Investigating Neuron Ablation in Attention Heads: The Case for Peak Activation Centering

Nicholas Pochinkov[1], Ben Pasero[2] and Skylar Shibayama[2]

[1]*Independent. Dublin, Ireland*
[2]*Independent. Seattle, WA, USA*

Abstract

The use of transformer-based models is growing rapidly throughout society. With this growth, it is important to understand how they work, and in particular, how the attention mechanisms represent concepts. Though there are many interpretability methods, many look at models through their neuronal activations, which are poorly understood. We describe different lenses though which to view neuron activations, and investigate the effectiveness in language models and vision transformers though various methods of neural ablation: zero ablation, mean ablation, activation resampling, and a novel approach we term 'peak ablation'. Through experimental analysis, we find that in different regimes and models, each method can offer the lowest degradation of model performance compared to other methods, with resampling usually causing the most significant performance deterioration. We make our code available at https://github.com/nickypro/investigating-ablation

Keywords

AI, LLMs, Transformers, Interpretability, Attention, Pruning.

1. Introduction

Understanding how language models make decisions is important to ensure that their use can be trusted. Mechanistic interpretability offers one lens through which to understand how transformer architecture models [1] perform the computations required to get an output. An oft-used tool in mechanistic interpretability is to attribute individual network parts to specific capabilities by ablating those parts and observing capability degradation.

However, choosing how to ablate neurons in language models is still an unsolved problem. The traditional closed-form methods are zero ablation and mean ablation [2, 3], as well as an additional, more randomised method of activation resampling in the case of causal scrubbing [3, 4], but little empirical analysis has been done to optimise these methods [4].

Understanding exactly how neuron activations deviate, and what baseline they deviate from, is a broadly applicable question that is underexplored, and has the potential to improve techniques for model pruning and analysis into model sparsity.

In this paper, we 1) describe a simplistic working model of neuron activations, 2) suggest an improved, closed-form method of neuron ablation using modal activation, called 'peak ablation', and 3) run experimental analysis on various ablation methods to compare the degree to which they harm model performance.

The 2nd World Conference on eXplainable Artificial Intelligence
work@nicky.pro (N. Pochinkov)

2. Related Work

Mechanistic interpretability is a field of research focusing on understanding how neural network models achieve their outputs. [5, 3, 6, 7, 8, 9]. A common method used in mechanistic interpretability, is 'ablate and measure' [3]. We investigate more precisely how different ablation methods affect performance, and propose 'peak ablation' as another possible method.

Most relevantly, recent research [4] investigates hyperparameter selection to optimise activation patching for causal scrubbing. Our research differs; instead of interpolating activations between similar inputs, we set neurons' values for all inputs, and do not limit only to resampling.

Pruning: Model pruning [10] is a common practice wherein reduced neural network parameter counts lessen memory and performance costs. In particular, structured pruning of large features [11] is interested in the removal on the scale of neurons and attention heads, and can often achieve a large reduction in parameter count [12]. Our work seeks to question the assumption of using masks that set neuron values to zero.

Modularity: Research into activation sparsity [13], modularity [14], mixture of experts [15, 14, 16], and unlearning by pruning [17, 18] all investigate how different subsets of activations are responsible for different tasks. These implicitly set activations to zero.

3. Method

3.1. Pre-Trained Models and Datasets

We work with two causal text models, Mistral 7B [19] and Meta's OPT 1.3B [20], a masked text model, RoBERTa Large [21], and a vision transformer, ViT Base Patch16 224 [22].

To get a general sense of performance, the above models were evaluated by looking at top1 prediction accuracy[1], as well as cross-entropy loss on various datasets. For text models, we assess on EleutherAI's 'The Pile' [23]. For image models, we assess on Imagenet-1k [24], an image dataset with 1000 different classes. We evaluate on deterministic subsets of 100,000 text tokens and 1000 images respectively

3.2. Neurons

The objects of study are attention pre-out neurons, sometimes called 'z'-hook activations. We define attention pre-out neuron activations $y_i = f(x_i) = \text{preout}(x_i)$ as $y_i = \sum_j A_{i,j} W_V x_j$, where $A_{i,j} = \text{softmax}((W_K x_i) \cdot (W_Q x_j))$, where W_Q, W_K, W_V are the attention query, key, and value matrices respectively. We focus on attention neurons rather than MLP neurons, as these do not have an activation function that privileges positive activations, making analysis more difficult. To ablate a neuron, we replace $y_i = f(x_i)$ with some constant.

