko et al. (2023) who show that "no-op" heads that place large attention weights on *shared* uninformative tokens encourage OFs: large attention weights on shared tokens also worsen signal propagation,[6] compared to identity-dominated attention, which can be seen as a "no-op" that instead encourages a token to attend to itself.

[5] Typically, the theory indicates that $\beta = O(\frac{1}{\sqrt{\text{depth}}})$ enables a well-behaved infinite-depth limit.

[6] Consider the extreme case when all attention weights are placed onto a single token (say the first one): all attention outputs will be equal to the first token's value representation so all token-wise cosine similarities are 1.

E. Optimisation Choices and OFE

So far, we have considered the impact of architectural hyperparameters for OFE, as the primary focus of our work. As OFE is a training phenomenon, it is important to also consider the role of optimsation choices, which we now explore.

Learning Rate Perhaps unsurprisingly, we find that using smaller LRs leads to reduced OFE during training, (Figs 6, 24 and 25), across different architectures. In these cases, slightly reducing the maximum LR in our scheduler (e.g. $0.001 \rightarrow 0.0003$ in Fig 6) did not lead to a loss in convergence speed (Fig 26), highlighting that one should use a smaller LR to avoid OFs if convergence is not affected. A direction for future work could be to explore the trade-off where one trains for more steps, but at lower LRs and precision.

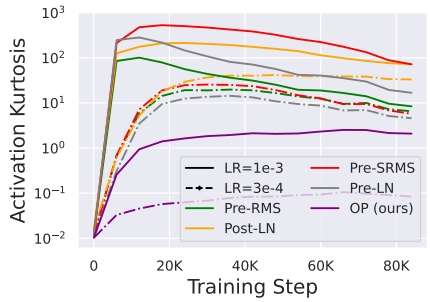


Figure 6. Smaller LRs lead to smaller OFs across different blocks.

Adaptivity As far as we are aware, the vast majority of modern NN architectures rely on the Adam optimiser (Kingma & Ba, 2014), which uses adaptive LRs for each parameter.

Given the importance of LR for OFs, we assess the effect of adaptive LRs through the ϵ hyperparameter in Adam, where the Adam update is $-\eta m_t/(\sqrt{v_t}+\epsilon)$, for learning rate η, and m_t and v_t denote first and second-moment estimates of each parameter's gradient, respectively. ϵ acts as a dampener to adaptive LRs, with larger ϵ reducing adaptivity for parameters with smaller v_t. In Figs 7, 28 and 30 we show that increasing ϵ also reduces OFE. Thus, one should increase Adam's ϵ to reduce OFE, if it does not impact convergence (like in Fig 27).

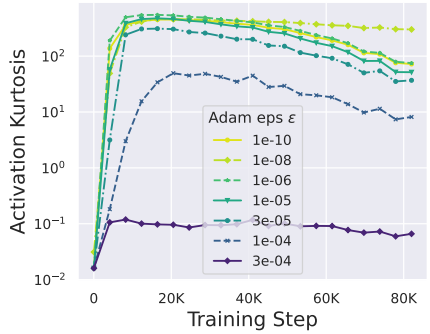


Figure 7. Larger Adam ϵ reduce OFs for 130M scale transformers.

Adam vs. SGD To push the question of adaptivity to the extreme, we consider the effect of replacing Adam with SGD in terms of OFE. As transformers are difficult to train (fast) with SGD, we consider OFs in a much simpler architecture and task: an MLP on CIFAR-10. In Figs 8 and 33 we see that SGD is not as susceptible to OFs, even with architectural changes that are OF prone, like Normalisation layers. In fact, with SGD the kurtosis can actually *decrease* during training. The model is a 6-layer Pre-Norm residual MLP with width 1024; we remove Pre-Norms for normless models. This also highlights that OFs are not specific to the Transformer model.

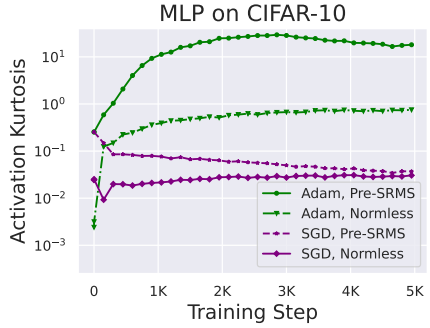


Figure 8. SGD has much reduced OFs, even when it can match Adam convergence speed (c.f. Fig 32).

The findings in this section, identifying key optimisation hyperparameters, point to the importance of (adaptive) LRs for OFE. This motivates us to break down the updates to kurtosis into terms of different powers in the learning rate η, in App I. There, we also consider settings without momentum, highlighting that momentum is not essential for OFE. We find that sub-leading order updates (in terms of LR) are the key driver in increasing kurtosis, providing a consistent mechanism for OFE that encapsulates our different observations concerning the roles of optimiser and architecture.

F. Limitations and future work

Though our work focuses on minimising Outlier Features (OFs) through understanding their emergence, in future work it would be interesting to assess if our suggested architectural and optimisation interventions do lead to practical improvements in low-precision training. Another practical limitation is the fact that while our experimental settings are sufficient to demonstrate and study the emergence of OFs, it remains to be seen if our Outlier Protected block continues to match Pre-Norm performance at larger scales beyond 1.2B parameters. Although we have no reason to believe otherwise, we are currently unable to test this due to compute constraints. Additional directions for future work could be to study OFs in other sequence modelling blocks besides the standard Transformer (such as those with gating (Gu & Dao, 2023)), combining our unnormalised Outlier Protected block with other simplified transformer blocks (He & Hofmann, 2024), as well as designing new optimisers with minimising OFs in mind. Our work also opens many new theoretical research questions for the community. Perhaps the most important is understanding how signal propagation evolves *during training*, and which factors affect this. Our results in App E indicate that this is a complex problem that depends not only on the architecture, like at initialisation, but also on the choice of optimiser.

G. Experimental Details

CodeParrot As discussed, all of our experiments at 130M scale (6 layers, width 768 transformers) are on next token prediction with the CodeParrot dataset, with 50K vocabulary size. We use a similar setup to He & Hofmann (2024), including their codebase.[7] We train with AdamW optimiser (Loshchilov & Hutter, 2017) and weight decay 0.1, betas=$(0.9, 0.999)$, and $\epsilon = 1e-8$ unless otherwise stated. We do not tie embeddings, and remove the final layer norm before unembedding layer. When we plot metrics (kurtosis, signal propagation, MMR etc) we plot the residual stream entering the attention sub-block (plots for the residual stream before the MLP sub-block are qualitatively the same). The only exception is the last layer, which is the input to the unembeddings. When we downweight residuals we set $\beta = 0.3$ in both attention and MLP sub-blocks unless otherwise stated. We do not train residual scalings β. Unless otherwise stated, we train with sequence length 128 and batch size 32 for 80K steps, with linear warmup to maximum learning rate $1e-3$, for 5% of the steps, before linear decay. We keep the standard parameter initialisations to $\mathcal{N}(0, \text{std} = 0.02)$ but upweight the input embeddings by a factor of 50 in order to make the average squared input 1 at initialisation, similar to considerations made by the Gemma model (Team et al., 2024). We use ReLU activations and do not scale inputs with an α, c.f. Fig 5, because ReLU is 1-homogeneous.

Languini For Languini (Stanić et al., 2023) our 100M, 320M, and 1.2B model sizes follow the "small" (depth 12, width 768), "medium" (depth 24, width 1024), and "XL" (depth 24, width 2048) model sizes provided by the authors, respectively. Our setup follows the authors in terms of codebase and tokeniser. We train with sequence length 512 and batch size 128, again with a maximum learning rate of $1e-3$ unless otherwise stated. We warm up the LR for the first 1.5% steps before linear decay. This learning rate was the largest stable and best performing choice on a logarithmic grid. We use linear warmup and linear decay after 1000 steps. We additionally use RoPE (Su et al., 2024), with GeLU nonlinearities in the MLPs. We use the same method as Brock et al. (2021) to calculate α to scale inputs to the GeLU. When we downweight residuals, we initialise $\beta = 0.1$ and allow them to be trainable. When we plot layer-wise metrics like kurtosis, we plot the outputs of the Pre-Normalisation layer (if there is one), otherwise, we treat the Normalisation layer as the identity and plot the residual stream going into the attention sub-block. We use tied embeddings. We also keep the standard parameter initialisations to $\mathcal{N}(0, \text{std} = 0.02)$ but upweight the input embeddings by a factor of 50 in order to make the average squared input 1 at initialisation.

CIFAR-10 For our MLP experiments on CIFAR-10, we train using batch size 2048 for 200 epochs. As described in App E, the model has 6 Pre-Norm layers with width 1024, giving 15M parameters. We zero initialise the last layer, and additionally downweight the output layer by $\sqrt{\text{width}}$ akin to μP (Yang & Hu, 2020), to encourage feature learning. We train with MSE loss and use LR 3 for SGD and 3e-3 for Adam. We use standard betas and epsilon for Adam and we do not use weight decay. We warm up the LR for 200 steps before cosine decay. We additionally found that it was important to whiten the inputs in order to observe OFE in the residual stream. We note that transformer embeddings are independently initialised, which can be thought of as implicitly whitening the embeddings for different tokens. Whitened inputs correspond to signal propagation with zero input correlations. This again suggests that signal propagation (and properties of the data) are important for OFs, but we leave further understanding of this to future work. We use PCA to whiten inputs.

[7] https://github.com/bobby-he/simplified_transformers

H. Additional Experiments

In this section, we include all additional experiments not included in the main paper.

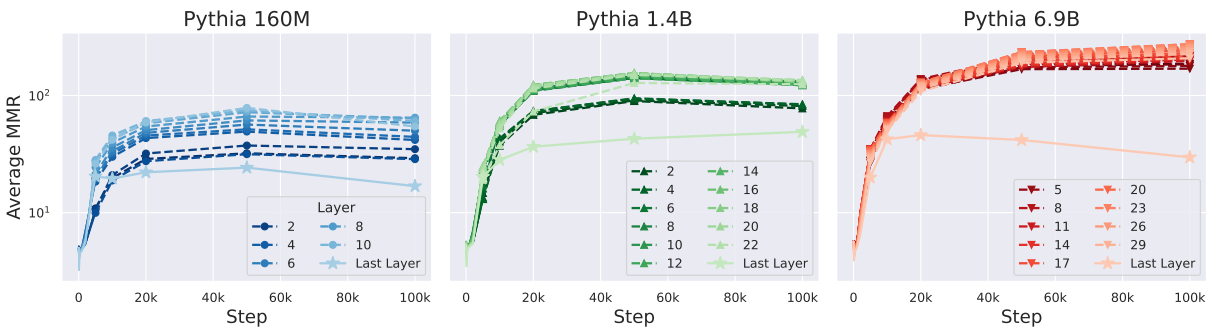


Figure 9. Max Median Ratio metric for Pythia, equivalent to Fig 1. We take the mean to aggregate over inputs

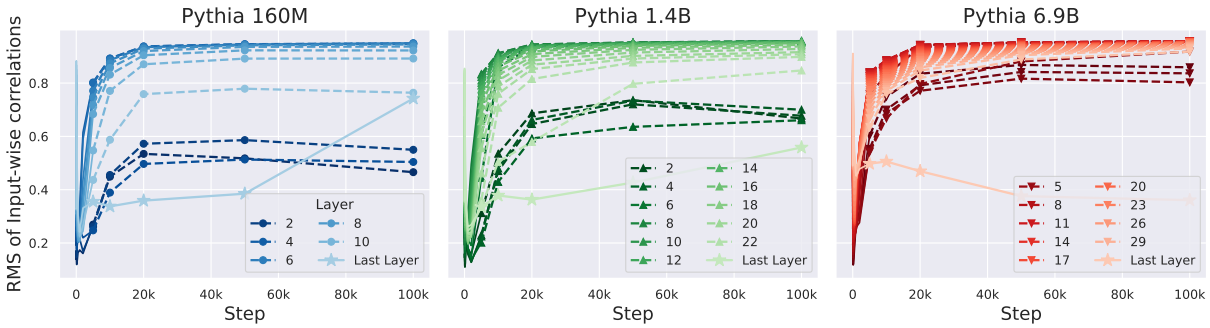


Figure 10. Signal Prop for Pythia, equivalent to Fig 1.

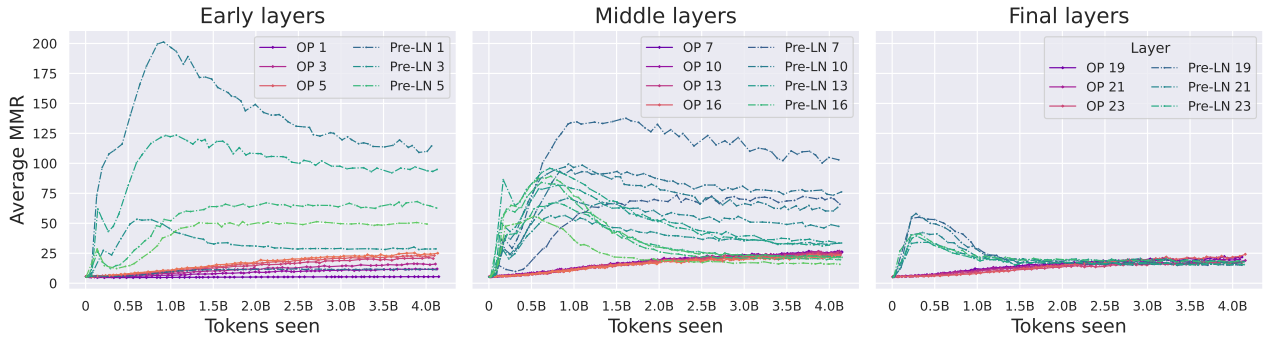


Figure 11. Average MMR metric comparing Pre-LN and OP blocks at 1.2B scale, equivalent to Fig 3.

