You can remove GPT2's LayerNorm by fine-tuning

Stefan Heimersheim
Apollo Research
`stefan@apolloresearch.ai`

Abstract

The LayerNorm (LN) layer in GPT-style transformer models has long been a hindrance to mechanistic interpretability. LN is a crucial component required to stabilize the training of large language models, and LN or the similar RMSNorm have been used in practically all large language models based on the transformer architecture. The non-linear nature of the LN layers is a hindrance for mechanistic interpretability as it hinders interpretation of the residual stream, and makes it difficult to decompose the model into circuits. Some researchers have gone so far as to name "reasons interpretability researchers hate layer norm".

In this paper we show that it is possible to remove the LN layers from a pre-trained GPT2-small model by fine-tuning on a fraction (500M tokens) of the training data. We demonstrate that this LN-free model achieves similar performance to the original model on the OpenWebText and ThePile datasets (-0.05 cross-entropy loss), and the Hellaswag benchmark (-0.5% accuracy). We provide our implementation at https://github.com/ApolloResearch/gpt2_noLN, and fine-tuned GPT2-small models at https://huggingface.co/apollo-research/gpt2_noLN.

Our work not only provides a simplified model for mechanistic interpretability research, but also provides evidence that the LN layers, at inference time, do not play a crucial role in transformer models.

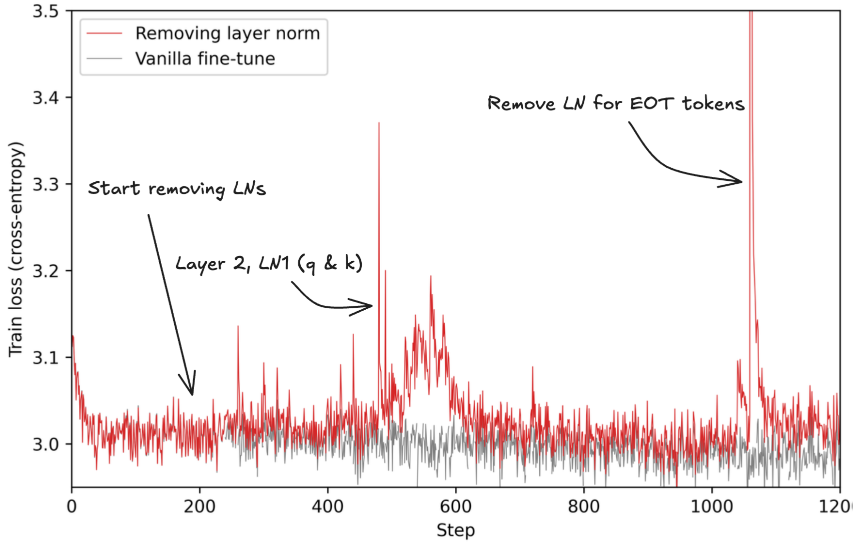

Figure 1: Removing LayerNorm while fine-tuning. The loss curve of a GPT2-small model being fine-tuned while gradually removing LN layers (red), compared to the loss from fine-tuning a vanilla GPT2-small model (gray).

Interpretable AI: Past, Present and Future Workshop at NeurIPS 2024

1 Introduction

Mechanistic interpretability aims to understand the inner workings of neural networks by analyzing individual network components and their interactions (circuits). Recent work based on sparse dictionary learning made progress in understanding the residual stream [Sharkey et al., 2022, Cunningham et al., 2023, Bricken et al., 2023, Templeton et al., 2024], the attention layers [Kissane et al., 2024a, Wynroe and Sharkey, 2024] and the feed-forward layers Dunefsky et al. [2024]. Attribution techniques such as direct logic attribution [nostalgebraist, 2020, Elhage et al., 2021, Wang et al., 2022], integrated gradients [Friedman, 2004, Sundararajan et al., 2017, Bushnaq et al., 2024], and activation- and attribution patching [e.g. Vig et al., 2020, Geiger et al., 2021, Meng et al., 2023, Nanda, 2023] have been used to understand which model internals are responsible for the model's behavior.

All frontier LLMs are transformer models [Brown et al., 2020, Touvron et al., 2023, OpenAI et al., 2024, Gemini Team et al., 2024]. Transformers consist of a residual stream, and a series of components (attention layers, feed-forward layers) that read and write from the residual stream. Of particular interest for this paper are the normalization layers that normalize the residual stream as it is read by the attention and feed-forward layers, and a final normalization layer that normalizes the residual stream before the unembedding. These normalization layers are introduced to stabilize and speed up training of models [as a replacement for batch normalization, Ioffe and Szegedy, 2015] and are active at inference time (unlike batch normalization layers). The two common choices are LayerNorm [LN, Lei Ba et al., 2016] or RMSNorm [Zhang and Sennrich, 2019]. Both operate on the embedding dimension of the residual stream.

$$\text{LN}(\mathbf{x}) = \frac{\mathbf{x} - \mu}{\sigma} \odot \boldsymbol{\gamma} + \boldsymbol{\beta} \qquad \text{RMSNorm}(\mathbf{x}) = \frac{\mathbf{x}}{\sigma} \odot \boldsymbol{\gamma} \tag{1}$$

$$\text{where} \quad \mu = \frac{1}{H}\sum_{h=1}^{H} x_h \qquad \sigma = \sqrt{\frac{1}{H}\sum_{h=1}^{H}(x_h - \mu)^2} \tag{2}$$

At inference time, the mean centering (μ), weight ($\boldsymbol{\gamma}$), and bias ($\boldsymbol{\beta}$) parameters can be folded into neighboring layers [1], so both normalizations are equivalent and can be simplified as a division by the standard deviation[2] (σ) of the embedding vector. For simplicity, we will refer to both normalization layers as LN in the following.

These LN layers have been a hindrance to mechanistic interpretability over the last years. The reasons mostly[3] fall into three categories:

1. Residual stream directions can not be directly interpreted as changes to the logits due to the final LN layer. This hinders logit lens analysis [also known as direct logit attribution, nostalgebraist, 2020, Elhage et al., 2021, Wang et al., 2022], as well as attribution patching [Nanda, 2023]. Olah et al. [2023] refer to this as "reason #78 for why interpretability researchers hate LayerNorm".
2. The transformer cannot be decomposed well into individual paths (circuits) without approximating LN layers. Elhage et al. [2021] and Sharkey [2023] have argued that decomposing transformer models into individual circuits would be much easier without LN. In practice, Bricken et al. [2023], McDougall et al. [2023], and Kissane et al. [2024a] all approximate (linearize) LN layers by freezing the normalization scale.
3. We do not know whether the LN layers play an important role in the model's computation. Recent work on toy models [Winsor, 2022] showed that LN can be used as the sole non-linearity, [Stolfo et al., 2024] suggest that LN might be used to implement confidence regularization in LLMs.

In brief, it is a common occurrence to hear the phrase "turns out that LayerNorm completely breaks things" [Nanda, 2023] among mechanistic interpretability researchers.

---

[1] See e.g. `fold_ln` in TransformerLens [Nanda and Bloom, 2022].

[2] Note that the "standard deviation" here is simply applied to an individual embedding vector, separately for each batch or token index.

[3] The extra LN layers introduced in Elhage et al. [2022] are introduced for a different reason, and have and unrelated though also-hindering effect on interpretability.

In this paper, we demonstrate that, in GPT2-small, the LN layers can be removed after pre-training by fine-tuning on a small fraction of the training data. Our primary goal is to show that an LLM of near-identical capability to GPT2-small can be achieved without any LN layers. We propose that such a model should be used as model organism for interpretability research, the role that is currently played by the original GPT2-small model. Previously, the only available language transformer models without LN were tiny models, such as the 4-layer TinyModel [Nabeshima, 2024].