In Figure 1, we showcase some plots of neuron probability distributions. We see an example of many attention pre-out neuron activation distributions within the same layer. We note that most neurons follow a roughly Gaussian or double-exponential distribution about zero, but note that there is a minority of neurons that are not distributed at zero. As most neurons are zero-centred and symmetric, it makes sense that zero and mean ablation work quite well.

[1] top1 token prediction accuracy for language models, top1 image classification accuracy for image models

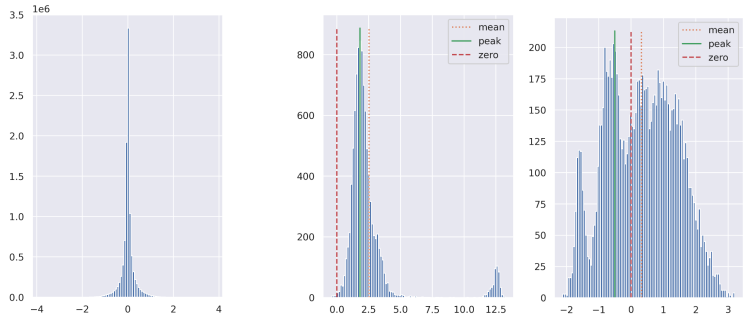


Figure 1: Un-normalised probability density functions (histograms) of attention neuron activations in RoBERTa. We see in (left) an average of distributions of all neurons in a layer, (centre) a bi-modal neuron with both peaks not at zero, and (right) another example of a neuron with an atypical distribution. X-axis shows neuron value, and Y-axis shows probability of a neuron taking that value.

3.3. A Working Model of Neuron Activation

Our hypothesis, based on activation profiles such as those seen in Figure 1 is that neurons have a 'baseline' or 'default-mode' activation (typically at zero), when the input contains no relevant features, which is then deviated from as neurons fire in proportion to various features they are tuned to pick up. In residual stream models [25], information is limited to the width of the residual stream [26], and as the residual stream typically grows exponentially in size [27], noise can become amplified. This is supported by the common redundancy of many circuits [28], even in transformer models trained without the use of dropout [29].

In particular, we expect that ablating neurons should have two contributors to reduced performance. These are 1) removing the relevant contextual computed information that the neuron is providing, and 2) taking the model activation out of distribution, by adding 'noise'. Ablating neurons to a constant value should cause some constant increase in loss for term 1, and different constant should contribute to different values of term 2. As we increase the distance from the 'default-mode' value, the neuron would further degrade the performance by taking the residual stream further out of distribution, thus in some sense, 'adding noise'.

3.4. Ablation Methods

We choose four main methods of ablating neurons, see Table 1 for a summary. These are:

Table 1
Comparison between the neural ablation methods described.

Method	Set the neuron activation...
Zero Ablation	...to zero
Mean Ablation	...to the mean value within the dataset D
Activation Resampling	...to some values from some different input
Naive Peak Ablation	...to the modal 'peak' activation value within the dataset D

Zero ablation: The most common form of ablation, which involves replacing a neuronal activation of any with a zeroed out activation. That is, setting $\forall x_j : f_i(x_j) = 0.0$

Mean Ablation: A still relatively-common method of ablation, which involves first collecting activations of various neurons on a distribution of inputs, and averaging the activations to find a mean activation. That is, for some dataset D, for $x_j \in D$, let $f_i(x_j) = \frac{1}{|D|} \sum_j f_i(x_j)$

Activation Resampling: Inspired by [3, 4], we also try general neuron resampling, by setting activations to those found by giving a randomised input. [2] For text model, we take activations by a) sampling random generated characters, b) sampling random tokens, and c) using OPT to generate a random text. For ViT, we use randomly generated pixel values.