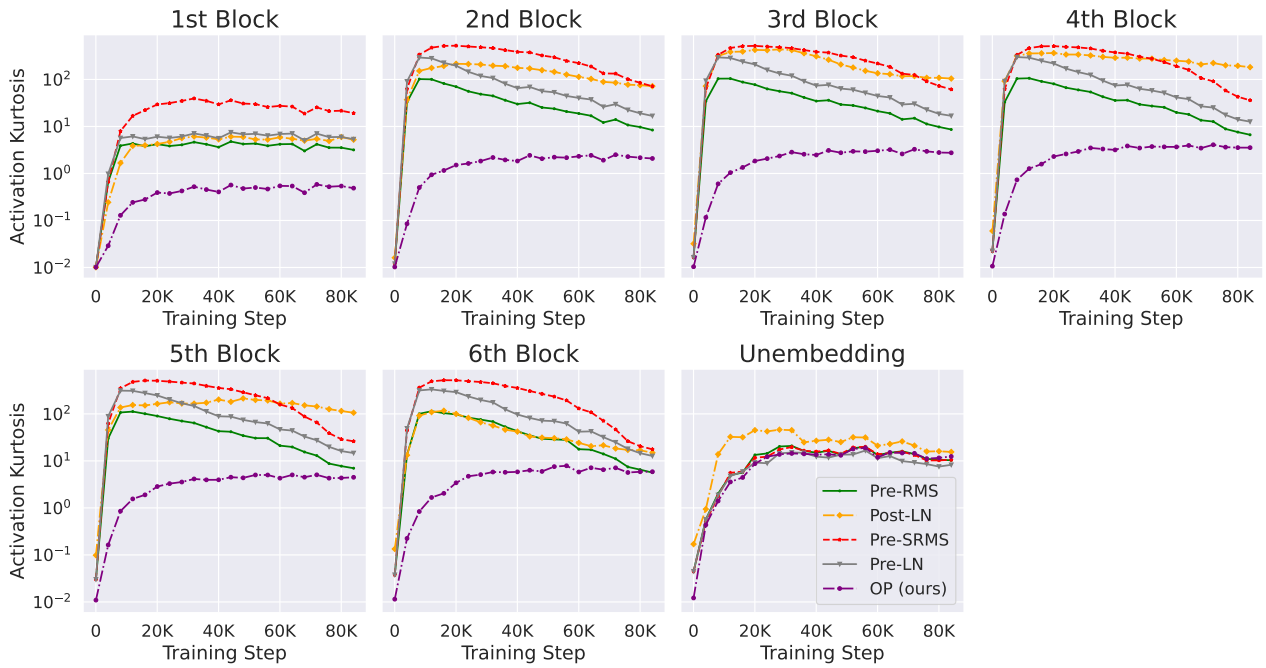


Figure 12. Kurtosis dynamics in different layers using different Norms and Norm locations on CodeParrot at 130M scale. Equivalent of Fig 2 but for the remaining layers. Fig 2 corresponds to the 2nd block.

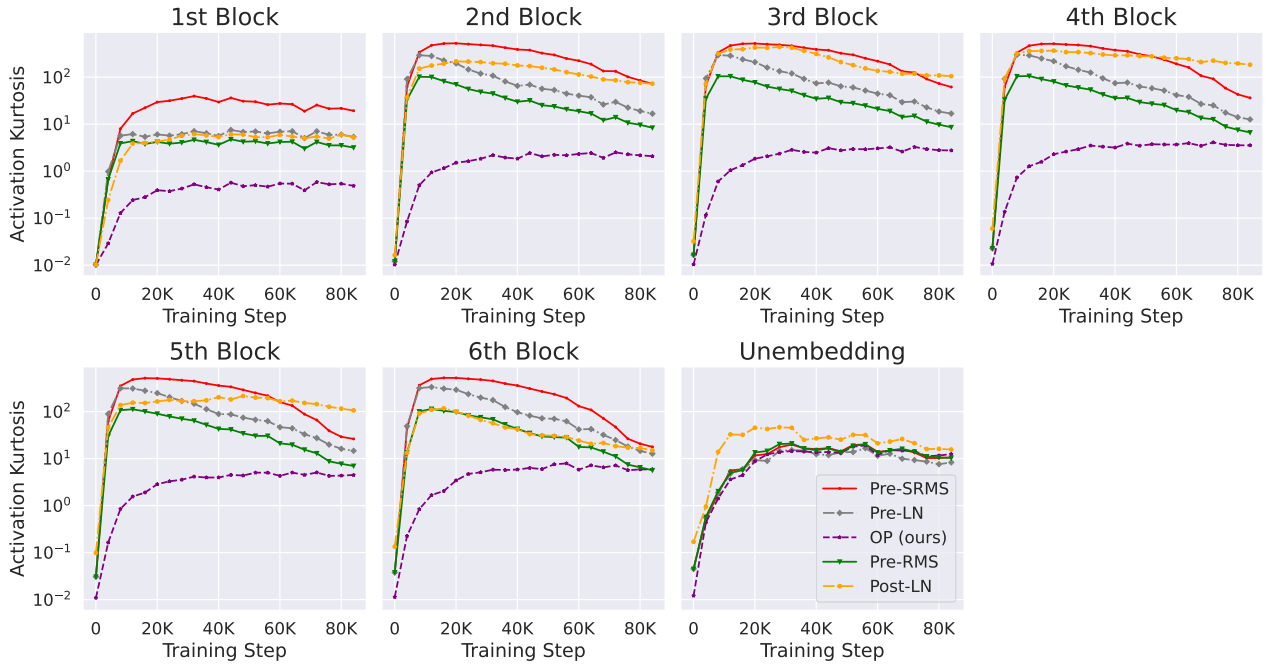


Figure 13. Equivalent of Fig 12 but with **centred activations** (centred along the width dimension). Notice there is no qualitative difference to kurtosis dynamics when centring activations.

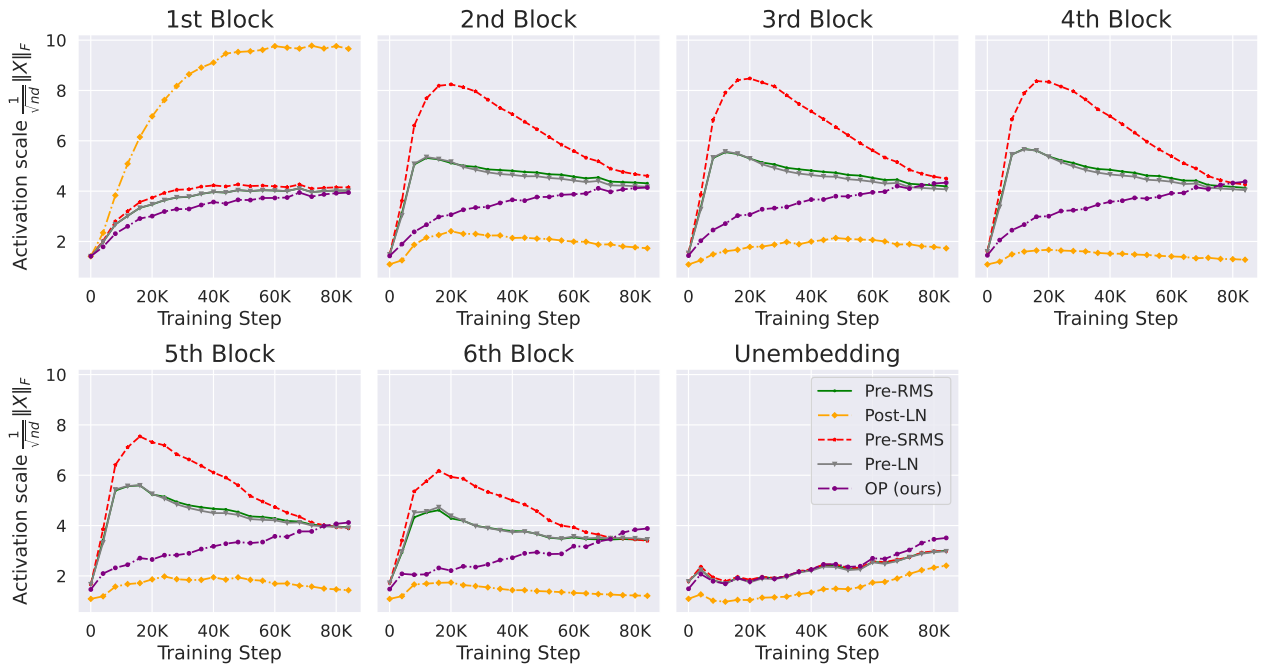


Figure 14. Equivalent of Fig 12 but for activation scale $\|\mathbf{X}\|_F$ trajectories through training. We see that activation scales do not correlate as well with OFs (Fig 12) as signal propagation (Fig 16). For example, Post-LN has smaller activation scales than the OP block in all blocks besides the first one, but much worse kurtosis in Fig 12.

Figure 15. Equivalent of Fig 12 but for the MMR metric (aggregated using maximum over the batch).

Figure 16. Equivalent of Fig 12 but for Signal Propagation (in terms of RMS of input correlations).

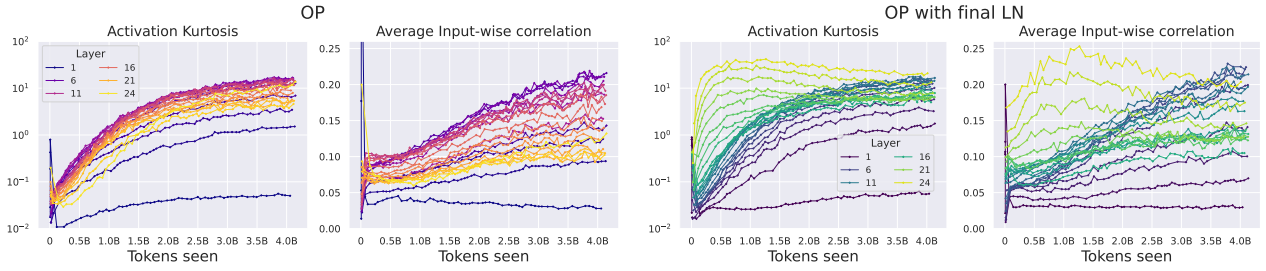


Figure 17. OP layers at 1.2B scale with worse Signal Propagation (i.e. higher input correlations) during training (centre left) have higher feature kurtosis (left). (**Right vs. left two plots**) Introducing a final LN before unembedding causes larger input correlations and feature kurtosis in later layers, even with the OP block. **NB**: y-axes values here are significantly smaller than Fig 4 with Pre-LN.

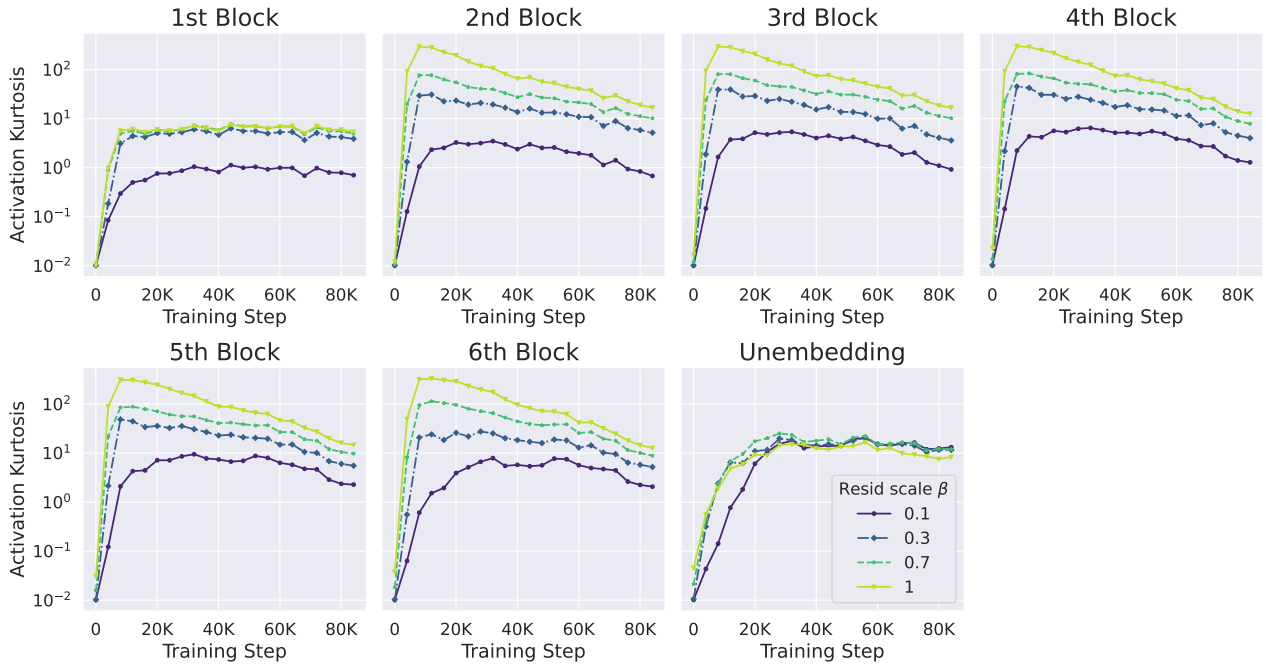


Figure 18. Downweighted residual scalings, $h(x) = x + \beta f(x)$ with $\beta < 1$, reduce OFs at 130M scale. All models are Pre-LN. We downweight both the MLP and Attention residuals.

