

Norm Growth and Stability Challenges in Localized Sequential Knowledge Editing

Akshat Gupta^{1*}, Christine Fang¹, Atahan Ozdemir¹, Maochuan Lu¹, Ahmed Alaa¹, Thomas Hartvigsen², Gopala Anumanchipalli¹

¹University of California Berkeley, ²University of Virginia

Abstract

This study investigates the impact of localized updates to large language models (LLMs), specifically in the context of knowledge editing - a task aimed at incorporating or modifying specific facts without altering broader model capabilities. We first show that across different post-training interventions like continuous pre-training, full fine-tuning and LORA-based fine-tuning, the Frobenius norm of the updated matrices always increases. This increasing norm is especially detrimental for localized knowledge editing, where only a subset of matrices are updated in a model. We reveal a consistent phenomenon across various editing techniques, including fine-tuning, hypernetwork-based approaches, and locate-and-edit methods: the norm of the updated matrix invariably increases with successive updates. Such growth disrupts model balance, particularly when isolated matrices are updated while the rest of the model remains static, leading to potential instability and degradation of downstream performance. Upon deeper investigations of the intermediate activation vectors, we find that the norm of internal activations decreases and is accompanied by shifts in the subspaces occupied by these activations, which shows that these activation vectors now occupy completely different regions in the representation space compared to the unedited model. With our paper, we highlight the technical challenges with continuous and localized sequential knowledge editing and their implications for maintaining model stability and utility.

Introduction

Recent advances in model interpretability have led to methods that perform localized updates to large language models by intervening at very specific locations (Meng et al. 2022a,b; Hernandez, Li, and Andreas 2023). A popular domain in which such methods are employed is knowledge editing (Yao et al. 2023) - a task where singular facts are added or updated inside of a model in a data and compute efficient manner. This paper starts with a simple yet powerful observation that sequential knowledge updates made to a model always leads to an increase in the norm of the matrix being updated. We first ask the question - is the increase in norm with continuous updates specific to localized knowledge editing methods or is this a general phenomenon?

*Correspondence to: akshat.gupta@berkeley.edu
Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Our experiments with numerous post-training interventions, including continual pretraining, full fine-tuning and LORA-based fine-tuning, present a very surprising result - **the norm of the weight matrices being updated always increases for all these post-training interventions**. While there has been dispersed work showing norm growth (Merrill et al. 2020), to the best of our knowledge, our study is the first to empirically evaluate this comprehensively for large number of important post-training interventions.

We next study the presence and implications of this phenomenon when performing localized updates. To do so we study the task of knowledge editing where localized updates are very common. Knowledge editing methods usually update specific parts of the model, for example the MLP sub-modules of certain layers, to add or update new information. This allows data and compute efficient updates to be made to a model. We perform continuous sequential knowledge edits to a model using various parameter modifying knowledge editing methods along with localized fine-tuning, and find that for all these scenarios, the norm of the edited matrix always increases with the number of updates. While the increasing norm may not be concerning in general, it is especially detrimental for performing localized updates. This is because disproportionate and continuous growth in the norm of one or few layers of a model, while the rest of the model remains frozen, will compromise the balance and stability of the entire system, eventually leading to a breaking point as observed in prior work (Gupta, Rao, and Anumanchipalli 2024; Gupta, Baskaran, and Anumanchipalli 2024). This disproportionate growth is shown in Figure 3.

We further analyze the effects of performing continuous localized knowledge updates to a model by studying how the hidden activations of the model changes. We find that contrary to the increasing norm of edited matrix, the norm of the activation vectors generated after the edited layers continuously decreases. We also show that these activation begin to occupy different regions in space when compared to the original. A follow-up to our work by Gupta et al. (2025) resolve the problems of disproportionate norm growth by using regularization methods and propose a more robust knowledge editing method called ENCORE.

To summarize, we make the following contributions in this paper:

1. We show that the frobenius norm of the updated weight

- matrices always increases during post training interventions.
2. The norm of edited matrix increases disproportionately for localized knowledge updates, leading to model collapse.
 3. This collapse is accompanied by a change in the norm and orientations of the resultant hidden activations, showing that the activations of the edited models now lie in a different region of representation space.