We provide details of our fine-tuning procedure in Section 2, present loss-curves and final model benchmarks in Section 3, and discuss applications and open questions in Section 4. We provide the fine-tuned GPT2-small model in this Hugging Face repository, including code to load the model into the TransformerLens [Nanda and Bloom, 2022] library.

2 Methodology

Previous works [e.g. Heimersheim and Turner, 2023] observed the residual stream standard deviation σ does not vary a lot between different forward passes (except for end-of-text (EOT) tokens used to indicate the beginning or end of sequences, and the first token in a prompt). This suggests that replacing the per-token standard deviation σ with a constant value $\bar{\sigma}$ calculated by averaging over a couple of prompts ["freezing" the normalization scale, as done in Bricken et al., 2023, McDougall et al., 2023, Kissane et al., 2024a] may be possible. We find that freezing all LN layers simultaneously breaks the model irreparably, resulting in either cross-entropy loss reaching NaN or remaining permanently $\gg 20$. However, a more gradual approach—freezing (parts of) the LN layers

Table 1: Training step at which we disable each LayerNorm layer. The digit before the dot indicates the transformer block, the name refers to the LN before the attention layer (`ln1qk` or `ln1v`), the feed-forward layer (`ln2`), or the unembedding (`lnf`). `eot` indicates the special case for the EOT tokens, and `bos` the special case for the first token.

Layer	**v1**	**v2**	**v3**	**v4**	**v5**	**Layer**	**v1**	**v2**	**v3**	**v4**	**v5**
`0.ln2`	50	50	200	200	180	`7.ln1v`	50	350	1230	890	610
`1.ln2`	50	50	240	220	190	`8.ln1v`	50	350	1240	920	620
`2.ln2`	50	50	280	240	200	`9.ln1v`	50	350	1250	950	630
`3.ln2`	50	50	320	260	210	`10.ln1v`	50	350	1260	980	640
`4.ln2`	50	50	360	280	220	`11.ln1v`	50	350	1270	1010	650
`5.ln2`	50	50	400	300	230	`lnf`	300	400	1640	1040	660
`6.ln2`	50	50	440	320	240	`0.eot`	200	500	1740	1060	680
`7.ln2`	50	50	480	340	250	`1.eot`	200	500	1740	1060	700
`8.ln2`	50	50	520	360	260	`2.eot`	200	500	1740	1060	720
`9.ln2`	50	50	560	380	270	`3.eot`	200	500	1740	1060	740
`10.ln2`	50	50	600	400	280	`4.eot`	200	500	1740	1060	760
`11.ln2`	50	50	640	420	290	`5.eot`	200	500	1740	1060	780
`0.ln1qk`	50	100	680	440	300	`6.eot`	200	500	1740	1060	800
`1.ln1qk`	50	120	720	460	320	`7.eot`	200	500	1740	1060	820
`2.ln1qk`	50	140	760	480	340	`8.eot`	200	500	1740	1060	840
`3.ln1qk`	50	160	800	500	360	`9.eot`	200	500	1740	1060	860
`4.ln1qk`	50	180	840	520	380	`10.eot`	200	500	1740	1060	880
`5.ln1qk`	50	200	880	540	400	`11.eot`	200	500	1740	1060	900
`6.ln1qk`	50	220	920	560	420	`0.bos`	200	700	2040	1160	920
`7.ln1qk`	50	240	960	580	440	`1.bos`	200	700	2040	1160	925
`8.ln1qk`	50	260	1000	600	460	`2.bos`	200	700	2040	1160	930
`9.ln1qk`	50	280	1040	620	480	`3.bos`	200	700	2040	1160	935
`10.ln1qk`	50	300	1080	640	500	`4.bos`	200	700	2040	1160	940
`11.ln1qk`	50	320	1120	660	520	`5.bos`	200	700	2040	1160	945
`0.ln1v`	50	350	1160	680	540	`6.bos`	200	700	2040	1160	950
`1.ln1v`	50	350	1170	710	550	`7.bos`	200	700	2040	1160	955
`2.ln1v`	50	350	1180	740	560	`8.bos`	200	700	2040	1160	960
`3.ln1v`	50	350	1190	770	570	`9.bos`	200	700	2040	1160	965
`4.ln1v`	50	350	1200	800	580	`10.bos`	200	700	2040	1160	970
`5.ln1v`	50	350	1210	830	590	`11.bos`	200	700	2040	1160	975
`6.ln1v`	50	350	1220	860	600	lr-sched.	const	const	const	var	var

incrementally—yields a recoverable state. In this case, the loss initially increases, sometimes spiking to ~ 20, but the model can be fine-tuned to restore the loss to approximately its original value.

Our fine-tuning procedure contains three key ingredients:

1. Disable one LN at a time. There are two LN layers in each transformer block, `ln1` before attention layer, and `ln2` before the feed-forward layer, plus the final layer norm `lnf`. We disable one LN layer in one block at a time, and fine-tune the model for a small number of steps.
2. Treat `ln1` before the query and key vectors ("`ln1qk`") separately from the `ln1` before the value vectors ("`ln1v`"). We noticed that the latter appeared to be more sensitive to freezing the LN scale, and disabling `ln1v` after `ln1qk` led to a more stable fine-tuning procedure.
3. Handle the first sequence position, and EOT tokens, separately. In these situation the standard deviation tends to be much larger [Heimersheim and Turner, 2023], so we use a second fixed $\bar{\sigma}_0$ value for these cases. We use the special case for the sequence position for all LN layers, but the special case for EOT tokens was only necessary for `ln1v`. Towards the end of the fine-tuning procedure, we remove these special cases one by one.

We collect the average standard deviations from 16 OpenWebText prompts, using the first token to calculate $\bar{\sigma}_0$, and the remaining tokens to calculate $\bar{\sigma}$. We then fine-tune GPT2-small on the OpenWebText dataset [Gokaslan and Cohen, 2019]. We use a batch size of 48, with 10 gradient accumulation steps, and a sequence length of 1024. We refer to 10 batches (i.e. one full gradient accumulation) as one step, containing 491,520 tokens. We use a base learning rate of $6 \cdot 10^{-4}$, and optionally use a linear learning rate warm-up for 100 steps and cosine decay schedule to decrease the learning rate to $6 \cdot 10^{-5}$ after 2000 steps. Most models are trained for around 1000 steps, which corresponds to around 500M tokens, and takes around 2 hours on a single A100 GPU.

We start with ~ 200 steps with all LN layers enabled, then disable LN layers one by one, and then disable the special case for the first token and EOT tokens (optionally for one layer at a time). We run a sweep of experiments using different LN removal schedules (sometimes removing LNs layers in multiple blocks at a time) and report the best-performing schedules in Table 1.

We evaluate the final LN-free models on the OpenWebText dataset, the ThePile [Gao et al., 2020, via `apollo-research/monology-pile-uncopyrighted-tokenizer-gpt2`] dataset (cross-entropy loss), and the Hellaswag [Zellers et al., 2019] benchmark [using the implementation by Karpathy, 2022]. To provide a fair comparison on the OpenWebText dataset—as our model was fine-tuned on (a different section of) this dataset—we also fine-tune a "vanilla" GPT2-small model (with all LN layers enabled) for 1000, 1200, and 2000 steps, and report its performance.

Table 2: Comparison of model performance between the fine-tuned no-LN model, the original GPT2-small model, and equivalently fine-tuned vanilla models. Note: Runs v4 and v5 used the variable learning rate schedule, so for comparison these should be compared to the starred vanilla model (which uses the same schedule).