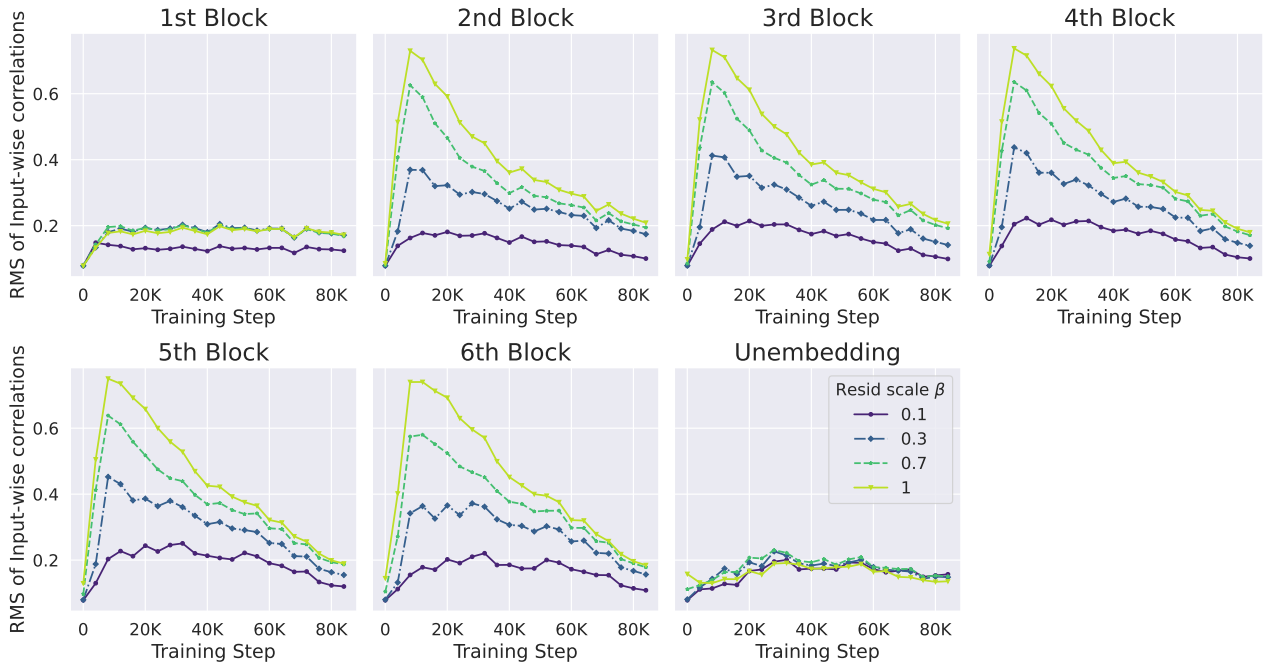


Figure 19. Residual scalings improve Signal Prop at 130M scale. Equivalent to Fig 18.

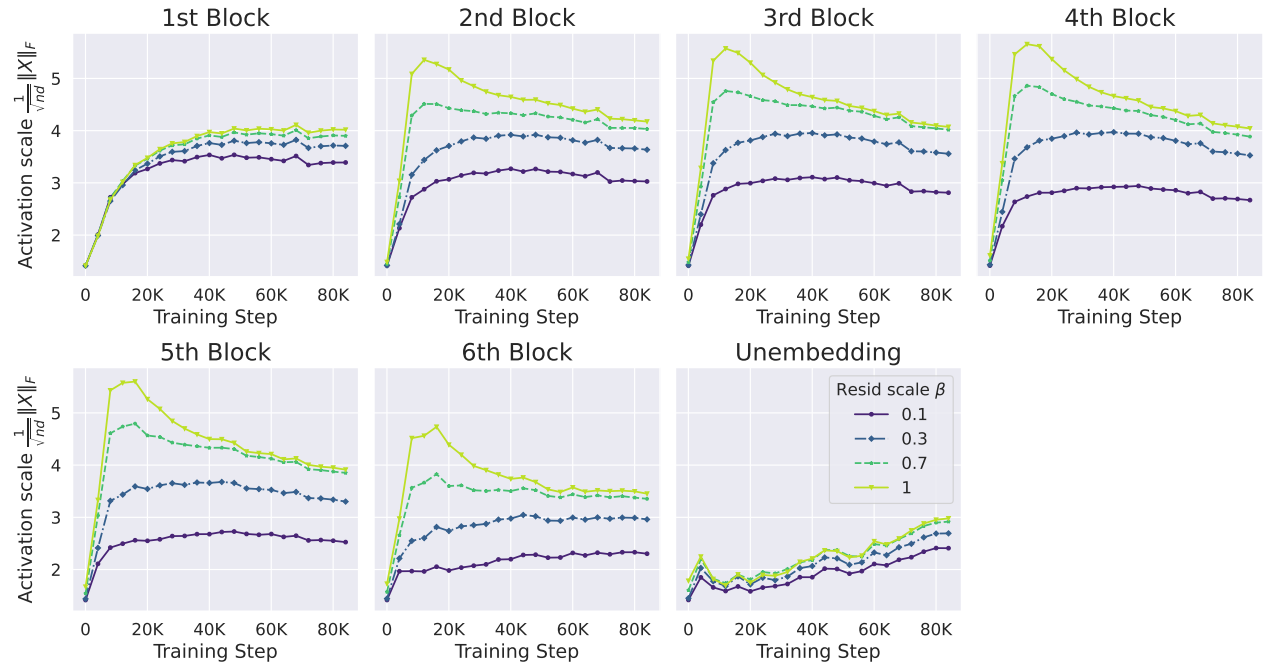


Figure 20. Residual scalings reduce activation scales at 130M scale. Equivalent to Fig 18.

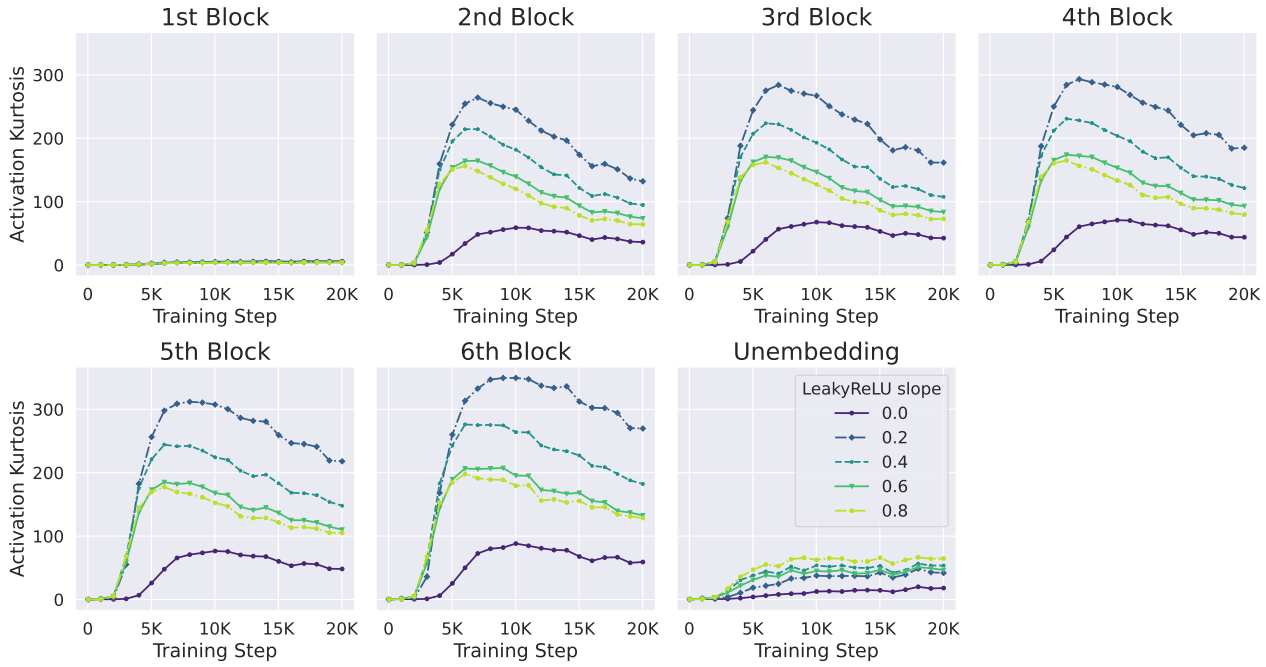


Figure 21. Increasing LeakyReLU slope, s, so that the nonlinearity is more linear mostly improves kurtosis during training, as one might expect from Signal Prop initialisation theory (Zhang et al., 2022a; Li et al., 2022). Here our notation is LeakyReLU$(x) = \max(x, sx)$ for slope $s < 1$. The exception is when the slope is 0, i.e. ReLU, the kurtosis is actually better during training, but this is reflected in the signal propagation during training too (Fig 22). We hypothesise this is because zero neurons get no gradient with ReLU, and this behaves fundamentally differently to a non-zero LeakyReLU slope. The plots show the average over 5 seeds, and we plot the first 20K steps (of 80K). The models are Pre-LN and we downweight the attention residual branch with a factor $\beta = 0.2$ to reduce kurtosis contributions from the attention sub-block, but do not downweight the MLP residual. Note we do not use a log-scaled y-axis to make the differences between LeakyReLU slopes clearer. Experiment is at 130M scale on CodeParrot.

Figure 22. Effect of different LeakyReLU slopes on signal propagation during training, equivalent to Fig 21. Surprisingly, ReLU (i.e. slope 0) has the best signal propagation (lowest input-wise correlations) during training in this setting, even though it has the worst signal prop at initialisation in later layers, compared to all other LeakyReLU variants. This initialisation effect was predicted by Zhang et al. (2022a), but our findings regarding training were previously unknown and require further research.

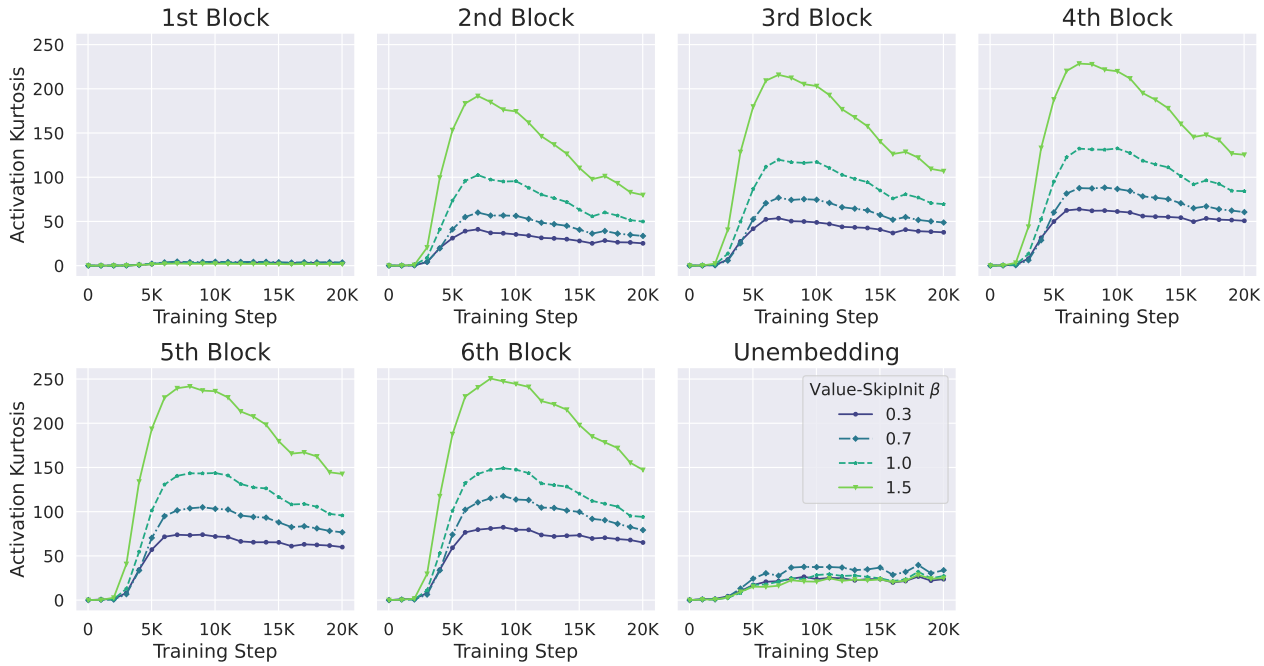


Figure 23. Reducing β in Value-SkipInit (He et al., 2023), which replaces Attention matrix $\mathbf{A} \leftarrow \alpha\mathbf{I} + \beta\mathbf{A}$ and makes attention more identity-like also reduces OFs. We do not train β in Value-SkipInit and fix $\alpha = 1$. The models are Pre-LN and we downweight the MLP residual branch with a factor 0.2 to reduce kurtosis contributions from the MLP sub-block, but do not downweight the attention residual. Each curve is an average over 5 seeds and we plot only the first 20K steps (of 80K). Experiment is at 130M scale on CodeParrot.