Naive Peak Ablation: Observing that neuronal activations frequently exhibit a prominent peak, we propose an ablation method targeting their modal activation. For bin size ϵ, the neuron i activations $f_i(x_j)$ for each $x_j \in D$ are sorted into bins $N_i[k]$ such that $y_k \leq f_i(x_j) < y_k + \epsilon$. The bin $N_i[k_{max}]$ with the highest occurrence is selected, and $f_i(x_j)$ is set to $y_{k_{max}} + \frac{\epsilon}{2}$.

3.5. Ablation Experiments

Under the working model described in Section 3.3, we expect that ablating neurons to different values should have different impacts to performance, with there being a value which leads to some minimal drop in performance due to minimal noise being added to the residual stream.

We randomly select attention neurons in increments of 10% and ablate them until the model is fully pruned, and at each step, assess performance by evaluating the Top1 accuracy and Cross-Entropy Loss in the chosen dataset with each ablation method, described in Table 1. The neurons are selected deterministically across three separate seeds, summarised in Table 2

4. Results

4.1. Causal Text Models

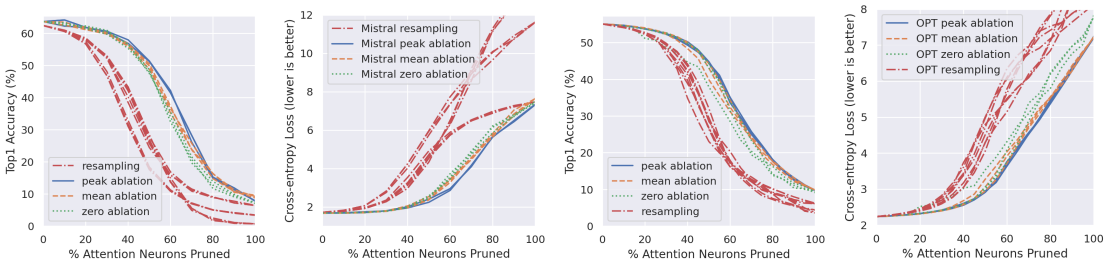
In Figure 2, we see the results for random pruning of OPT 1.3b and Mistral 7b with the different methods of ablation. We can see that Peak ablation has the most consistent pattern, causing the lowest amount of degradation, with mean ablation and zero ablation coming a close second and third, and Random resampling causes by far the most degradation. Of the three resampling methods, choosing random tokens causes the lowest degradation.

4.2. Other Transformers

In Figure 3, we see that for ViT, zero, mean, and peak ablation have statistically insignificant differences in performance, while resampling causes some small additional degradation. We can see that almost all of the performance loss is based on the specific neurons being selected rather than the ablation method being chosen, even between zero, mean, and peak ablation.

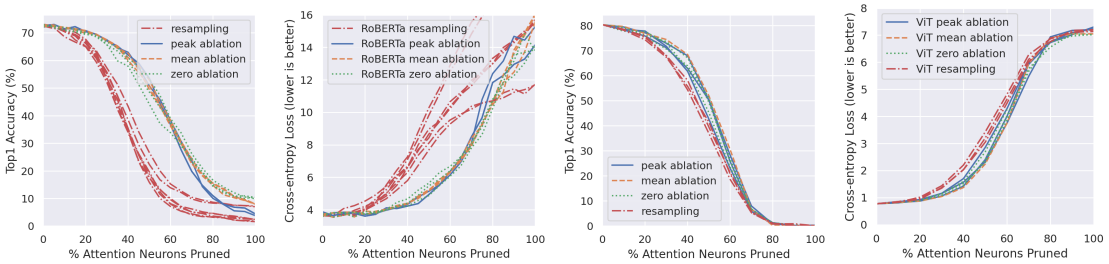
In RoBERTa, we see that in the first 75% of pruning, the three main methods of peak, mean and zero ablation are very close, with Peak edging slightly better performance. Beyond 75%, the three methods become more noisy; resampling of IDs ends up having the best performance overall in both Top1 and Cross-Entropy Loss at the task of token unmasking and de-randomization.

[2] This differs slightly to the original description, as in other research, they use a specific task, like circuit analysis [3] for the activation resampling, where the specific prompt template already exists.