Version	OpenWebText	ThePile	Hellaswag
no-LN v1 (850 steps)	3.130	3.057	29.17%
no-LN v2 (1000 steps)	3.014	2.926	29.54%
no-LN v3 (2450 steps)	3.010	2.931	29.39%
no-LN v4 (1200 steps*)	**3.000**	**2.900**	29.51%
no-LN v5 (1100 steps*)	3.006	2.930	29.52%
Original (no fine-tuning)	3.095	2.856	29.56%
Vanilla (1000 steps)	2.989	2.880	29.82%
Vanilla (2000 steps)	2.978	2.905	29.64%
Vanilla (1200 steps*)	**2.966**	**2.850**	30.01%

3 Results

We are able to successfully train a GPT2-small model without any LN layers ("no-LN" model). Our best model (v4) achieves a cross-entropy loss of 3.000 on the OpenWebText dataset (compared to 2.966 for the fine-tuned vanilla GPT2-small model with LN). On ThePile our same model achieves a loss of 2.9000 (compared to 2.850 for the baseline with LN). Additionally we report the accuracy on the Hellaswag benchmark, which is 29.5% (compared to 30.0% for the original GPT2-small model). Table 2 shows the performance of the 5 different fine-tuned models (v1-v5) as described in Table 1. We also provide generations from a no-LN model, compared to the vanilla model, in Appendix A. There is no notable difference in the quality of the generations.

To put this loss difference into perspective, we estimate what model size would be needed to achieve this loss at the same compute budget. Following the Chinchilla-based scaling laws [Hoffmann et al., 2024] we roughly estimate [using the graphs from Korbak, 2022] that a 0.05 cross-entropy loss difference corresponds to a 93M (rather than 117M) parameter model.

Figure 2 shows the training loss curves for all 5 no-LN models, and Figure 1 shows the loss curve for the best no-LN model (v4) compared to equivalently fine-tuning a vanilla GPT2-small model. We observe jumps in the loss curve when disabling the LN layers, as expected, but in most cases (the ones shown here) the model is able to recover from these jumps. Empirically we found that disabling many LN layers at once, or in quick succession, leads to a significant loss spike from which the model does not fully recover. For example, in the no-LN v1 we disabled many LN layers at once, and the OpenWebText validation loss never drops below 3.1 after that.

The last run, no-LN v5, is a run where we spread out the LN removal as widely as possible (every removal happens at a different step). We find that we can indeed avoid large loss spikes, although the model does not end up being our best-performing model (possibly due to still removing LN layers in quick succession).

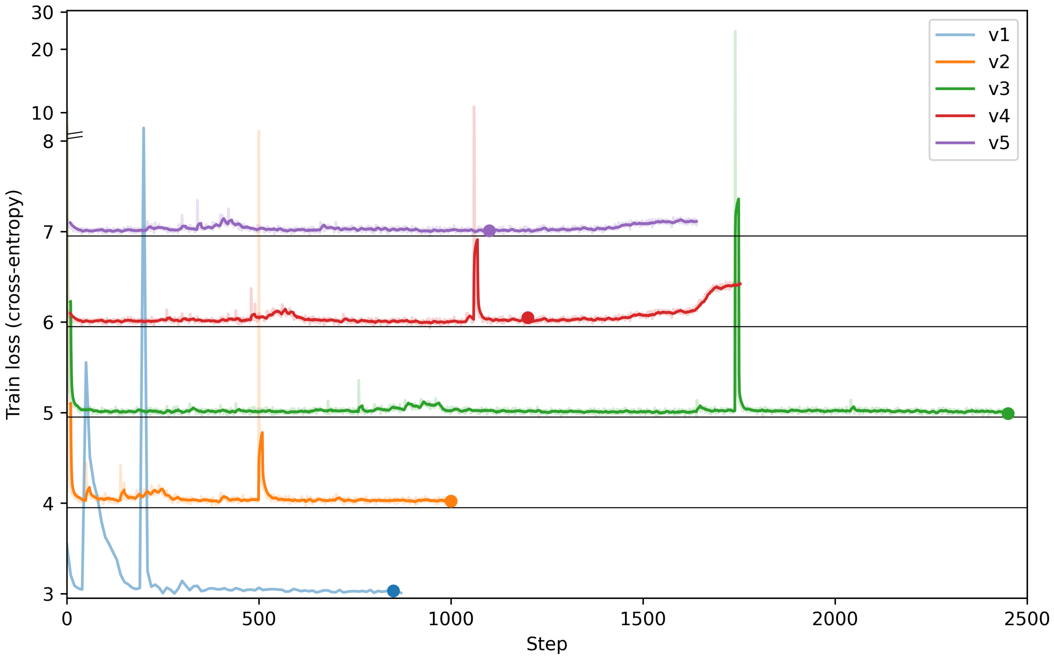

Figure 2: Training loss curves for all 5 no-LN models. The dot indicates the snapshot of best validation loss, which is the one we use for the benchmarks. Note that the y-axis is offset by 1 for each version, and log-scaled for $y > 8$.

4 Discussion

We want to discuss the application of this technique, and the loss penalty for removing LayerNorm. The cross-entropy loss drop of 0.05 is a significant drop in terms of modern LLMs, and we would not expect a production model to remove LN layers. However, our goal is not primarily to make SOTA model interpretable, but to work towards a full mechanistic understand of *any* large language model. We don't know whether mechanistic interpretability insights from the no-LN version of a model would transfer to the original model, i.e. whether the no-LN model is sufficiently faithful to the original model. However, the transfer of other techniques [e.g. Kissane et al., 2024b] suggests that they might. This would allow us to reach a secondary goal of (eventually) understanding SOTA LLMs.

The high performance of the no-LN models suggest that the LN layers do not play an important role in language modelling. This provides evidence that the common practice of linearizing LN [Bricken et al., 2023, McDougall et al., 2023, Kissane et al., 2024a] probably does not obscure important model behaviour.[4]

In this work we only consider the GPT2-small model, which leads to two limitations: (1) It is possible that LN layers play a more important role in larger models and it is not possible to remove LN there. (2) Training larger models is harder, so this fine-tuning procedure might be more difficult or more expensive for larger models.

We also want to highlight some confusing aspects of our results: (1) We found that the loss on ThePile drops more than on OpenWebText when removing LN, suggesting perhaps a worse generalization. However this also brings the losses on ThePile and OpenWebText closer together, so we are unsure about the conclusion on generalization. (2) We noticed that in the runs with variable learning rate schedule (v4 and v5), the the loss curves started to rise towards the end of training (long after removing the LNs). We do not understand why this happens when the learning rate is decreased.

There are a few improvements that we would like to see in future works. First, it might be helpful to collect more data to compute the averages $\bar{\sigma}$ and $\bar{\sigma}_0$, and possibly to separate $\bar{\sigma}_0$ into separate averages for position 0 tokens and EOT tokens. Second, we only briefly explored the idea of gradually turning off the individual LN layers (not shown in this paper).[5] Such a gradual removal might help reduce the loss spikes further. Finally, and most importantly, we would like to see this technique applied to much larger models.

5 Conclusion

This paper demonstrates that the LayerNorm layers in GPT2-small can be removed after pre-training by fine-tuning on a small fraction of the training data (500M tokens, 2 GPU-hours). We make our trained models available in a Hugging Face repository, and the models are already being used in mechanistic interpretability research [work in progress, Giglemiani, 2024, Janiak, 2024]

The removal of LayerNorm allows researchers to leverage the success of sparse dictionary learning in understanding individual components of the transformer model [Sharkey, 2024, Templeton et al., 2024, Wynroe and Sharkey, 2024, Dunefsky et al., 2024] and analyze the interaction between these components, putting it all together.

While frontier models will likely continue to be trained with LN (or a similar normalization layer), we believe that understanding any capable LLM (e.g. a GPT2-small model) would be a major success of mechanistic interpretability. Techniques or insights transferring to larger models with LN would be a bonus.