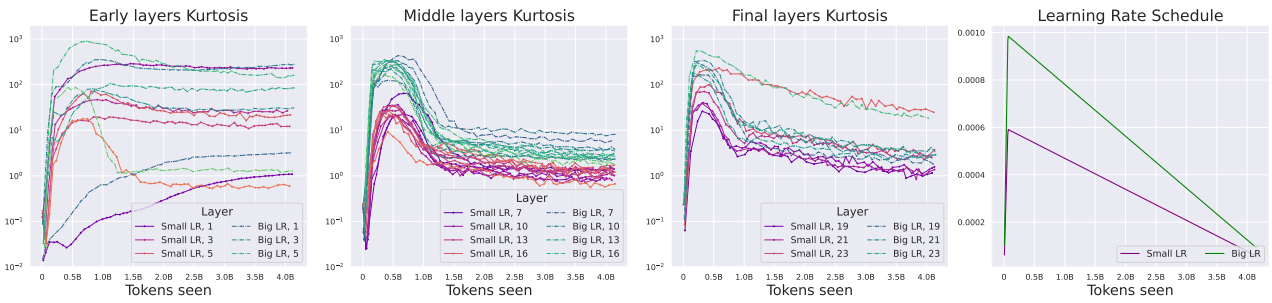


Figure 24. Smaller LR (max value from $0.001 \rightarrow 0.0006$) reduces OFE in a Pre-LN model at 1.2B scale on Languini (Stanić et al., 2023). Models are slightly different from the Pre-LN model in Fig 3 as we do not upweight the input embeddings as described in App G. Still, we do also observe large increases in kurtosis during training, and that a smaller LR reduces this. In this experiment, reducing the max LR to 0.0006 did not impact convergence speed.

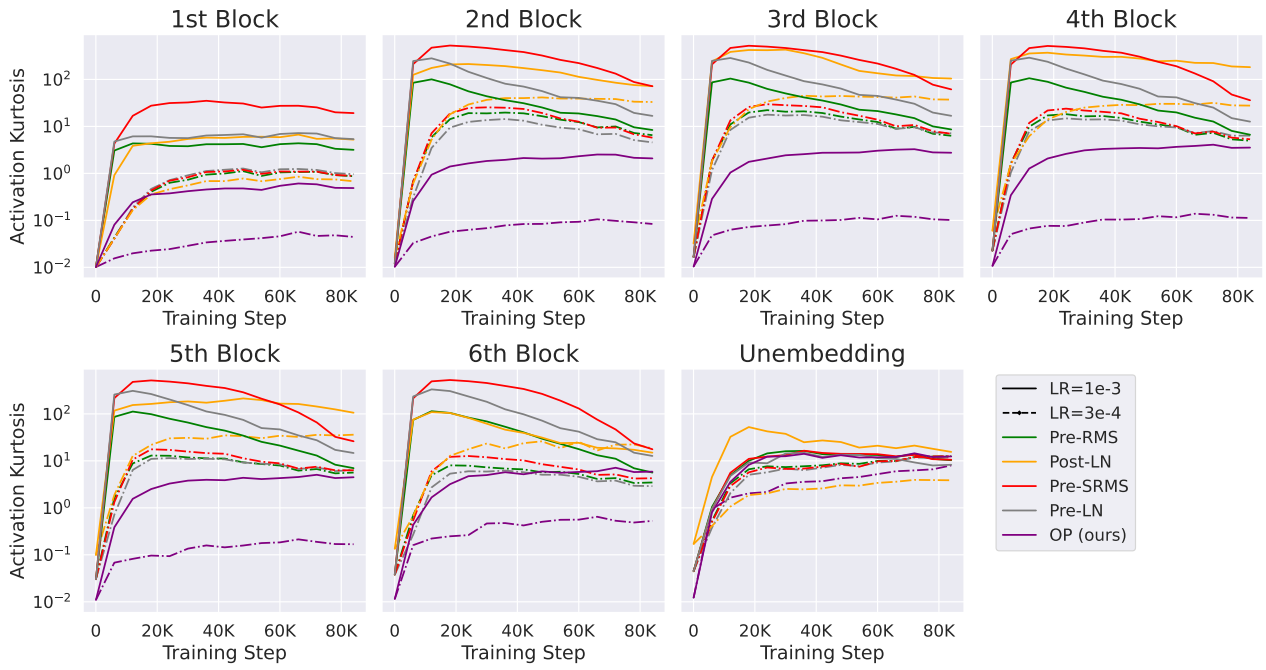


Figure 25. Smaller LRs means reduced OFs, for different Norms and Norm locations. Equivalent of Fig 6, but with all layers. Experiment is on CodeParrot at 130M scale.

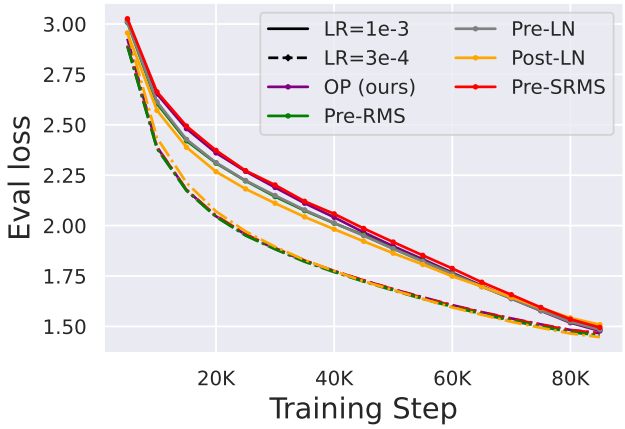


Figure 26. Convergence speed for the runs in Figs 6 and 25 comparing the effect of reduced LRs.

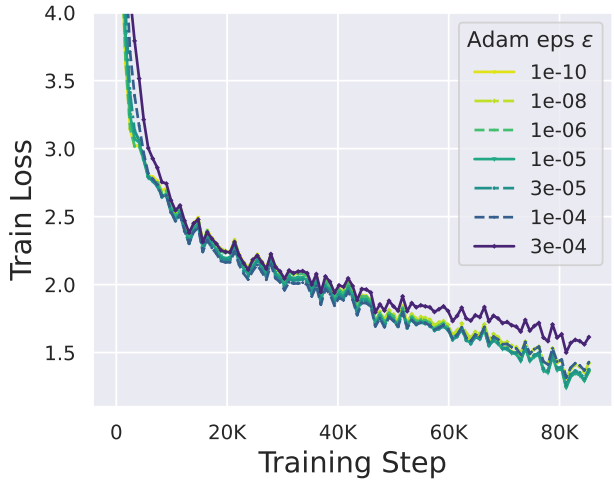


Figure 27. Train loss plot with different Adam epsilon, equivalent to Fig 28. There is not a noticeable difference in convergence speed for $\epsilon < 3e-4$ in this experiment.

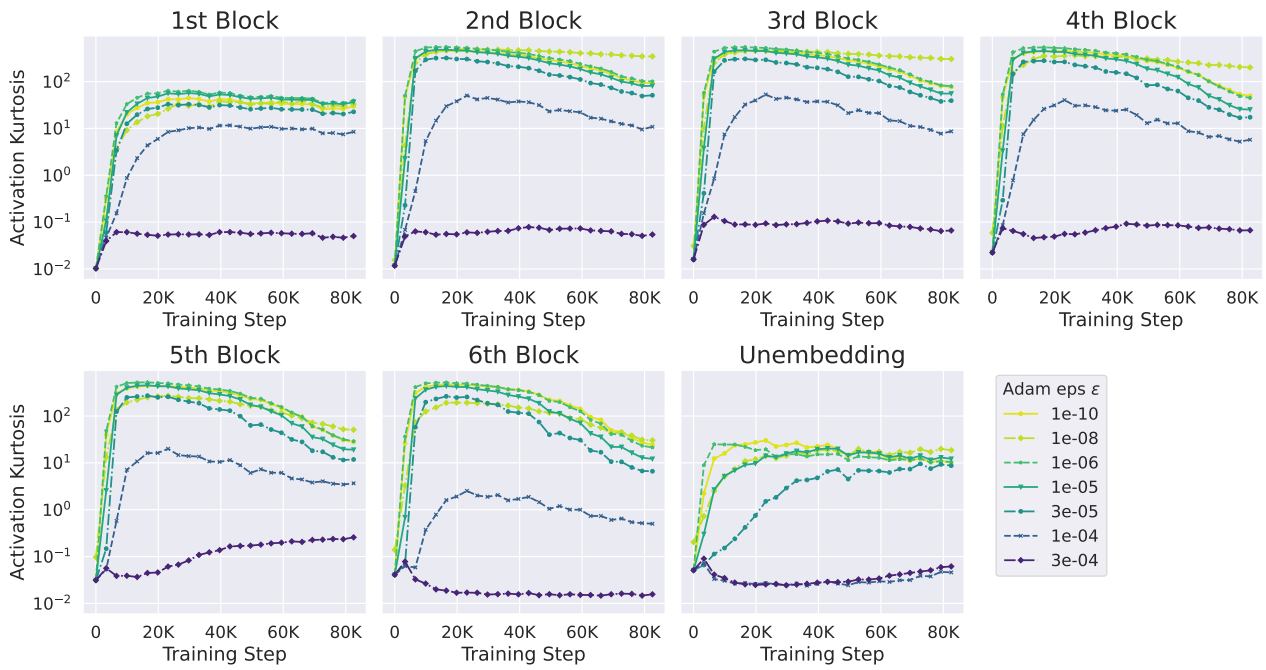


Figure 28. Kurtosis plot with different Adam epsilons on CodeParrot at 130M scale. Each curve is an average over 3 seeds. We see that increasing ϵ from $1e-6$ to $3e-4$ monotonically decreases OFE. At values of ϵ smaller than $1e-6$ there is less of a difference in OFE between different ϵ values.

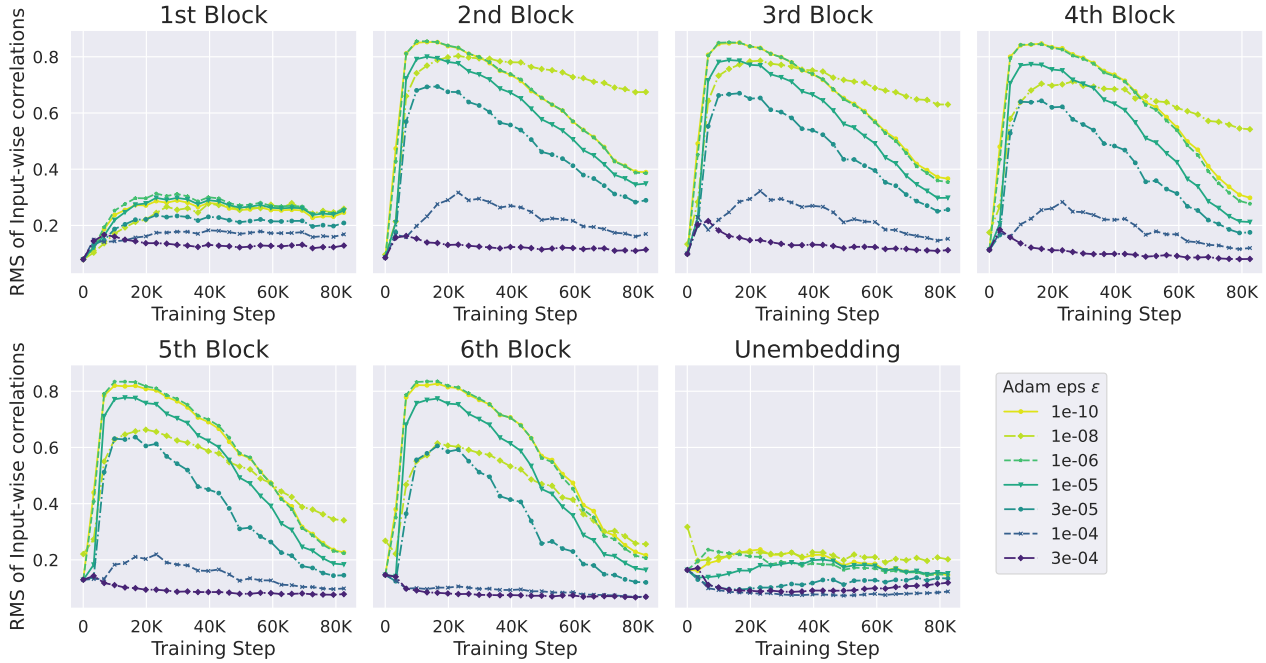


Figure 29. Signal Prop plot with different Adam epsilon. Equivalent of Fig 28.

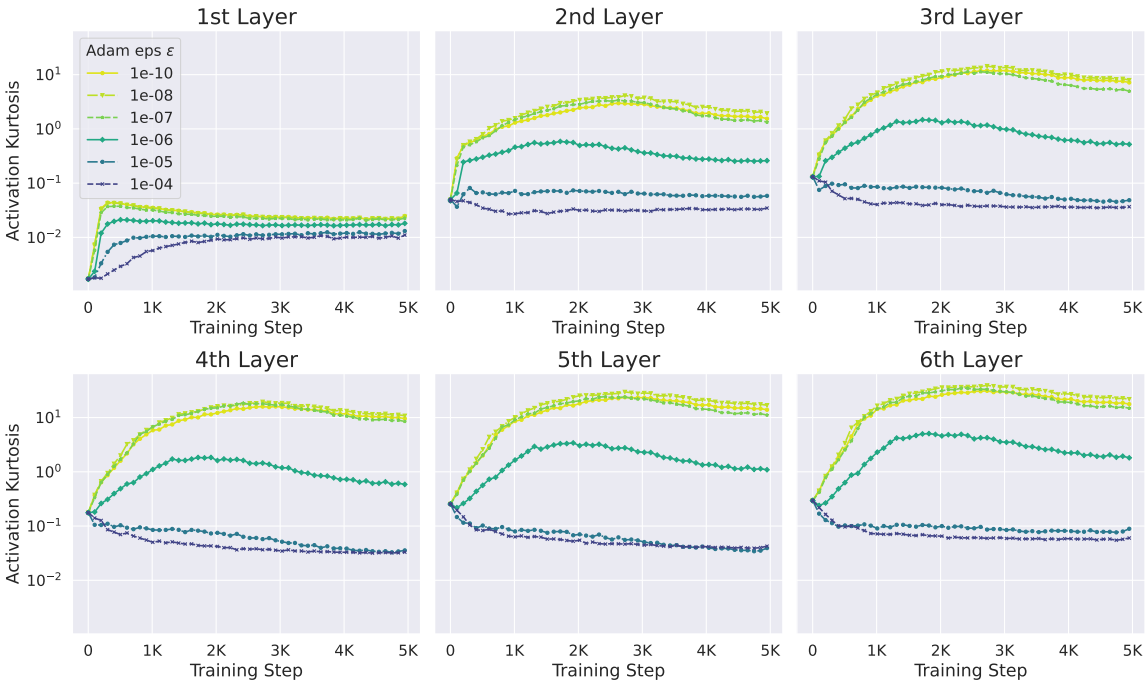


Figure 30. Kurtosis plot with different Adam ϵ with an MLP on CIFAR-10. The model uses Pre-Norm structure with SRMSNorm normalisation. Like in Fig 28, we see that larger ϵ generally leads to smaller OFs.