(a) Mistral 7B Top1 (b) Mistral 7B CE Loss (c) OPT 1.3B Top1 (d) Opt 1.3B CE Loss

Figure 2: Change in Top1 next-token prediction accuracy (Top1) and cross-entropy loss (CE Loss) at different fractions of model pruned with different methods of ablation for Mistral 7B and OPT 1.3B



(a) RoBERTa Top1 (b) RoBERTa CE Loss (c) ViT Top1 (d) ViT CE Loss

Figure 3: Change in Top1 next-token prediction accuracy (Top1) and cross-entropy loss (CE Loss) at different fractions of model pruned with different methods of ablation for ViT 7B and RoBERTa

4.3. Overall Comparison

In Table 2, we see that in different models and in different regimes, the different methods have different merits in reducing performance, with Peak ablation working overall best in the most cases. Surprisingly, although Random resampling seems to add a lot of noise to the activations, random token ID resampling can sometimes work well, such as in RoBERTa.

5. Discussion

The analysis presented seems to suggest that when evaluating and understanding neurons in the attention layers of language models, the ideal centring method seems to depend significantly on the model. In decoder models, a good method is to find the largest peak, with a close second being zero ablation. This similarity is expected, as most neurons are centred at zero. This has downstream effects on improving the way we can look at one of the most crucial aspects of how neural networks work - their activations.

We have seen that neurons can have activations that are: non-Gaussian, non-symmetric, multi-modal, non-zero-centred. We hypothesise that taking into consideration this fact has the potential to make interpretability analysis into more fruitful, and centring activations by their peak seems a potential natural method.

Future work could: 1) investigate other potential better methods for neuron recentring, 2)

Table 2
Performance impact of neural ablation methods on the attention neurons of OPT, Mistral, ViT and RoBERTa. Ablation methods are Peak, Mean and Zero ablation, as well as Resampling (RS) with random characters (RS1), token IDs (RS2) and generated text (RS3) for text models, and random pixels (RS1) for ViT. Models are pruned by randomly selecting 50% and 90% of neurons

Top1 Accuracy		OPT	Mistral	ViT	RoBERTa
Baseline		55.05 ± 0.00	60.05 ± 0.00	80.32 ± 0.00	73.04 ± 0.00
50 % Pruning	Peak	$\mathbf{44.54 \pm 0.19}$	$\mathbf{51.30 \pm 0.05}$	$\mathbf{47.72 \pm 3.58}$	$\mathbf{52.55 \pm 1.27}$
	Mean	42.39 ± 1.84	49.71 ± 0.70	$\mathbf{48.65 \pm 5.38}$	51.10 ± 1.41
	Zero	41.77 ± 2.76	48.03 ± 0.50	$\mathbf{48.65 \pm 2.54}$	48.69 ± 3.75
	RS1	25.97 ± 2.24	27.01 ± 1.12	39.29 ± 1.84	18.57 ± 0.93
	RS2	27.51 ± 1.02	26.10 ± 1.34	-	24.83 ± 1.84
	RS3	29.90 ± 1.65	17.93 ± 0.33	-	16.10 ± 1.12
90% Pruning	Peak	$\mathbf{12.81 \pm 0.29}$	$\mathbf{11.70 \pm 0.26}$	0.20 ± 0.00	6.37 ± 0.62
	Mean	12.59 ± 0.39	10.84 ± 0.32	0.17 ± 0.17	10.46 ± 0.22
	Zero	11.05 ± 0.53	9.53 ± 0.34	$\mathbf{0.50 \pm 0.37}$	$\mathbf{11.20 \pm 0.60}$
	RS1	6.18 ± 0.43	1.03 ± 0.12	$\mathbf{0.37 \pm 0.39}$	3.19 ± 0.33
	RS2	7.35 ± 0.46	7.41 ± 0.25	-	7.55 ± 0.06
	RS3	5.55 ± 0.17	4.11 ± 0.09	-	2.26 ± 0.18
Cross-Entopy Loss		OPT	Mistral	ViT	RoBERTa
Baseline		2.24 ± 0.00	1.89 ± 0.00	0.77 ± 0.00	3.75 ± 0.00
50% Pruning	Peak	$\mathbf{2.93 \pm 0.01}$	$\mathbf{2.35 \pm 0.08}$	$\mathbf{2.52 \pm 0.23}$	$\mathbf{5.00 \pm 0.09}$
	Mean	3.09 ± 0.14	2.43 ± 0.07	$\mathbf{2.47 \pm 0.33}$	5.19 ± 0.03
	Zero	3.19 ± 0.24	2.49 ± 0.06	$\mathbf{2.46 \pm 0.17}$	5.33 ± 0.33
	RS1	4.53 ± 0.20	4.52 ± 0.16	3.22 ± 0.16	8.68 ± 0.22
	RS2	4.89 ± 0.13	4.71 ± 0.15	-	7.85 ± 0.18
	RS3	4.26 ± 0.16	5.88 ± 0.10	-	10.09 ± 0.20
90% Pruning	Peak	$\mathbf{6.33 \pm 0.04}$	$\mathbf{6.45 \pm 0.03}$	7.10 ± 0.06	13.53 ± 0.81
	Mean	$\mathbf{6.35 \pm 0.06}$	6.87 ± 0.12	7.07 ± 0.04	13.31 ± 0.60
	Zero	6.90 ± 0.10	6.75 ± 0.04	$\mathbf{6.99 \pm 0.04}$	12.99 ± 0.54
	RS1	8.40 ± 0.15	12.71 ± 0.18	7.13 ± 0.03	14.37 ± 0.06
	RS2	7.80 ± 0.11	7.25 ± 0.03	-	$\mathbf{11.32 \pm 0.09}$
	RS3	8.67 ± 0.09	10.80 ± 0.08	-	20.85 ± 0.19