---

[4] To be clear, the model (without fine-tuning) absolutely does not work when LN is linearized, but the fact that it can adjust to this change (with fine-tuning) relatively easily is a good sign.

[5] This can be done by using a combination of σ and $\bar{\sigma}$ in the LN formula, and slowly increasing the weight of $\bar{\sigma}$ while fine-tuning.

Acknowledgments

Thanks to Alice Rigg, Bilal Chughtai, Leo Gao, Neel Nanda, and Rudolf Laine for comments and feedback on the draft and idea. The nanoGPT repository and accompanying video by Andrej Karpathy helped me implement this idea quickly.

References

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.

Lucius Bushnaq, Jake Mendel, Kaarel Hänni, and Stefan Heimersheim. Interpretability: Integrated gradients is a decent attribution method, May 2024. URL `https://www.lesswrong.com/posts/Rv6ba3CMhZGZzNH7x/interpretability-integrated-gradients-is-a-decent`.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*, 2023.

Jacob Dunefsky, Philippe Chlenski, and Neel Nanda. Transcoders enable fine-grained interpretable circuit analysis for language models, Apr 2024. URL `https://www.alignmentforum.org/posts/YmkjnWtZGLbHRbzrP/transcoders-enable-fine-grained-interpretable-circuit`.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. https://transformer-circuits.pub/2021/framework/index.html.

Nelson Elhage, Tristan Hume, Catherine Olsson, Neel Nanda, Tom Henighan, Scott Johnston, Sheer ElShowk, Nicholas Joseph, Nova DasSarma, Ben Mann, Danny Hernandez, Amanda Askell, Kamal Ndousse, Andy Jones, Dawn Drain, Anna Chen, Yuntao Bai, Deep Ganguli, Liane Lovitt, Zac Hatfield-Dodds, Jackson Kernion, Tom Conerly, Shauna Kravec, Stanislav Fort, Saurav Kadavath, Josh Jacobson, Eli Tran-Johnson, Jared Kaplan, Jack Clark, Tom Brown, Sam McCandlish, Dario Amodei, and Christopher Olah. Softmax linear units. *Transformer Circuits Thread*, 2022. https://transformer-circuits.pub/2022/solu/index.html.

Eric Friedman. Paths and consistency in additive cost sharing. *International Journal of Games Theory*, 32: 501–518, 08 2004. doi: 10.1007/s001820400173.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. The Pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. *Advances in Neural Information Processing Systems*, 34:9574–9586, 2021.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul R. Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, Jack Krawczyk, Cosmo Du, Ed Chi, Heng-Tze Cheng, Eric Ni, Purvi Shah, Patrick Kane, Betty Chan, Manaal Faruqui, Aliaksei Severyn, Hanzhao Lin, YaGuang Li, Yong Cheng, Abe Ittycheriah, Mahdis Mahdieh, Mia Chen, Pei Sun, Dustin Tran, Sumit Bagri, Balaji Lakshminarayanan, Jeremiah Liu, Andras Orban, Fabian Güra, Hao Zhou, Xinying Song, Aurelien Boffy, Harish Ganapathy, Steven Zheng, HyunJeong Choe, Ágoston Weisz, Tao Zhu, Yifeng Lu, Siddharth Gopal, Jarrod Kahn, Maciej