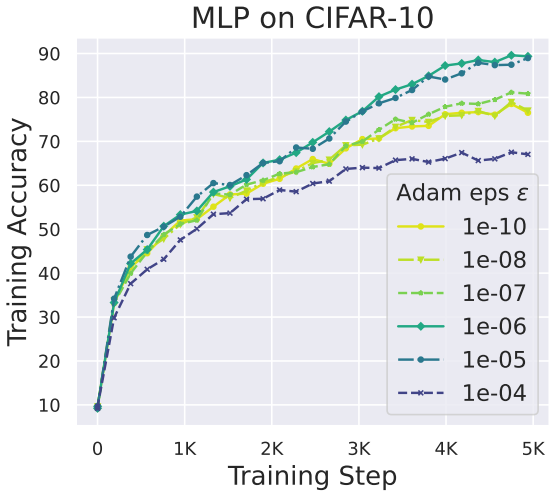


Figure 31. Train accuracy plot with different Adam ϵ of MLP on CIFAR-10, equivalent to Fig 30. In this experiment, milder values of $\epsilon \in \{1e-5, 1e-6\}$ converge fastest.

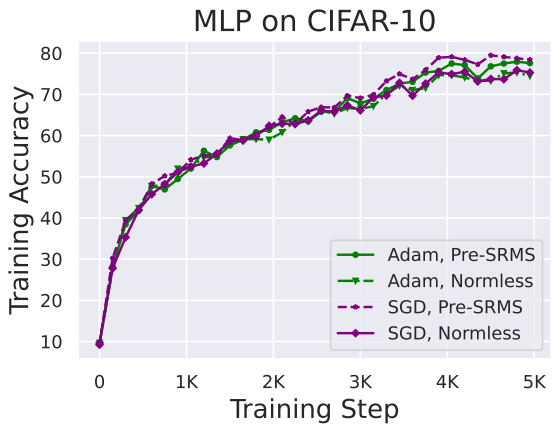


Figure 32. Train accuracy plot with SGD vs Adam of MLP on CIFAR-10, corresponding to Fig 8. Adam ϵ is the default value of $1e-8$.

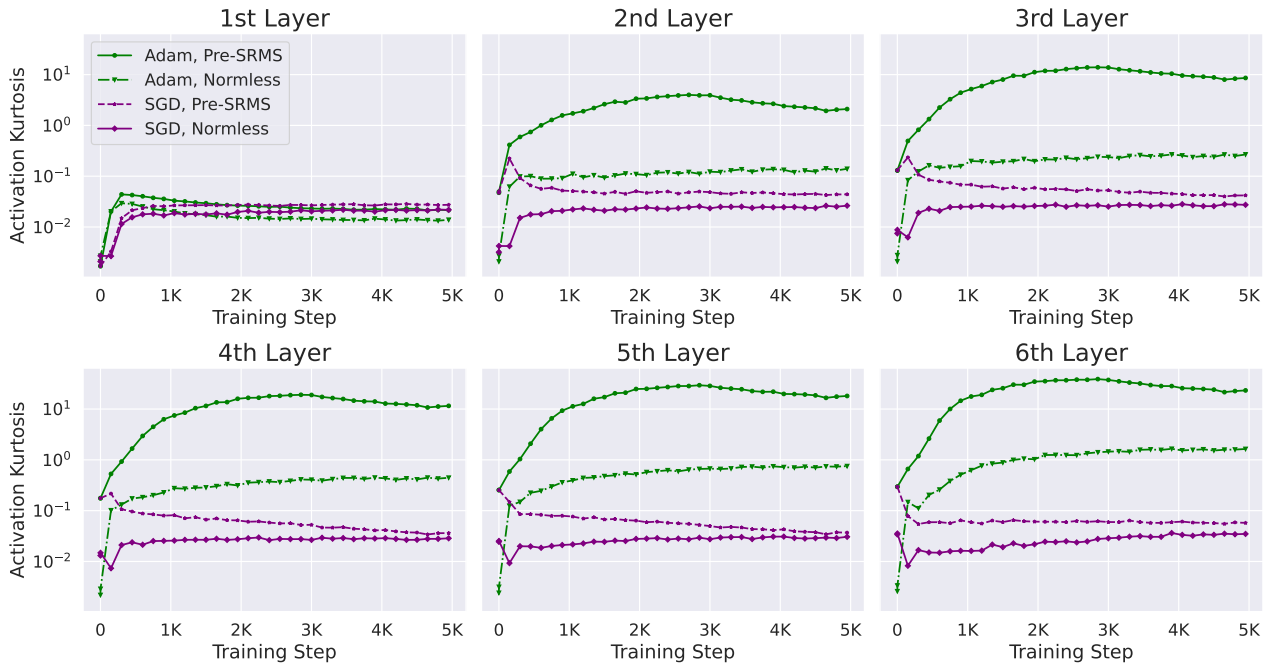


Figure 33. OFs of SGD vs Adam in an MLP on CIFAR-10. Although normalisation layers lead to higher kurtosis for a given optimiser, Adam always has higher OFs than SGD. Fig 8 corresponds to the 6th layer.

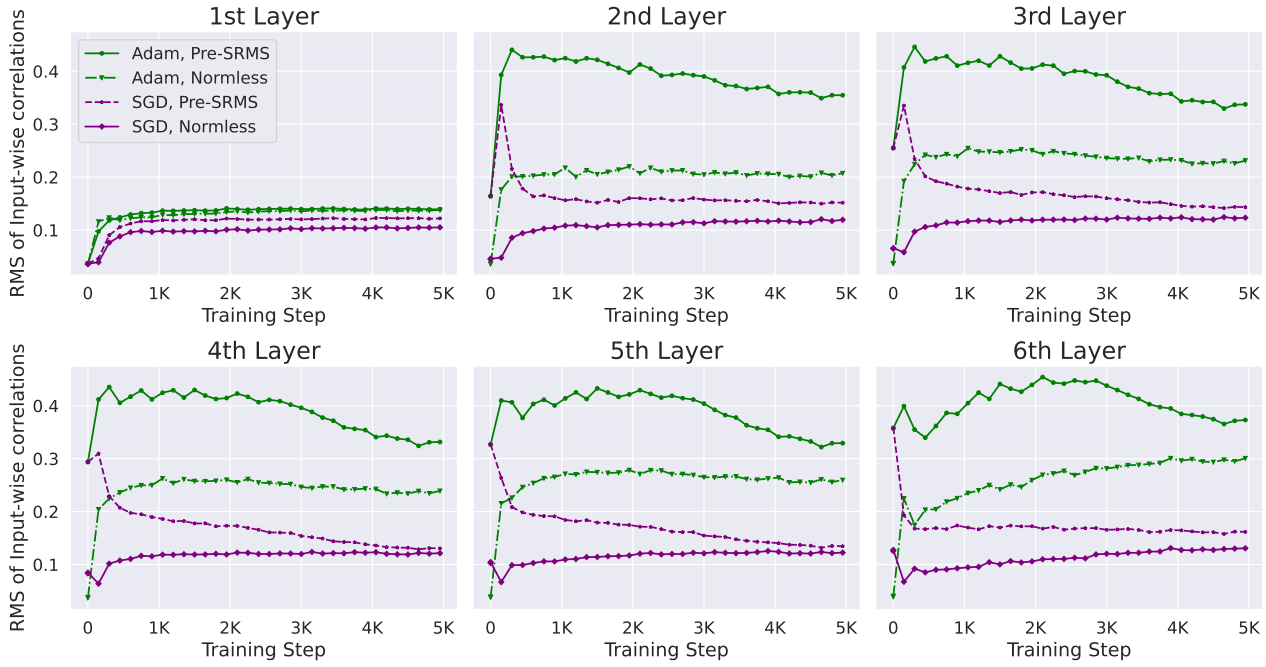


Figure 34. Effect of SGD vs Adam on Signal Prop, for models plotted in Fig 33.

H.1. Ablating the components of the OP block

In Tabs 2 and 3 and Fig 35 we ablate the components of our OP block. Tab 2 assesses the impact of not having an EntReg mechanism on training stability and convergence speed on the Languini dataset (Stanić et al., 2023) at 320M scale. Fig 35 confirms the loss of EntReg causes entropy collapse on CodeParrot at 130M scale, which is shown to lead to unstable training in Fig 36. In these experiments, we also try the tanh thresholding as an alternative EntReg mechanism to QK-Norm. Tab 3 goes from Pre-LN to OP one step at a time, assessing the impact of different norms and downweighted residuals, in terms of OFE.

Table 2. Ablating the convergence and training benefits of the OP block. The asterisk * denotes that training failed without Flash Attention (Dao et al., 2022), which centres pre-softmax logits based on their max value and is therefore more stable. This highlights the training instability of not having some entropy regulating (EntReg) mechanism, where smaller LRs are required for stability. At a smaller (but stable) LR, the naive unnormalised model without EntReg converges much slower (17.4 vs 16.2 ppl) in this example. Even with larger LR, the EntReg mechanism in the OP block improves convergence (16.6 vs 16.2 ppl for QK-RMSNorm) compared to the naive unnormalised model. Tanh thresholding (from Grok-1) also works as an example of an alternative EntReg mechanism to QK-Norm. Because Pre-Norms appear before Query/Key weights, they already provide an implicit EntReg mechanism. As a result, adding EntReg to Pre-Norm models results in only minor changes to convergence speed in this experiment (though ViT-22B shows in other settings Pre-Norm alone is not enough (Dehghani et al., 2023)). Models are 320M parameters, trained also for 3.3B tokens on Languini (Stanić et al., 2023) as in Tab 1.

Model	MLP/Attn Pre-Norm	EntReg	Scaled Residual	LR	Eval PPL
Pre-LN	LN	None	Implicit	1e-3	16.2
Pre-RMSNorm	RMS	None	Implicit	1e-3	16.3
Pre-LN+QK-Norm	LN	QK-RMS	Implicit	1e-3	16.0
Pre-LN+Tanh	LN	Tanh	Implicit	1e-3	16.2
Naive unnormalised	None	None	Yes	3e-4	17.4
Naive unnormalised	None	None	Yes	1e-3	16.6*
OP (QK-Norm)	None	QK-RMS	Yes	1e-3	16.2
OP (Tanh)	None	Tanh	Yes	1e-3	16.4

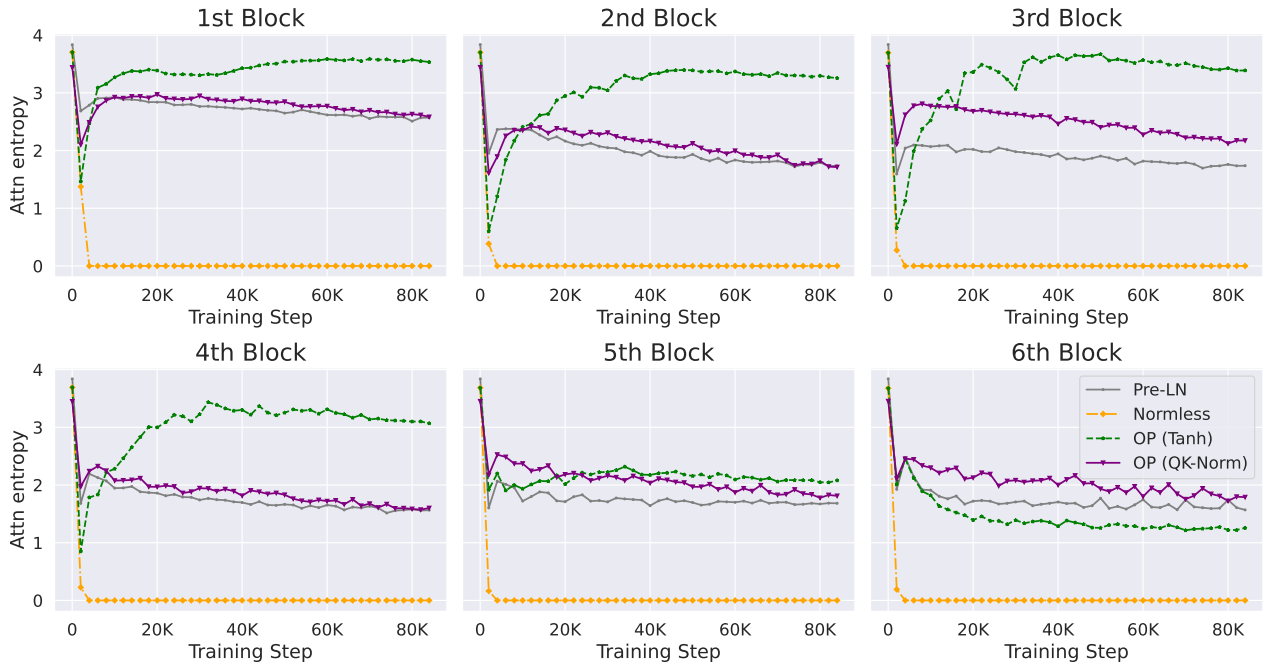


Figure 35. No EntReg leads to entropy collapse without Pre-Norms, which means training fails (as seen in Fig 36).