more thoroughly investigate the differences between 'well-behaved' symmetric zero-centred distributions, and those that deviate from this norm, 3) find more efficient ways of computing the peak activations for larger models.

References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing 30 (2017).

[2] R. Meyes, M. Lu, C. W. de Puiseau, T. Meisen, Ablation studies in artificial neural networks, arXiv preprint arXiv:1901.08644 (2019).
[3] L. Chan, A. Garriga-Alonso, N. Goldowsky-Dill, R. Greenblatt, J. Nitishinskaya, A. Radhakrishnan, B. Shlegeris, N. Thomas, Causal scrubbing: a method for rigorously testing interpretability hypotheses, AI Alignment Forum, 2022. URL: https://www.alignmentforum.org/posts/JvZhhzycHu2Yd57RN.
[4] F. Zhang, N. Nanda, Towards best practices of activation patching in language models: Metrics and methods, arXiv preprint arXiv:2309.16042 (2023).
[5] M. Geva, R. Schuster, J. Berant, O. Levy, Transformer feed-forward layers are key-value memories, in: M. Moens, X. Huang, L. Specia, S. W. Yih (Eds.), Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021, Association for Computational Linguistics, 2021, pp. 5484–5495.
[6] A. Conmy, A. N. Mavor-Parker, A. Lynch, S. Heimersheim, A. Garriga-Alonso, Towards automated circuit discovery for mechanistic interpretability, CoRR abs/2304.14997 (2023). arXiv:2304.14997.
[7] nostalgebraist, interpreting gpt: the logit lens, LessWrong (2020). URL: https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens.
[8] N. Belrose, Z. Furman, L. Smith, D. Halawi, I. Ostrovsky, L. McKinney, S. Biderman, J. Steinhardt, Eliciting latent predictions from transformers with the tuned lens, arXiv preprint arXiv:2303.08112 (2023).
[9] C. Olsson, N. Elhage, N. Nanda, N. Joseph, N. DasSarma, T. Henighan, B. Mann, A. Askell, Y. Bai, A. Chen, et al., In-context learning and induction heads, arXiv preprint arXiv:2209.11895 (2022).
[10] D. W. Blalock, J. J. G. Ortiz, J. Frankle, J. V. Guttag, What is the state of neural network pruning?, in: I. S. Dhillon, D. S. Papailiopoulos, V. Sze (Eds.), Proceedings of Machine Learning and Systems 2020, MLSys 2020, Austin, TX, USA, March, 2020, mlsys.org, 2020.
[11] Z. Wang, J. Wohlwend, T. Lei, Structured pruning of large language models, in: B. Webber, T. Cohn, Y. He, Y. Liu (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020, Association for Computational Linguistics, 2020, pp. 6151–6162.
[12] E. Frantar, D. Alistarh, Sparsegpt: Massive language models can be accurately pruned in one-shot, in: International Conference on Machine Learning, PMLR, 2023, pp. 10323–10337.
[13] Z. Liu, J. Wang, T. Dao, T. Zhou, B. Yuan, Z. Song, A. Shrivastava, C. Zhang, Y. Tian, C. Ré, B. Chen, Deja vu: Contextual sparsity for efficient llms at inference time, in: A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, J. Scarlett (Eds.), International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA, volume 202 of *Proceedings of Machine Learning Research*, PMLR, 2023, pp. 22137–22176.
[14] Z. Zhang, Z. Zeng, Y. Lin, C. Xiao, X. Han, Z. Liu, M. Sun, J. Zhou, Emergent modularity in pre-trained transformers, 2023. URL: https://openreview.net/forum?id=XHuQacT6sa6.
[15] Z. Zhang, Y. Lin, Z. Liu, P. Li, M. Sun, J. Zhou, Moefication: Transformer feed-forward layers are mixtures of experts, in: S. Muresan, P. Nakov, A. Villavicencio (Eds.), Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022, Association for Computational Linguistics, 2022, pp. 877–890. URL: https://doi.org/