Kula, Jeff Pitman, Rushin Shah, Emanuel Taropa, Majd Al Merey, Martin Baeuml, Zhifeng Chen, Laurent El Shafey, Yujing Zhang, Olcan Sercinoglu, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, Alexandre Frechette, Charlotte Smith, Laura Culp, Lev Proleev, Yi Luan, Xi Chen, James Lottes, Nathan Schucher, Federico Lebron, Alban Rrustemi, Natalie Clay, Phil Crone, Tomas Kocisky, Jeffrey Zhao, Bartek Perz, Dian Yu, Heidi Howard, Adam Bloniarz, Jack W. Rae, Han Lu, Laurent Sifre, Marcello Maggioni, Fred Alcober, Dan Garrette, Megan Barnes, Shantanu Thakoor, Jacob Austin, Gabriel Barth-Maron, William Wong, Rishabh Joshi, Rahma Chaabouni, Deeni Fatiha, Arun Ahuja, Gaurav Singh Tomar, Evan Senter, Martin Chadwick, Ilya Kornakov, Nithya Attaluri, Iñaki Iturrate, Ruibo Liu, Yunxuan Li, Sarah Cogan, Jeremy Chen, Chao Jia, Chenjie Gu, Qiao Zhang, Jordan Grimstad, Ale Jakse Hartman, Xavier Garcia, Thanumalayan Sankaranarayana Pillai, Jacob Devlin, Michael Laskin, Diego de Las Casas, Dasha Valter, Connie Tao, Lorenzo Blanco, Adrià Puigdomènech Badia, David Reitter, Mianna Chen, Jenny Brennan, Clara Rivera, Sergey Brin, Shariq Iqbal, Gabriela Surita, Jane Labanowski, Abhi Rao, Stephanie Winkler, Emilio Parisotto, Yiming Gu, Kate Olszewska, Ravi Addanki, Antoine Miech, Annie Louis, Denis Teplyashin, Geoff Brown, Elliot Catt, Jan Balaguer, Jackie Xiang, Pidong Wang, Zoe Ashwood, Anton Briukhov, Albert Webson, Sanjay Ganapathy, Smit Sanghavi, Ajay Kannan, Ming-Wei Chang, Axel Stjerngren, Josip Djolonga, Yuting Sun, Ankur Bapna, Matthew Aitchison, Pedram Pejman, Henryk Michalewski, Tianhe Yu, Cindy Wang, Juliette Love, Junwhan Ahn, Dawn Bloxwich, Kehang Han, Peter Humphreys, Thibault Sellam, James Bradbury, Varun Godbole, Sina Samangooei, Bogdan Damoc, Alex Kaskasoli, Sébastien M. R. Arnold, Vijay Vasudevan, Shubham Agrawal, Jason Riesa, Dmitry Lepikhin, Richard Tanburn, Srivatsan Srinivasan, Hyeontaek Lim, Sarah Hodkinson, Pranav Shyam, Johan Ferret, Steven Hand, Ankush Garg, Tom Le Paine, Jian Li, Yujia Li, Minh Giang, Alexander Neitz, Zaheer Abbas, Sarah York, Machel Reid, Elizabeth Cole, Aakanksha Chowdhery, Dipanjan Das, Dominika Rogozińska, Vitaliy Nikolaev, Pablo Sprechmann, Zachary Nado, Lukas Zilka, Flavien Prost, Luheng He, Marianne Monteiro, Gaurav Mishra, Chris Welty, Josh Newlan, Dawei Jia, Miltiadis Allamanis, Clara Huiyi Hu, Raoul de Liedekerke, Justin Gilmer, Carl Saroufim, Shruti Rijhwani, Shaobo Hou, Disha Shrivastava, Anirudh Baddepudi, Alex Goldin, Adnan Ozturel, Albin Cassirer, Yunhan Xu, Daniel Sohn, Devendra Sachan, Reinald Kim Amplayo, Craig Swanson, Dessie Petrova, Shashi Narayan, Arthur Guez, Siddhartha Brahma, Jessica Landon, Miteyan Patel, Ruizhe Zhao, Kevin Villela, Luyu Wang, Wenhao Jia, Matthew Rahtz, Mai Giménez, Legg Yeung, James Keeling, Petko Georgiev, Diana Mincu, Boxi Wu, Salem Haykal, Rachel Saputro, Kiran Vodrahalli, James Qin, Zeynep Cankara, Abhanshu Sharma, Nick Fernando, Will Hawkins, Behnam Neyshabur, Solomon Kim, Adrian Hutter, Priyanka Agrawal, Alex Castro-Ros, George van den Driessche, Tao Wang, Fan Yang, Shuo yiin Chang, Paul Komarek, Ross McIlroy, Mario Lučić, Guodong Zhang, Wael Farhan, Michael Sharman, Paul Natsev, Paul Michel, Yamini Bansal, Siyuan Qiao, Kris Cao, Siamak Shakeri, Christina Butterfield, Justin Chung, Paul Kishan Rubenstein, Shivani Agrawal, Arthur Mensch, Kedar Soparkar, Karel Lenc, Timothy Chung, Aedan Pope, Loren Maggiore, Jackie Kay, Priya Jhakra, Shibo Wang, Joshua Maynez, Mary Phuong, Taylor Tobin, Andrea Tacchetti, Maja Trebacz, Kevin Robinson, Yash Katariya, Sebastian Riedel, Paige Bailey, Kefan Xiao, Nimesh Ghelani, Lora Aroyo, Ambrose Slone, Neil Houlsby, Xuehan Xiong, Zhen Yang, Elena Gribovskaya, Jonas Adler, Mateo Wirth, Lisa Lee, Music Li, Thais Kagohara, Jay Pavagadhi, Sophie Bridgers, Anna Bortsova, Sanjay Ghemawat, Zafarali Ahmed, Tianqi Liu, Richard Powell, Vijay Bolina, Mariko Iinuma, Polina Zablotskaia, James Besley, Da-Woon Chung, Timothy Dozat, Ramona Comanescu, Xiance Si, Jeremy Greer, Guolong Su, Martin Polacek, Raphaël Lopez Kaufman, Simon Tokumine, Hexiang Hu, Elena Buchatskaya, Yingjie Miao, Mohamed Elhawaty, Aditya Siddhant, Nenad Tomasev, Jinwei Xing, Christina Greer, Helen Miller, Shereen Ashraf, Aurko Roy, Zizhao Zhang, Ada Ma, Angelos Filos, Milos Besta, Rory Blevins, Ted Klimenko, Chih-Kuan Yeh, Soravit Changpinyo, Jiaqi Mu, Oscar Chang, Mantas Pajarskas, Carrie Muir, Vered Cohen, Charline Le Lan, Krishna Haridasan, Amit Marathe, Steven Hansen, Sholto Douglas, Rajkumar Samuel, Mingqiu Wang, Sophia Austin, Chang Lan, Jiepu Jiang, Justin Chiu, Jaime Alonso Lorenzo, Lars Lowe Sjösund, Sébastien Cevey, Zach Gleicher, Thi Avrahami, Anudhyan Boral, Hansa Srinivasan, Vittorio Selo, Rhys May, Konstantinos Aisopos, Léonard Hussenot, Livio Baldini Soares, Kate Baumli, Michael B. Chang, Adrià Recasens, Ben Caine, Alexander Pritzel, Filip Pavetic, Fabio Pardo, Anita Gergely, Justin Frye, Vinay Ramasesh, Dan Horgan, Kartikeya Badola, Nora Kassner, Subhrajit Roy, Ethan Dyer, Víctor Campos Campos, Alex Tomala, Yunhao Tang, Dalia El Badawy, Elspeth White, Basil Mustafa, Oran Lang, Abhishek Jindal, Sharad Vikram, Zhitao Gong, Sergi Caelles, Ross Hemsley, Gregory Thornton, Fangxiaoyu Feng, Wojciech Stokowiec, Ce Zheng, Phoebe Thacker, Çağlar Ünlü, Zhishuai Zhang, Mohammad Saleh, James Svensson, Max Bileschi, Piyush Patil, Ankesh Anand, Roman Ring, Katerina Tsihlas, Arpi Vezer, Marco Selvi, Toby Shevlane, Mikel Rodriguez, Tom Kwiatkowski, Samira Daruki, Keran Rong, Allan Dafoe, Nicholas FitzGerald, Keren Gu-Lemberg, Mina Khan, Lisa Anne Hendricks, Marie Pellat, Vladimir Feinberg, James Cobon-Kerr, Tara Sainath, Maribeth Rauh, Sayed Hadi Hashemi, Richard Ives, Yana Hasson, Eric Noland, Yuan Cao, Nathan Byrd, Le Hou, Qingze Wang, Thibault Sottiaux, Michela Paganini, Jean-Baptiste Lespiau, Alexandre Moufarek, Samer Hassan, Kaushik Shivakumar, Joost van Amersfoort, Amol Mandhane, Pratik Joshi, Anirudh Goyal, Matthew Tung, Andrew Brock, Hannah Sheahan, Vedant Misra, Cheng Li, Nemanja Rakićević, Mostafa Dehghani, Fangyu Liu, Sid Mittal, Junhyuk Oh, Seb Noury, Eren Sezener, Fantine Huot, Matthew Lamm, Nicola De Cao, Charlie Chen, Sidharth Mudgal, Romina Stella, Kevin Brooks, Gautam Vasudevan, Chenxi Liu, Mainak Chain, Nivedita Melinkeri, Aaron Cohen, Venus Wang, Kristie Seymore, Sergey Zubkov, Rahul Goel, Summer