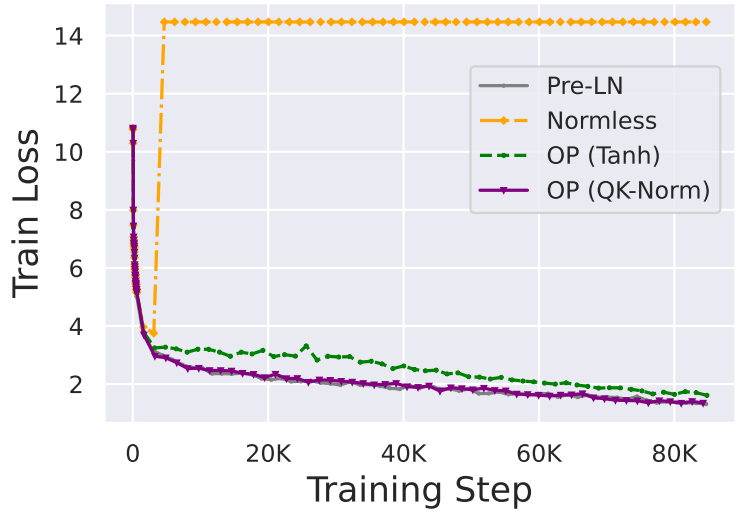


Figure 36. Entropy collapse leads to failed training. OP with tanh does not fail but does converge slower in this setting. Note this is a different task (Code prediction) to language modelling in Tab 2 and we use learnt positional encodings in the input embedding layer, not RoPE, which may account for this difference. We tuned a few values of the max_attn_val hyperparameter with tanh thresholding: $f(x) = max\_attn\_val \cdot \tanh(x/max\_attn\_val)$, which is set by default to 30 in Grok-1, but they did not close the convergence speed loss.

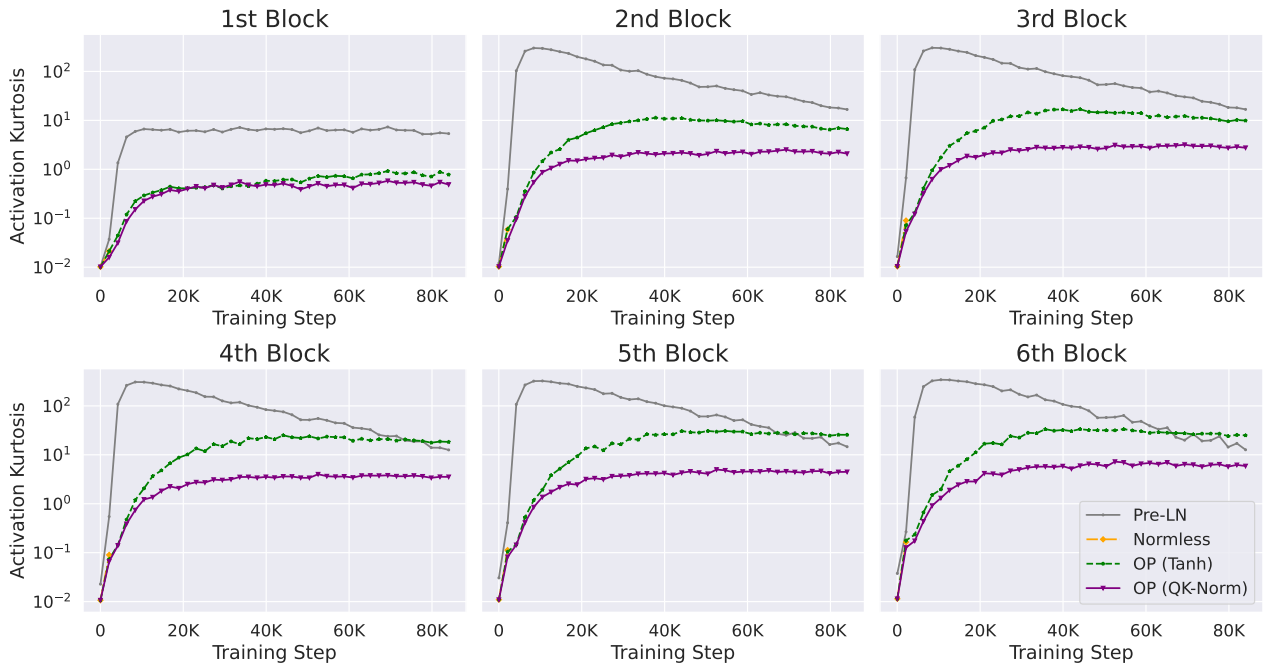


Figure 37. OP with Tanh still has reduced peak OFs compared to Pre-LN. This plot corresponds to the models shown in Fig 35.

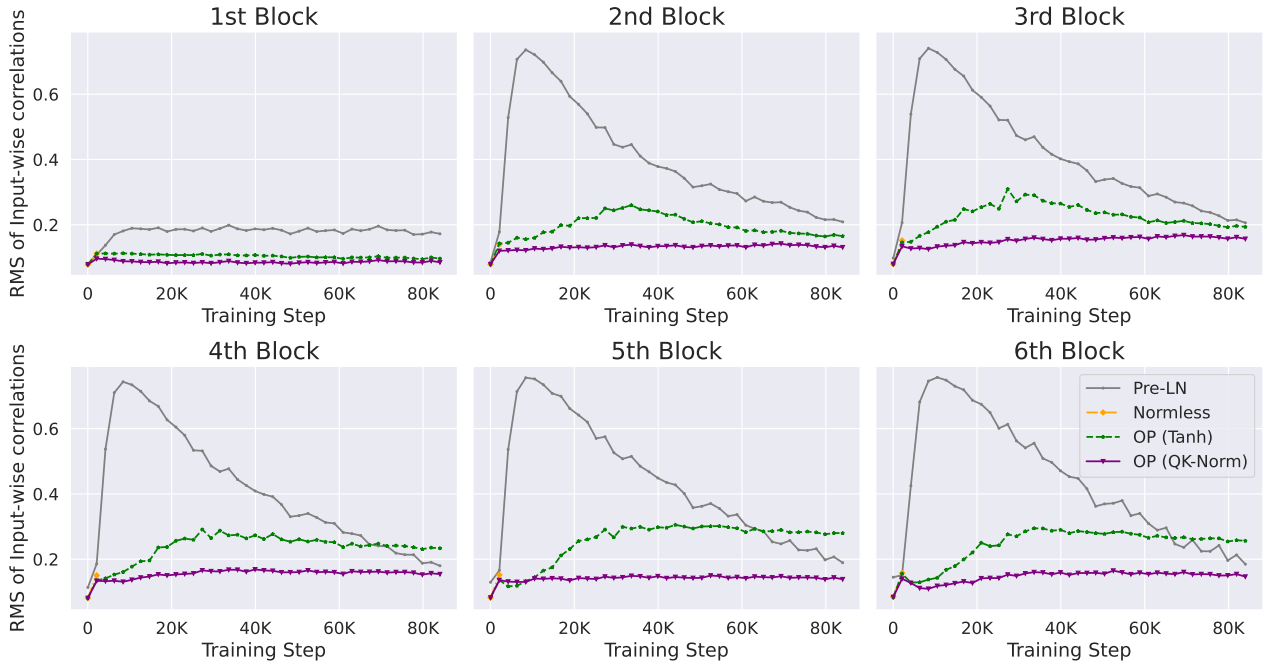


Figure 38. Signal Prop plot with OP Tanh. This plot corresponds to the models shown in Fig 35.

Table 3. Going from Pre-Norm to OP step by step. We remove or add Norms one by one, with different Norm locations depicted in Fig 40. All models trained well (at similar speeds), as they all have some form of entropy regulation (either explicit or implicit) and downweighted residuals. We present the peak Kurtosis (Eq (1)), Signal Propagation (RMS of input-wise correlations), and activation RMS ($\|\mathbf{X}\|_F$) over the training run, with mean and standard deviation over three seeds. We present results where activations $\mathbf{X}$ are the input to the second transformer block. We see that that preventing attention entropy collapse through QK-Norm helps reduce OFs (which we see coincides with improved signal propagation). On the other hand, peak activation RMS does not correlate well as a metric with peak kurtosis, across the different models. In addition, the 2 best models in terms of OFs (our OP and also the third last row, which has no Pre-V or Pre-MLP Norms) are 1-homogeneous (at least at initialisation), which implies that the fact that Pre-V or Pre-MLP Norms make the residual stream scale independent is detrimental for OFE. This is corroborated by Fig 39, which plots the trajectories for the three models (1. Post-QK+Pre-V, 2. QK Norms only and 3. OP) that achieved peak kurtosis lower than 10. Fig 39 shows that the non-homogeneity (due to a Pre-V Norm) leads to a large initial increase in kurtosis and signal propagation in this setting, like we consistently see with Pre-Norm blocks e.g. Fig 4. Models are 130M scale on CodeParrot.

Model	Norm				Scaled Resid	Homog.?	Act RMS	Signal Prop	Kurtosis
	Post-QK	Pre-QK	Pre-V	Pre-MLP					
Pre-RMS	None	RMS	RMS	RMS	Implicit	No	$5.45_{\pm 0.13}$	$0.72_{\pm 0.03}$	$131.8_{\pm 21.2}$
Scaled Resids	None	RMS	RMS	RMS	Yes	No	$3.97_{\pm 0.09}$	$0.47_{\pm 0.04}$	$46.4_{\pm 14.0}$
All Norms	RMS	RMS	RMS	RMS	Yes	No	$3.92_{\pm 0.07}$	$0.24_{\pm 0.05}$	$12.7_{\pm 10.2}$
Attn Norms only	RMS	RMS	RMS	None	Yes	No	$4.38_{\pm 0.07}$	$0.29_{\pm 0.04}$	$11.8_{\pm 8.03}$
Post-QK+Pre-V	RMS	None	RMS	None	Yes	No	$4.40_{\pm 0.06}$	$0.27_{\pm 0.01}$	$6.4_{\pm 1.32}$
QK Norms only	RMS	RMS	None	None	Yes	Yes	$4.32_{\pm 0.06}$	$0.15_{\pm 0.01}$	$2.5_{\pm 0.93}$
Pre-QK only	None	RMS	None	None	Yes	Yes	$4.38_{\pm 0.01}$	$0.37_{\pm 0.05}$	$64.0_{\pm 49.5}$
OP (ours)	RMS	None	None	None	Yes	Yes	$4.46_{\pm 0.09}$	$0.17_{\pm 0.01}$	$4.3_{\pm 1.49}$

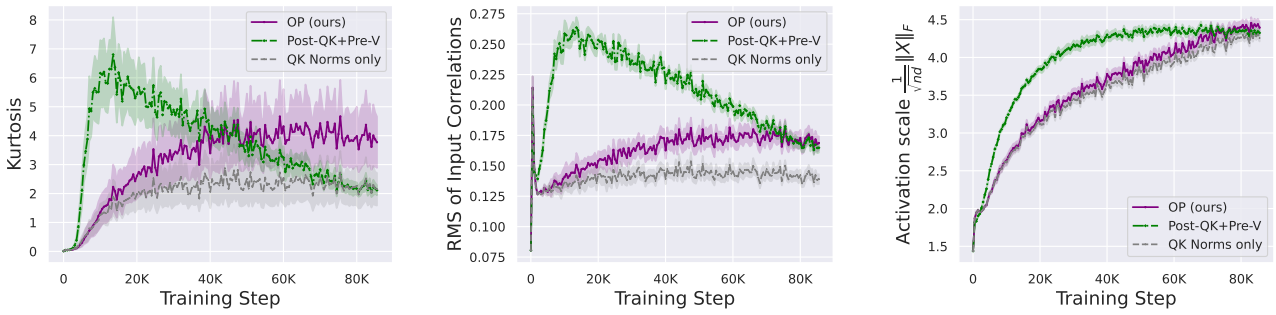


Figure 39. Training trajectories of kurtosis, signal propagation and activation scales for the three best configurations in Tab 3. The setting with Pre-V Norm (which is not 1-homogeneous) sees a large initial increase in all metrics, with kurtosis and input correlations peaking within 10K steps before reducing during training.

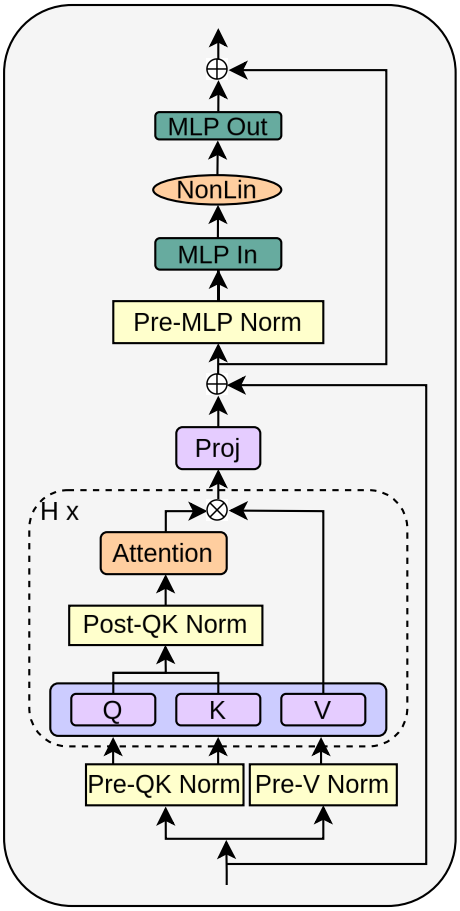


Figure 40. A transformer block with many different Norm layers depicted, to help parse the ablations we consider in Tab 3. Note we break down the standard attention Pre-Norm into Pre-QK Norm and Pre-V Norm because removal of Pre-V Norm makes the attention sub-block homogeneous (i.e. $f(x)$ is homogeneous if $f(kx) = kf(x)$ for some scalar $k > 0$), hence acts differently to Pre-QK Norm, which acts as an implicit regulator for attention entropy.