10.18653/v1/2022.findings-acl.71. doi:10.18653/v1/2022.findings-acl.71.

[16] J. Pfeiffer, S. Ruder, I. Vulic, E. M. Ponti, Modular deep learning, CoRR abs/2302.11529 (2023). URL: https://doi.org/10.48550/arXiv.2302.11529. arXiv:2302.11529.

[17] N. Pochinkov, N. Schoots, Dissecting large language models, in: Socially Responsible Language Modelling Research, 2023.

[18] J. Foster, S. Schoepf, A. Brintrup, Fast machine unlearning without retraining through selective synaptic dampening, CoRR abs/2308.07707 (2023). arXiv:2308.07707.

[19] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, W. E. Sayed, Mistral 7b (2023).

[20] S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. T. Diab, X. Li, X. V. Lin, T. Mihaylov, M. Ott, S. Shleifer, K. Shuster, D. Simig, P. S. Koura, A. Sridhar, T. Wang, L. Zettlemoyer, OPT: open pre-trained transformer language models, CoRR abs/2205.01068 (2022). URL: https://doi.org/10.48550/arXiv.2205.01068. arXiv:2205.01068.

[21] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized BERT pretraining approach, CoRR abs/1907.11692 (2019). URL: http://arxiv.org/abs/1907.11692. arXiv:1907.11692.

[22] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, in: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021, OpenReview.net, 2021. URL: https://openreview.net/forum?id=YicbFdNTTy.

[23] L. Gao, S. Biderman, S. Black, L. Golding, T. Hoppe, C. Foster, J. Phang, H. He, A. Thite, N. Nabeshima, S. Presser, C. Leahy, The Pile: An 800gb dataset of diverse text for language modeling, arXiv preprint arXiv:2101.00027 (2020).

[24] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, L. Fei-Fei, ImageNet Large Scale Visual Recognition Challenge, International Journal of Computer Vision (IJCV) 115 (2015) 211–252. doi:10.1007/s11263-015-0816-y.

[25] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[26] M. Thorpe, Y. van Gennip, Deep limits of residual neural networks, arXiv preprint arXiv:1810.11741 (2018).

[27] S. Heimersheim, A. Turner, Residual stream norms grow exponentially over the forward pass, https://www.alignmentforum.org/posts/8mizBCm3dyc432nK8/residual-stream-norms-grow-exponentially-over-the-forward, 2023. 14 min read.

[28] K. Wang, A. Variengien, A. Conmy, B. Shlegeris, J. Steinhardt, Interpretability in the wild: a circuit for indirect object identification in gpt-2 small, arXiv preprint arXiv:2211.00593 (2022).

[29] T. McGrath, M. Rahtz, J. Kramár, V. Mikulik, S. Legg, The hydra effect: Emergent self-repair in language model computations, CoRR abs/2307.15771 (2023). URL: https://doi.org/10.48550/arXiv.2307.15771. doi:10.48550/arXiv.2307.15771. arXiv:2307.15771.