Yue, Sai Krishnakumaran, Brian Albert, Nate Hurley, Motoki Sano, Anhad Mohananey, Jonah Joughin, Egor Filonov, Tomasz Kępa, Yomna Eldawy, Jiawern Lim, Rahul Rishi, Shirin Badiezadegan, Taylor Bos, Jerry Chang, Sanil Jain, Sri Gayatri Sundara Padmanabhan, Subha Puttagunta, Kalpesh Krishna, Leslie Baker, Norbert Kalb, Vamsi Bedapudi, Adam Kurzrok, Shuntong Lei, Anthony Yu, Oren Litvin, Xiang Zhou, Zhichun Wu, Sam Sobell, Andrea Siciliano, Alan Papir, Robby Neale, Jonas Bragagnolo, Tej Toor, Tina Chen, Valentin Anklin, Feiran Wang, Richie Feng, Milad Gholami, Kevin Ling, Lijuan Liu, Jules Walter, Hamid Moghaddam, Arun Kishore, Jakub Adamek, Tyler Mercado, Jonathan Mallinson, Siddhinita Wandekar, Stephen Cagle, Eran Ofek, Guillermo Garrido, Clemens Lombriser, Maksim Mukha, Botu Sun, Hafeezul Rahman Mohammad, Josip Matak, Yadi Qian, Vikas Peswani, Pawel Janus, Quan Yuan, Leif Schelin, Oana David, Ankur Garg, Yifan He, Oleksii Duzhyi, Anton Älgmyr, Timothée Lottaz, Qi Li, Vikas Yadav, Luyao Xu, Alex Chinien, Rakesh Shivanna, Aleksandr Chuklin, Josie Li, Carrie Spadine, Travis Wolfe, Kareem Mohamed, Subhabrata Das, Zihang Dai, Kyle He, Daniel von Dincklage, Shyam Upadhyay, Akanksha Maurya, Luyan Chi, Sebastian Krause, Khalid Salama, Pam G Rabinovitch, Pavan Kumar Reddy M, Aarush Selvan, Mikhail Dektiarev, Golnaz Ghiasi, Erdem Guven, Himanshu Gupta, Boyi Liu, Deepak Sharma, Idan Heimlich Shtacher, Shachi Paul, Oscar Akerlund, François-Xavier Aubet, Terry Huang, Chen Zhu, Eric Zhu, Elico Teixeira, Matthew Fritze, Francesco Bertolini, Liana-Eleonora Marinescu, Martin Bölle, Dominik Paulus, Khyatti Gupta, Tejasi Latkar, Max Chang, Jason Sanders, Roopa Wilson, Xuewei Wu, Yi-Xuan Tan, Lam Nguyen Thiet, Tulsee Doshi, Sid Lall, Swaroop Mishra, Wanming Chen, Thang Luong, Seth Benjamin, Jasmine Lee, Ewa Andrejczuk, Dominik Rabiej, Vipul Ranjan, Krzysztof Styrc, Pengcheng Yin, Jon Simon, Malcolm Rose Harriott, Mudit Bansal, Alexei Robsky, Geoff Bacon, David Greene, Daniil Mirylenka, Chen Zhou, Obaid Sarvana, Abhimanyu Goyal, Samuel Andermatt, Patrick Siegler, Ben Horn, Assaf Israel, Francesco Pongetti, Chih-Wei "Louis" Chen, Marco Selvatici, Pedro Silva, Kathie Wang, Jackson Tolins, Kelvin Guu, Roey Yogev, Xiaochen Cai, Alessandro Agostini, Maulik Shah, Hung Nguyen, Noah Ó Donnaile, Sébastien Pereira, Linda Friso, Adam Stambler, Adam Kurzrok, Chenkai Kuang, Yan Romanikhin, Mark Geller, ZJ Yan, Kane Jang, Cheng-Chun Lee, Wojciech Fica, Eric Malmi, Qijun Tan, Dan Banica, Daniel Balle, Ryan Pham, Yanping Huang, Diana Avram, Hongzhi Shi, Jasjot Singh, Chris Hidey, Niharika Ahuja, Pranab Saxena, Dan Dooley, Srividya Pranavi Potharaju, Eileen O'Neill, Anand Gokulchandran, Ryan Foley, Kai Zhao, Mike Dusenberry, Yuan Liu, Pulkit Mehta, Ragha Kotikalapudi, Chalence Safranek-Shrader, Andrew Goodman, Joshua Kessinger, Eran Globen, Prateek Kolhar, Chris Gorgolewski, Ali Ibrahim, Yang Song, Ali Eichenbaum, Thomas Brovelli, Sahitya Potluri, Preethi Lahoti, Cip Baetu, Ali Ghorbani, Charles Chen, Andy Crawford, Shalini Pal, Mukund Sridhar, Petru Gurita, Asier Mujika, Igor Petrovski, Pierre-Louis Cedoz, Chenmei Li, Shiyuan Chen, Niccolò Dal Santo, Siddharth Goyal, Jitesh Punjabi, Karthik Kappaganthu, Chester Kwak, Pallavi LV, Sarmishta Velury, Himadri Choudhury, Jamie Hall, Premal Shah, Ricardo Figueira, Matt Thomas, Minjie Lu, Ting Zhou, Chintu Kumar, Thomas Jurdi, Sharat Chikkerur, Yenai Ma, Adams Yu, Soo Kwak, Victor Ähdel, Sujeevan Rajayogam, Travis Choma, Fei Liu, Aditya Barua, Colin Ji, Ji Ho Park, Vincent Hellendoorn, Alex Bailey, Taylan Bilal, Huanjie Zhou, Mehrdad Khatir, Charles Sutton, Wojciech Rzadkowski, Fiona Macintosh, Konstantin Shagin, Paul Medina, Chen Liang, Jinjing Zhou, Pararth Shah, Yingying Bi, Attila Dankovics, Shipra Banga, Sabine Lehmann, Marissa Bredesen, Zifan Lin, John Eric Hoffmann, Jonathan Lai, Raynald Chung, Kai Yang, Nihal Balani, Arthur Bražinskas, Andrei Sozanschi, Matthew Hayes, Héctor Fernández Alcalde, Peter Makarov, Will Chen, Antonio Stella, Liselotte Snijders, Michael Mandl, Ante Kärrman, Paweł Nowak, Xinyi Wu, Alex Dyck, Krishnan Vaidyanathan, Raghavender R, Jessica Mallet, Mitch Rudominer, Eric Johnston, Sushil Mittal, Akhil Udathu, Janara Christensen, Vishal Verma, Zach Irving, Andreas Santucci, Gamaleldin Elsayed, Elnaz Davoodi, Marin Georgiev, Ian Tenney, Nan Hua, Geoffrey Cideron, Edouard Leurent, Mahmoud Alnahlawi, Ionut Georgescu, Nan Wei, Ivy Zheng, Dylan Scandinaro, Heinrich Jiang, Jasper Snoek, Mukund Sundararajan, Xuezhi Wang, Zack Ontiveros, Itay Karo, Jeremy Cole, Vinu Rajashekhar, Lara Tumeh, Eyal Ben-David, Rishub Jain, Jonathan Uesato, Romina Datta, Oskar Bunyan, Shimu Wu, John Zhang, Piotr Stanczyk, Ye Zhang, David Steiner, Subhajit Naskar, Michael Azzam, Matthew Johnson, Adam Paszke, Chung-Cheng Chiu, Jaume Sanchez Elias, Afroz Mohiuddin, Faizan Muhammad, Jin Miao, Andrew Lee, Nino Vieillard, Jane Park, Jiageng Zhang, Jeff Stanway, Drew Garmon, Abhijit Karmarkar, Zhe Dong, Jong Lee, Aviral Kumar, Luowei Zhou, Jonathan Evens, William Isaac, Geoffrey Irving, Edward Loper, Michael Fink, Isha Arkatkar, Nanxin Chen, Izhak Shafran, Ivan Petrychenko, Zhe Chen, Johnson Jia, Anselm Levskaya, Zhenkai Zhu, Peter Grabowski, Yu Mao, Alberto Magni, Kaisheng Yao, Javier Snaider, Norman Casagrande, Evan Palmer, Paul Suganthan, Alfonso Castaño, Irene Giannoumis, Wooyeol Kim, Mikołaj Rybiński, Ashwin Sreevatsa, Jennifer Prendki, David Soergel, Adrian Goedeckemeyer, Willi Gierke, Mohsen Jafari, Meenu Gaba, Jeremy Wiesner, Diana Gage Wright, Yawen Wei, Harsha Vashisht, Yana Kulizhskaya, Jay Hoover, Maigo Le, Lu Li, Chimezie Iwuanyanwu, Lu Liu, Kevin Ramirez, Andrey Khorlin, Albert Cui, Tian LIN, Marcus Wu, Ricardo Aguilar, Keith Pallo, Abhishek Chakladar, Ginger Perng, Elena Allica Abellan, Mingyang Zhang, Ishita Dasgupta, Nate Kushman, Ivo Penchev, Alena Repina, Xihui Wu, Tom van der Weide, Priya Ponnapalli, Caroline Kaplan, Jiri Simsa, Shuangfeng Li, Olivier Dousse, Fan Yang, Jeff Piper, Nathan Ie, Rama Pasumarthi, Nathan Lintz, Anitha Vijayakumar, Daniel Andor, Pedro Valenzuela, Minnie Lui, Cosmin Paduraru, Daiyi Peng, Katherine Lee, Shuyuan Zhang, Somer Greene, Duc Dung Nguyen, Paula Kurylowicz, Cassidy Hardin, Lucas Dixon, Lili Janzer, Kiam Choo, Ziqiang Feng, Biao Zhang, Achintya Singhal, Dayou Du, Dan McKinnon, Natasha Antropova, Tolga Bolukbasi, Orgad Keller, David Reid, Daniel Finchelstein, Maria Abi Raad, Remi Crocker,