I. Orders of Activation Updates for Kurtosis

To better appreciate the effect of different optimiser hyperparameters on OFs, we now consider how the updates that arise during training to a representation matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ can lead to increasing kurtosis (and OFs). In general, a training step (e.g. with a gradient/Adam update on trainable parameters earlier in the forward pass than $\mathbf{X}$) will lead to an update $\mathbf{X} \leftarrow \mathbf{X} + \Delta^{\mathbf{X}}$.

Recall that Kurt$(\mathbf{X})$ is an defined through comparing the fourth $m_4(\mathbf{X})$ and second $m_2(\mathbf{X})$ moments of neuron RMS $\sqrt{\frac{1}{n}\sum_{\alpha=1}^{n} \mathbf{X}_{\alpha,j}^2}$ for different j. As such, it is natural to ask how updating $\mathbf{X} \leftarrow \mathbf{X} + \Delta^{\mathbf{X}}$ updates these moment statistics. We first study the second moment update u_2:

$$u_2 \stackrel{\text{def}}{=} m_2(\mathbf{X} + \Delta^{\mathbf{X}}) - m_2(\mathbf{X}) = \frac{1}{d}\sum_{j=1}^{d}\left(\frac{1}{n}\sum_{\alpha=1}^{n}(\mathbf{X} + \Delta^{\mathbf{X}})_{\alpha,j}^2\right) - \frac{1}{d}\sum_{j=1}^{d}\left(\frac{1}{n}\sum_{\alpha=1}^{n}\mathbf{X}_{\alpha,j}^2\right) \tag{6}$$

$$= \frac{1}{nd}\big(u_{2,1} + u_{2,2}\big), \quad \text{with} \tag{7}$$

$$u_{2,1} \stackrel{\text{def}}{=} \sum_{j=1}^{d}\sum_{\alpha=1}^{n} 2\Delta_{\alpha,j}^{\mathbf{X}}\mathbf{X}_{\alpha,j}, \quad u_{2,2} \stackrel{\text{def}}{=} \sum_{j=1}^{d}\sum_{\alpha=1}^{n}(\Delta_{\alpha,j}^{\mathbf{X}})^2. \tag{8}$$

Likewise for the fourth moment update u_4:

$$u_4 \stackrel{\text{def}}{=} m_4(\mathbf{X} + \Delta^{\mathbf{X}}) - m_4(\mathbf{X}) = \frac{1}{d}\sum_{j=1}^{d}\left(\frac{1}{n}\sum_{\alpha=1}^{n}(\mathbf{X} + \Delta^{\mathbf{X}})_{\alpha,j}^2\right)^2 - \frac{1}{d}\sum_{j=1}^{d}\left(\frac{1}{n}\sum_{\alpha=1}^{n}\mathbf{X}_{\alpha,j}^2\right)^2 \tag{9}$$

$$= \frac{1}{n^2 d}\big(u_{4,1} + u_{4,2} + u_{4,3} + u_{4,4}\big), \quad \text{with} \tag{10}$$

$$u_{4,1} \stackrel{\text{def}}{=} \sum_{j=1}^{d}\sum_{\alpha,\beta=1}^{n} 4\Delta_{\alpha,j}^{\mathbf{X}}\mathbf{X}_{\alpha,j}\mathbf{X}_{\beta,j}^2, \quad u_{4,2} \stackrel{\text{def}}{=} \sum_{j=1}^{d}\sum_{\alpha,\beta=1}^{n} 2(\Delta_{\alpha,j}^{\mathbf{X}})^2\mathbf{X}_{\beta,j}^2 + 4\Delta_{\alpha,j}^{\mathbf{X}}\mathbf{X}_{\alpha,j}\Delta_{\beta,j}^{\mathbf{X}}\mathbf{X}_{\beta,j}, \tag{11}$$

$$u_{4,3} \stackrel{\text{def}}{=} \sum_{j=1}^{d}\sum_{\alpha,\beta=1}^{n} 4\mathbf{X}_{\alpha,j}\Delta_{\alpha,j}^{\mathbf{X}}(\Delta_{\beta,j}^{\mathbf{X}})^2, \quad u_{4,4} \stackrel{\text{def}}{=} \sum_{j=1}^{d}\sum_{\alpha,\beta=1}^{n}(\Delta_{\alpha,j}^{\mathbf{X}})^2(\Delta_{\beta,j}^{\mathbf{X}})^2. \tag{12}$$

Above, we have broken down the p^{th} moment update u_p into $(u_{p,l})_l$, where $u_{p,l}$ denotes the contribution to u_p that is order l in $\Delta^{\mathbf{X}}$. The reason for this is that, typically, a learning rate parameter η is used such $\Delta^{\mathbf{X}}$ is linear in η, and so $u_{p,l}$ is order l in η.[8] Usually, η is chosen to be small such that $\Delta^{\mathbf{X}}$ is small elementwise relative to $\mathbf{X}$. Note that the quadratic update terms $u_{p,2}$ are always positive,[9] whereas the linear terms $u_{p,1}$ are not necessarily positive, so we might expect quadratic terms to drive any increase in the p^{th} moment m_p.

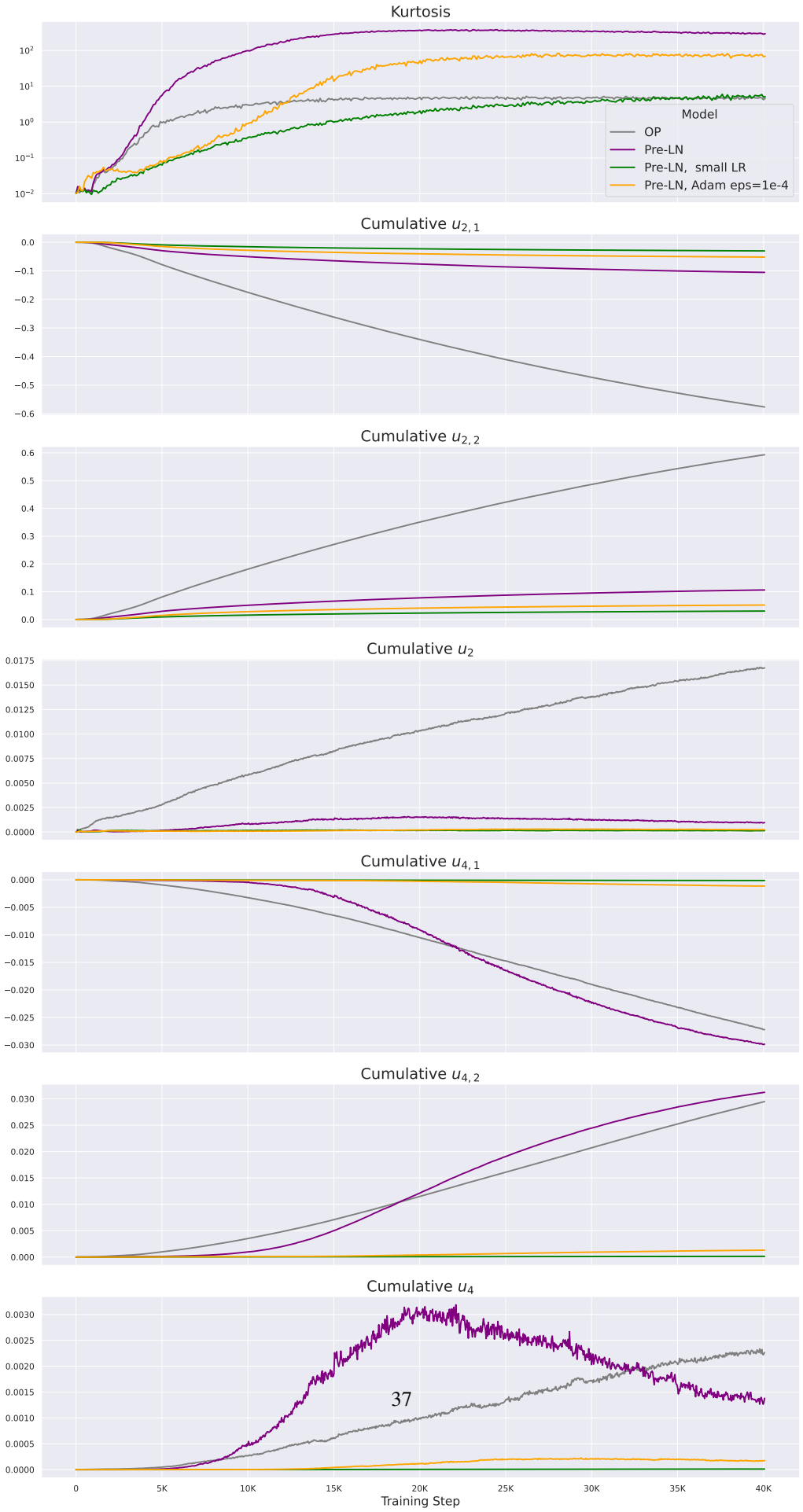
In Fig 41, we plot the cumulative sum of these $(u_{p,l})$ terms, for our OP block, a default Pre-LN block, and also two modifications that reduce OFs in Pre-LN (increasing Adam epsilon from $1e-8$ to $1e-4$ and also reducing maximum LR from $1e-3$ to $3e-4$) trained on CodeParrot. We see indeed that the cumulative $u_{4,2}$ quadratic term dominates the update to u_4 and drives the increase in m_4 in the default Pre-LN model. Both reducing LR and also increasing Adam ϵ reduce this term, which also reduces the growth in fourth moment and kurtosis. In particular, in the small LR $\eta \to 0$ limit the linear first order term $u_{4,1}$ will dominate and the effect of quadratic $u_{4,2}$ can be ignored. The impact of sub-leading order terms like $u_{4,2}$ in OFE is related to the discretisation drift between discrete-time gradient descent and continuous-time gradient flow (Rosca, 2023). Fig 42 plots the non-cumulative version of Fig 41.

On the other hand, in Fig 41 the OP block has a large increase in u_4 that is matched by a large increase in u_2, which means the kurtosis (which is the ratio m_4/m_2^2) does not increase as much as Pre-LN. Fig 43 shows that $u_{4,2}$ dominates the cubic $u_{4,3}$ and quartic $u_{4,4}$ update terms to the fourth moment, so we can focus on studying $u_{4,2}$. We plot the moment updates for the input to the second attention block (out of six).

[8] For example, if we have $\mathbf{X} = \mathbf{H}\mathbf{W}$ for a previous layer $\mathbf{H}$ that is fixed (e.g. embedding layer in a transformer). Then we usually update weights $\mathbf{W} + \Delta^{\mathbf{W}}$ linearly in η, and so $\Delta^{\mathbf{X}} = \mathbf{H}\Delta^{\mathbf{W}}$ is also linear in η. For other layers we need to consider the change in $\mathbf{H}$ too, but this will also be linear in η to leading order.

[9] This is straightforward to see for $u_{2,2}$. For $u_{4,2}$ the second summand can be factorised as $\sum_j \left(\sum_\alpha \mathbf{X}_{\alpha,j}\Delta_{\alpha,j}\right)^2$ which is positive.

The models presented in Figs 41 to 43 were trained using Adam without momentum, akin to RMSProp (Tieleman & Hinton, 2012): we set $\beta_1 = 0$ and $\beta_2 = 0.95$ in Adam. The reason for this was to separate out the contribution of individual training steps on the kurtosis updates. If instead we re-introduce momentum with $\beta_1 = 0.9$, then the different update steps become mixed and the leading order $u_{4,1}$ dominates the updates to the kurtosis for the Pre-LN model, as seen in Fig 44.


Kurtosis
Model
OP
Pre-LN
Pre-LN, small LR
Pre-LN, Adam eps=1e-4
Cumulative $u_{2,1}$
Cumulative $u_{2,2}$
Cumulative u_2
Cumulative $u_{4,1}$
Cumulative $u_{4,2}$
Cumulative u_4
Training Step

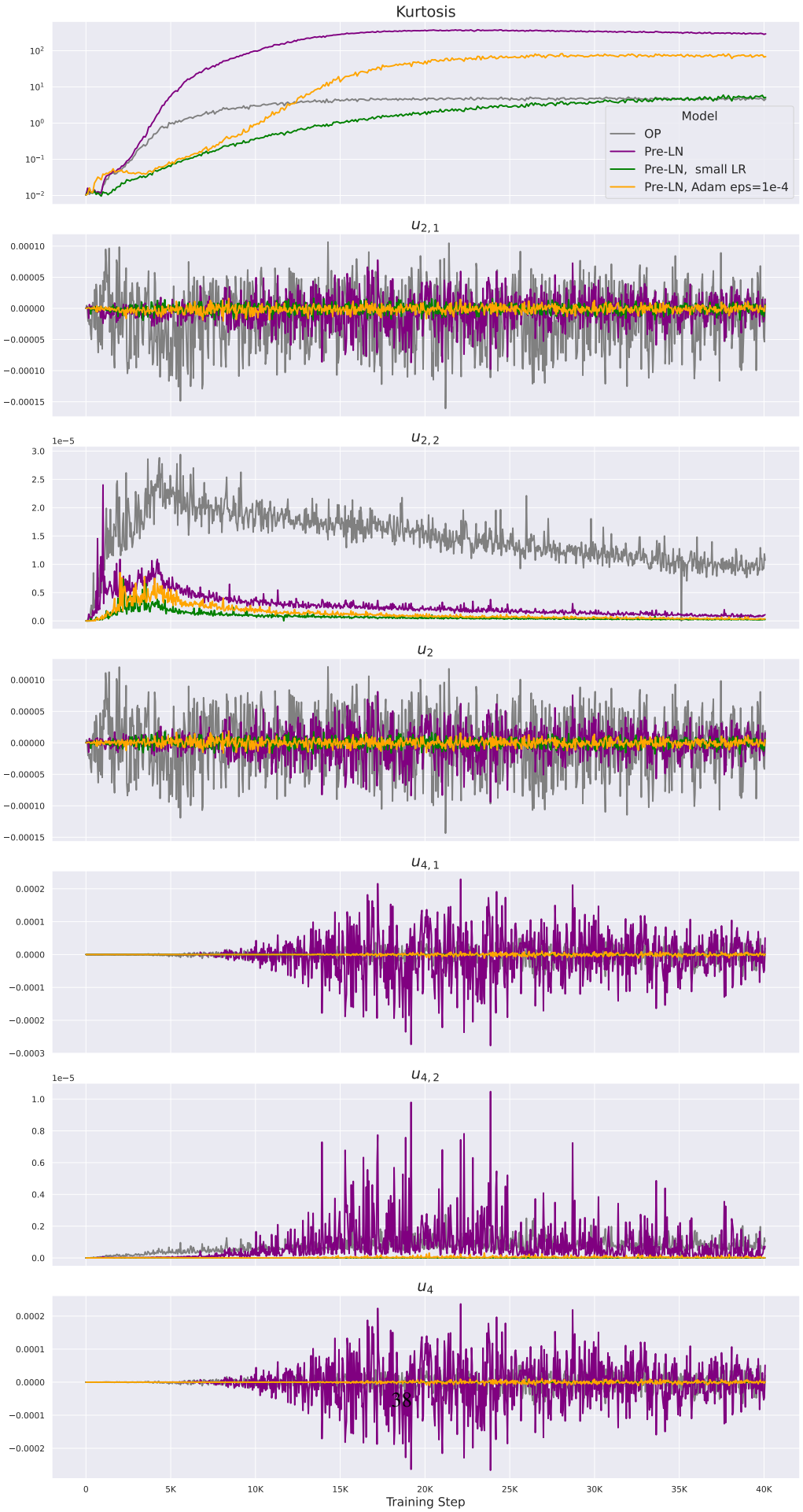
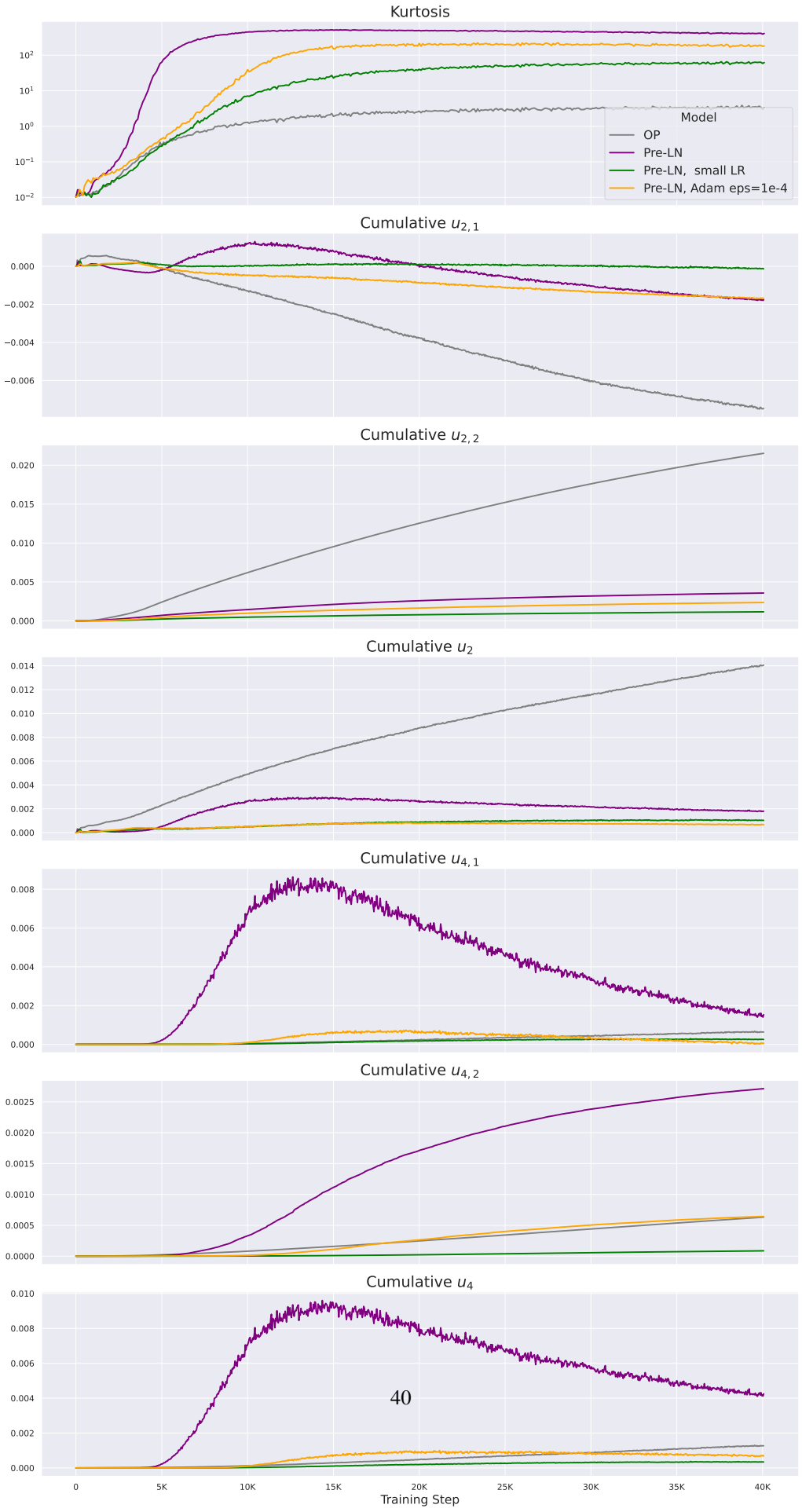

Kurtosis
Model
OP
Pre-LN
Pre-LN, small LR
Pre-LN, Adam eps=1e-4
$u_{2,1}$
$u_{2,2}$
u_2
$u_{4,1}$
$u_{4,2}$
u_4
1e−5
Training Step

Figure 43. Sub-leading order terms are dominated by $u_{4,2}$.


Kurtosis
Model
OP
Pre-LN
Pre-LN, small LR
Pre-LN, Adam eps=1e-4
Cumulative $u_{2,1}$
Cumulative $u_{2,2}$
Cumulative u_2
Cumulative $u_{4,1}$
Cumulative $u_{4,2}$
Cumulative u_4
Training Step

J. Worse Signal Prop Means Higher Activation Kurtosis in Gaussian Features

Proposition J.1 (Bad Signal Propagation implies higher kurtosis for Gaussian features)**.** *Suppose we have $\mathbf{X} \in \mathbb{R}^{n \times d}$ zero-mean Gaussian distributed with all inputs uniformly correlated with some $\rho > 0$, and independent features (across columns). That is: $\mathbb{E}[\mathbf{X}] = \mathbf{0}$ and $\mathbb{E}[\mathbf{X}_{\alpha,j}\mathbf{X}_{\beta,k}] = \rho \cdot \mathbf{1}\{j = k\} + (1-\rho) \cdot \mathbf{1}\{j = k\} \cdot \mathbf{1}\{\alpha = \beta\}$.*[10]

Then, if we consider the feature-wise Gram matrix $\Sigma_F = \frac{1}{n}\mathbf{X}^\top\mathbf{X}$, we have that the expected squared diagonal entry of Σ_F is $\mathbb{E}[(\Sigma_F)^2_{1,1}] = 1 + 2\rho^2 + o_n(1)$ increases as ρ increases, whereas the expected diagonal entry is $\mathbb{E}[(\Sigma_F)_{1,1}] = 1$ is independent of ρ.

Proof. As Gaussians are determined by their first two moments, let us suppose that $\mathbf{X}_{\alpha,j} = \sqrt{1-\rho}u_{\alpha,j} + \sqrt{\rho}v_j$, where $(u_{\alpha,j})_{\alpha,j}$ and $(v_j)_j$ are independent standard Gaussians. Then, for two neuron indices $k, l \leq d$ we have:

$$\left(\mathbf{X}^T\mathbf{X}\right)_{k,l} = (1-\rho)\sum_{\alpha \leq n} u_{\alpha,k}u_{\alpha,l} \tag{13}$$

$$+ \rho n v_k v_l \tag{14}$$

$$+ \sqrt{\rho(1-\rho)}\sum_{\alpha \leq n} u_{\alpha,k}v_k + u_{\alpha,l}v_l. \tag{15}$$

We are interested in the diagonal elements of $\Sigma_{\text{F}} = \frac{1}{n}\mathbf{X}^\top\mathbf{X}$, when $k = l$ above. In this case, we have $(u^2_{\alpha,k})_\alpha$ and v_k^2 are all independent chi-squared χ^2 distributed with 1 degree of freedom. For $Z \sim \chi^2_1$, we have $\mathbb{E}[Z] = 1$ and $\mathbb{E}[Z^2] = 3$.

For the first moment, we take the expectation above and note that the summands of Eq (15) are products of independent zero-mean Gaussians (so zero mean). This gives $\mathbb{E}[\mathbf{X}^T\mathbf{X}_{k,k}] = n$ and hence $\mathbb{E}[(\Sigma_{\text{F}})_{1,1}] = 1$, as required.

For the second moment, we note that all cross products in $\left(\mathbf{X}^T\mathbf{X}\right)^2_{k,k}$ will disappear in expectation when we square besides the one involving Eqs (13) and (14), as both terms will be χ^2_1 distributed (hence not zero-mean). On the other hand, all cross products involving Eq (15) will be an odd order in at least one zero-mean independent Gaussian (hence zero-mean).

The square of Eq (13) is $(1-\rho)^2 n(n+2)$ in expectation, which can be seen by the fact that $\sum_{\alpha \leq n} u^2_{\alpha,k}$ is actually a χ^2_n distribution, with mean n and variance $2n$. Hence for $Z \sim \chi^2_n$, we have $\mathbb{E}[Z^2] = \mathbb{E}[Z]^2 + \text{Var}(Z) = n^2 + 2n$.

The square of Eq (14) is $3\rho^2 n^2$ in expectation, again by properties of χ^2_1 random variables.

The square of Eq (15) is $O(n)$ (in fact $4\rho(1-\rho)n$) in expectation and will be dominated by the $O(n^2)$ terms. To see this, we note that Eq (15) is a sum of n zero mean i.i.d. random variables, so one can use the additive property of variances for independent random variables.

Finally, the cross term between Eqs (13) and (14) is $2\rho(1-\rho)n^2$ in mean. One factor of n comes from the sum of inputs $\alpha \leq n$ and the other comes from Eq (14) already. The product of two independent χ^2_1 random variables is 1 in expectation.

Putting this all together, we have

$$\mathbb{E}[\mathbf{X}^T\mathbf{X}^2_{k,k}] = (1-\rho)^2 n(n+2) + 3\rho^2 n^2 + 4\rho(1-\rho)n + 2\rho(1-\rho)n^2 \tag{16}$$

$$= \left((1-\rho)^2 + 3\rho^2 + 2\rho - 2\rho^2\right)n^2 + O(n) \tag{17}$$

$$= (1 + 2\rho^2)n^2 + O(n) \tag{18}$$

As $\Sigma_{\text{F}} = \frac{1}{n}\mathbf{X}^T\mathbf{X}$, we divide Eq (16) by n^2, and obtain our desired result. □

Above, we note that $\mathbb{E}[(\Sigma_{\text{F}})^2_{1,1}]$ is equivalent to the fourth moment m_4 in our feature-wise kurtosis definition Eq (1), while $\mathbb{E}[(\Sigma_{\text{F}})_{1,1}]$ corresponds to the second moment m_2. Hence, Prop J.1 demonstrates that worse signal propagation (in terms of higher ρ) leads to higher kurtosis.

[10] Note this covariance gives a "uniform" correlation structure $\mathbb{E}[\frac{1}{d}\mathbf{X}\mathbf{X}^\top] = (1-\rho)\mathbf{I}_n + \rho\mathbf{1}_n\mathbf{1}_n^\top$, which has been studied before in Noci et al. (2022); He et al. (2023) as a way to study signal propagation in sequences. Rank collapse (Dong et al., 2021) is when $\rho = 1$.

We note that the result is restricted to a Gaussian setting with independent features. This is an accurate description of large-width NN initialisation (Matthews et al., 2018; Lee et al., 2018; Yang, 2019), but does not capture training dynamics as we discuss in the main paper. Indeed, the maximum kurtosis $(1 + 2\rho^2)$ is 3 when $\rho = 1$, whereas in our experiments we obtain much higher values during training (and the maximum is the width d, which is considerably larger than 3 in practice). This represents a gap in our theoretical understanding and practice, which we leave for future study.