Peter Hawkins, Robert Dadashi, Colin Gaffney, Ken Franko, Anna Bulanova, Rémi Leblond, Shirley Chung, Harry Askham, Luis C. Cobo, Kelvin Xu, Felix Fischer, Jun Xu, Christina Sorokin, Chris Alberti, Chu-Cheng Lin, Colin Evans, Alek Dimitriev, Hannah Forbes, Dylan Banarse, Zora Tung, Mark Omernick, Colton Bishop, Rachel Sterneck, Rohan Jain, Jiawei Xia, Ehsan Amid, Francesco Piccinno, Xingyu Wang, Praseem Banzal, Daniel J. Mankowitz, Alex Polozov, Victoria Krakovna, Sasha Brown, MohammadHossein Bateni, Dennis Duan, Vlad Firoiu, Meghana Thotakuri, Tom Natan, Matthieu Geist, Ser tan Girgin, Hui Li, Jiayu Ye, Ofir Roval, Reiko Tojo, Michael Kwong, James Lee-Thorp, Christopher Yew, Danila Sinopalnikov, Sabela Ramos, John Mellor, Abhishek Sharma, Kathy Wu, David Miller, Nicolas Sonnerat, Denis Vnukov, Rory Greig, Jennifer Beattie, Emily Caveness, Libin Bai, Julian Eisenschlos, Alex Korchemniy, Tomy Tsai, Mimi Jasarevic, Weize Kong, Phuong Dao, Zeyu Zheng, Frederick Liu, Fan Yang, Rui Zhu, Tian Huey Teh, Jason Sanmiya, Evgeny Gladchenko, Nejc Trdin, Daniel Toyama, Evan Rosen, Sasan Tavakkol, Linting Xue, Chen Elkind, Oliver Woodman, John Carpenter, George Papamakarios, Rupert Kemp, Sushant Kafle, Tanya Grunina, Rishika Sinha, Alice Talbert, Diane Wu, Denese Owusu-Afriyie, Cosmo Du, Chloe Thornton, Jordi Pont-Tuset, Pradyumna Narayana, Jing Li, Saaber Fatehi, John Wieting, Omar Ajmeri, Benigno Uria, Yeongil Ko, Laura Knight, Amélie Héliou, Ning Niu, Shane Gu, Chenxi Pang, Yeqing Li, Nir Levine, Ariel Stolovich, Rebeca Santamaria-Fernandez, Sonam Goenka, Wenny Yustalim, Robin Strudel, Ali Elqursh, Charlie Deck, Hyo Lee, Zonglin Li, Kyle Levin, Raphael Hoffmann, Dan Holtmann-Rice, Olivier Bachem, Sho Arora, Christy Koh, Soheil Hassas Yeganeh, Siim Põder, Mukarram Tariq, Yanhua Sun, Lucian Ionita, Mojtaba Seyedhosseini, Pouya Tafti, Zhiyu Liu, Anmol Gulati, Jasmine Liu, Xinyu Ye, Bart Chrzaszcz, Lily Wang, Nikhil Sethi, Tianrun Li, Ben Brown, Shreya Singh, Wei Fan, Aaron Parisi, Joe Stanton, Vinod Koverkathu, Christopher A. Choquette-Choo, Yunjie Li, TJ Lu, Abe Ittycheriah, Prakash Shroff, Mani Varadarajan, Sanaz Bahargam, Rob Willoughby, David Gaddy, Guillaume Desjardins, Marco Cornero, Brona Robenek, Bhavishya Mittal, Ben Albrecht, Ashish Shenoy, Fedor Moiseev, Henrik Jacobsson, Alireza Ghaffarkhah, Morgane Rivière, Alanna Walton, Clément Crepy, Alicia Parrish, Zongwei Zhou, Clement Farabet, Carey Radebaugh, Praveen Srinivasan, Claudia van der Salm, Andreas Fidjeland, Salvatore Scellato, Eri Latorre-Chimoto, Hanna Klimczak-Plucińska, David Bridson, Dario de Cesare, Tom Hudson, Piermaria Mendolicchio, Lexi Walker, Alex Morris, Matthew Mauger, Alexey Guseynov, Alison Reid, Seth Odoom, Lucia Loher, Victor Cotruta, Madhavi Yenugula, Dominik Grewe, Anastasia Petrushkina, Tom Duerig, Antonio Sanchez, Steve Yadlowsky, Amy Shen, Amir Globerson, Lynette Webb, Sahil Dua, Dong Li, Surya Bhupatiraju, Dan Hurt, Haroon Qureshi, Ananth Agarwal, Tomer Shani, Matan Eyal, Anuj Khare, Shreyas Rammohan Belle, Lei Wang, Chetan Tekur, Mihir Sanjay Kale, Jinliang Wei, Ruoxin Sang, Brennan Saeta, Tyler Liechty, Yi Sun, Yao Zhao, Stephan Lee, Pandu Nayak, Doug Fritz, Manish Reddy Vuyyuru, John Aslanides, Nidhi Vyas, Martin Wicke, Xiao Ma, Evgenii Eltyshev, Nina Martin, Hardie Cate, James Manyika, Keyvan Amiri, Yelin Kim, Xi Xiong, Kai Kang, Florian Luisier, Nilesh Tripuraneni, David Madras, Mandy Guo, Austin Waters, Oliver Wang, Joshua Ainslie, Jason Baldridge, Han Zhang, Garima Pruthi, Jakob Bauer, Feng Yang, Riham Mansour, Jason Gelman, Yang Xu, George Polovets, Ji Liu, Honglong Cai, Warren Chen, XiangHai Sheng, Emily Xue, Sherjil Ozair, Christof Angermueller, Xiaowei Li, Anoop Sinha, Weiren Wang, Julia Wiesinger, Emmanouil Koukoumidis, Yuan Tian, Anand Iyer, Madhu Gurumurthy, Mark Goldenson, Parashar Shah, MK Blake, Hongkun Yu, Anthony Urbanowicz, Jennimaria Palomaki, Chrisantha Fernando, Ken Durden, Harsh Mehta, Nikola Momchev, Elahe Rahimtoroghi, Maria Georgaki, Amit Raul, Sebastian Ruder, Morgan Redshaw, Jinhyuk Lee, Denny Zhou, Komal Jalan, Dinghua Li, Blake Hechtman, Parker Schuh, Milad Nasr, Kieran Milan, Vladimir Mikulik, Juliana Franco, Tim Green, Nam Nguyen, Joe Kelley, Aroma Mahendru, Andrea Hu, Joshua Howland, Ben Vargas, Jeffrey Hui, Kshitij Bansal, Vikram Rao, Rakesh Ghiya, Emma Wang, Ke Ye, Jean Michel Sarr, Melanie Moranski Preston, Madeleine Elish, Steve Li, Aakash Kaku, Jigar Gupta, Ice Pasupat, Da-Cheng Juan, Milan Someswar, Tejvi M., Xinyun Chen, Aida Amini, Alex Fabrikant, Eric Chu, Xuanyi Dong, Amruta Muthal, Senaka Buthpitiya, Sarthak Jauhari, Nan Hua, Urvashi Khandelwal, Ayal Hitron, Jie Ren, Larissa Rinaldi, Shahar Drath, Avigail Dabush, Nan-Jiang Jiang, Harshal Godhia, Uli Sachs, Anthony Chen, Yicheng Fan, Hagai Taitelbaum, Hila Noga, Zhuyun Dai, James Wang, Chen Liang, Jenny Hamer, Chun-Sung Ferng, Chenel Elkind, Aviel Atias, Paulina Lee, Vít Listík, Mathias Carlen, Jan van de Kerkhof, Marcin Pikus, Krunoslav Zaher, Paul Müller, Sasha Zykova, Richard Stefanec, Vitaly Gatsko, Christoph Hirnschall, Ashwin Sethi, Xingyu Federico Xu, Chetan Ahuja, Beth Tsai, Anca Stefanoiu, Bo Feng, Keshav Dhandhania, Manish Katyal, Akshay Gupta, Atharva Parulekar, Divya Pitta, Jing Zhao, Vivaan Bhatia, Yashodha Bhavnani, Omar Alhadlaq, Xiaolin Li, Peter Danenberg, Dennis Tu, Alex Pine, Vera Filippova, Abhipso Ghosh, Ben Limonchik, Bhargava Urala, Chaitanya Krishna Lanka, Derik Clive, Yi Sun, Edward Li, Hao Wu, Kevin Hongtongsak, Ianna Li, Kalind Thakkar, Kuanysh Omarov, Kushal Majmundar, Michael Alverson, Michael Kucharski, Mohak Patel, Mudit Jain, Maksim Zabelin, Paolo Pelagatti, Rohan Kohli, Saurabh Kumar, Joseph Kim, Swetha Sankar, Vineet Shah, Lakshmi Ramachandruni, Xiangkai Zeng, Ben Bariach, Laura Weidinger, Tu Vu, Alek Andreev, Antoine He, Kevin Hui, Sheleem Kashem, Amar Subramanya, Sissie Hsiao, Demis Hassabis, Koray Kavukcuoglu, Adam Sadovsky, Quoc Le, Trevor Strohman, Yonghui Wu, Slav Petrov, Jeffrey Dean, and Oriol Vinyals. Gemini: A family of highly capable multimodal models, 2024. URL `https://arxiv.org/abs/2312.11805`.

Giorgi Giglemiani. Personal correspondence, Aug 2024.

Aaron Gokaslan and Vanya Cohen. Openwebtext corpus. `http://Skylion007.github.io/OpenWebTextCorpus`, 2019.

Stefan Heimersheim and Alex Turner. Residual stream norms grow exponentially over the forward pass, May 2023. URL `https://www.alignmentforum.org/posts/8mizBCm3dyc432nK8/residual-stream-norms-grow-exponentially-over-the-forward`.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Oriol Vinyals, Jack W. Rae, and Laurent Sifre. Training compute-optimal large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 9781713871088.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 448–456, Lille, France, 07–09 Jul 2015. PMLR. URL `https://proceedings.mlr.press/v37/ioffe15.html`.

Jett Janiak. Personal correspondence, Aug 2024.

Andrej Karpathy. NanoGPT. `https://github.com/karpathy/nanoGPT`, 2022.

Connor Kissane, Robert Krzyzanowski, Arthur Conmy, and Neel Nanda. Sparse autoencoders work on attention layer outputs, Jan 2024a. URL `https://www.alignmentforum.org/posts/DtdzGwFh9dCfsekZZ/sparse-autoencoders-work-on-attention-layer-outputs`.

Connor Kissane, robertzk, Arthur Conmy, and Neel Nanda. Saes (usually) transfer between base and chat models. *Alignment Forum*, 2024b. URL `https://www.alignmentforum.org/posts/fmwk6qxrpW8d4jvbd/saes-usually-transfer-between-base-and-chat-models`.

Tomek Korbak. Training a compute-optimal gpt-2 small, Oct 2022. URL `https://tomekkorbak.com/2022/10/10/compute-optimal-gpt2/`.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer Normalization. *arXiv e-prints*, art. arXiv:1607.06450, July 2016. doi: 10.48550/arXiv.1607.06450.

Callum McDougall, Arthur Conmy, Cody Rushing, Thomas McGrath, and Neel Nanda. Copy Suppression: Comprehensively Understanding an Attention Head. *arXiv e-prints*, art. arXiv:2310.04625, October 2023. doi: 10.48550/arXiv.2310.04625.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt, 2023.

Noa Nabeshima. Tinymodel, Jul 2024. URL `https://github.com/noanabeshima/tiny_model/`. GitHub repository.

Neel Nanda. Attribution patching: Activation patching at industrial scale, Mar 2023. URL `https://www.alignmentforum.org/posts/gtLLBhzQTG6nKTeCZ/attribution-patching-activation-patching-at-industrial-scale`.

Neel Nanda and Joseph Bloom. Transformerlens. `https://github.com/TransformerLensOrg/TransformerLens`, 2022.

nostalgebraist. interpreting gpt: the logit lens, Aug 2020. URL `https://www.alignmentforum.org/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens`.

Chris Olah, Josh Batson, and Shan Carter. Circuits updates — may 2023. *Transformer Circuits Thread*, May 2023. https://transformer-circuits.pub/2023/may-update/index.html#external-research.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2024.

Lee Sharkey. A technical note on bilinear layers for interpretability. *arXiv e-prints*, art. arXiv:2305.03452, May 2023. doi: 10.48550/arXiv.2305.03452.

Lee Sharkey. Sparsify: A mechanistic interpretability research agenda. *Alignment Forum*, Apr 2024. URL `https://www.alignmentforum.org/posts/64MizJXzyvrYpeKqm/sparsify-a-mechanistic-interpretability-research-agenda`.

Lee Sharkey, Dan Braun, and Beren Millidge. Taking features out of superposition with sparse autoencoders, Dec 2022. URL `https://www.alignmentforum.org/posts/z6QQJbtpkEAX3Aojj/interim-research-report-taking-features-out-of-superposition`.

Alessandro Stolfo, Ben Wu, Wes Gurnee, Yonatan Belinkov, Xingyi Song, Mrinmaya Sachan, and Neel Nanda. Confidence regulation neurons in language models. *arXiv preprint arXiv:2406.16254*, 2024.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks, 2017.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L Turner, Callum McDougall, Monte MacDiarmid, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Chris Olah, and Tom Henighan. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*, 2024. URL `https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html`.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023. URL `https://arxiv.org/abs/2302.13971`.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Simas Sakenis, Jason Huang, Yaron Singer, and Stuart Shieber. Causal Mediation Analysis for Interpreting Neural NLP: The Case of Gender Bias. *arXiv e-prints*, art. arXiv:2004.12265, April 2020. doi: 10.48550/arXiv.2004.12265.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in gpt-2 small. *arXiv preprint arXiv:2211.00593*, 2022.

Eric Winsor. Re-examining layernorm. *Alignment Forum*, 2022. URL `https://www.alignmentforum.org/posts/jfG6vdJZCwTQmG7kb/re-examining-layernorm`.

Keith Wynroe and Lee Sharkey. Decomposing the qk circuit with bilinear sparse dictionary learning, Jul 2024. URL `https://www.alignmentforum.org/posts/2ep6FGjTQoGDRnhrq/decomposing-the-qk-circuit-with-bilinear-sparse-dictionary`.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019.

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

A Appendix

We provide some samples of text generated by the no-LN GPT2-small model, and the vanilla GPT2-small model.

A.1 GPT2-small without LayerNorm (v2 model)

Two example generated texts from the no-LN GPT2-small model (v2) are shown below.

> As the last leaf fell from the tree, John realized that he was going to have to use it. “We’ve all been there. We’ve all been there,” he said. “It’s been a long time. But, it’s good to be back. It’s good to be back.” “It’s good to be back. It’s good to be back
>
> As the last leaf fell from the tree, John realized that a large amount of weight had been lifted from him. "I had a little panic attack. I was afraid that I could not walk," he said. "I felt like my legs were going to break." John has since gone back to the tree. "I have to tell you that I’m sorry I did that, but I don’t know if that will ever happen," he said.

A.2 Vanilla GPT2-small model

Two example generated texts from the vanilla GPT2-small model are shown below.

> As the last leaf fell from the tree, John realized that it was empty. He took the leaf and turned it over to his wife, who told him that it was still there and that he would have to go to the church to find it. John went to the church, and found that it was empty. He said, "I am going to the church and I am going to find the rest of the leaves, and I am going to look for them and find out where they
>
> As the last leaf fell from the tree, John realized that the tree had been torn down. As he turned his head, the other trees started to fall. "Come on," John said, "we’re going to get out of here!" The next tree was a wildflower. "How is it?" John asked, "do you see any other way?" "It’s a good thing," the other trees replied.