Norm Growth During Post-Training Interventions

We focus on the following common interventions that are applied on a model after the pretraining step - continual pre-training, full fine-tuning, LORA based fine-tuning (Hu et al. 2021). We discuss our experiment settings for each of the following below:

1. **Continued pretraining (CPT)** - We consider continued pretraining as a separate case from full fine-tuning although in both cases all the weights of the model are updated using a next-token prediction loss. We define continued pretraining as a process where the foundational knowledge of a model is extended by training on a large domain-specific corpora. In our experiments, we present the results for performing CPT on 20 billion tokens for Python programming (Li et al. 2023a) on Llama-2 (7B) (Touvron et al. 2023).
2. **Full Fine-Tuning (FFT)** - We define full fine-tuning as task specific next-token prediction training of a model to optimize the model's parameters on a particular task. We present the results for fine-tuning Llama-2 (7B) model on 110k question answer pairs for programming (Wei et al. 2024).
3. **LORA based full fine-tuning (LFFT)** - Here we use LORA (Hu et al. 2021) to fine-tune all the model weights in the same setting as FFT.

For each of the above, we present post-training intervention results using the checkpoints provided in the study by Biderman et al. (2024). In each of the above cases, the model update equation can be written as:

$$W_{new} = W_{old} + \Delta W \quad (1)$$

Thus, the norm of the new weight matrix does not necessarily have to decrease and can lie in the range as shown below using the triangle inequality:

$$\|W_{old}\|_F - \|\Delta W\|_F \leq \|W_{new}\|_F \leq \|W_{old}\|_F + \|\Delta W\|_F \quad (2)$$

which means that after each update, the norm of the matrix being updated can both decrease or increase. Yet we find that the norm of the updated matrix in each of these interventions always increases. This can be seen in Figure 1. The frobenius norm of all three MLP¹ and attention matrices

¹Note that the Llama architecture has three MLP matrices instead of the common practice of two.

in Llama-2 (7B) can be seen to increase during CPT, FFT and LFFT. While this increase in norm is not detrimental to model performance, we believe this is because the norm of all the weight matrices involved increases in conjunction with each other.

While our study is not exhaustive across different types of models or datasets used, we use the above findings to set the stage for the coming sections and to motivate studying norms of edited matrices when performing localized updates to a model. As only a few weight matrices are updated during parameter modifying knowledge editing methods, this phenomenon can have adverse consequences when the norm of some intermediate matrices grow while the remaining model remains frozen.

Localized Updates during Knowledge Editing

Knowledge editing is defined as the task of making data and compute efficient knowledge updates to large language model without compromising their general ability (Yao et al. 2023; Kolbeinsson et al. 2024). In this paper, we focus on parameter-modifying knowledge editing methods, where knowledge is updated by changing the weights of the model (Meng et al. 2022a,b; Gupta, Sajnani, and Anumanchipalli 2024). The compute efficient component of knowledge editing comes from the fact that usually only one or a few layers of a model are updated when incorporating new knowledge. In this paper we focus on four popular knowledge editing methods - **ROME** (Meng et al. 2022a), **MEMIT** (Meng et al. 2022b), **MEND** (Mitchell et al. 2021) and **PMET** (Li et al. 2023b). We perform sequential edits using these methods on GPT2-XL (1.5B) (Radford et al. 2019) and GPT-J (6B) (Wang and Komatsuaki 2021) for 2000 edits and analyze the different properties of the updated matrices. The list of layers updated for the different model editing algorithms can be seen in Table 1.

Algorithm	GPT2-XL	GPT-J
ROME	17	5
MEMIT	13-17	3-8
MEND	45-47	25-27
PMET	13-17	3-8

Table 1: List of layers edited when using each of the above algorithms for GPT2-XL and GPT-J.

As can be seen in Figure 2, the norm of the edited matrices always increases for all four types of model editing methods used. Figure 2 also shows evidences of model degradation as a function of sequential model editing evaluated on 8 downstream tasks including MMLU (Hendrycks et al. 2020) and a few tasks from the GLUE benchmark (Wang et al. 2018). A more detailed account of downstream measurement is provided in appendix . To put the norm growth in perspective of the rest of the model, we plot the norm of the edited matrix after 100, 500 and 2000 edits along with the norm of other matrices present inside the model in Figure 3. We clearly see the anomalous growth in norm for the edited layers, while the norm of the remaining layers remains constant.

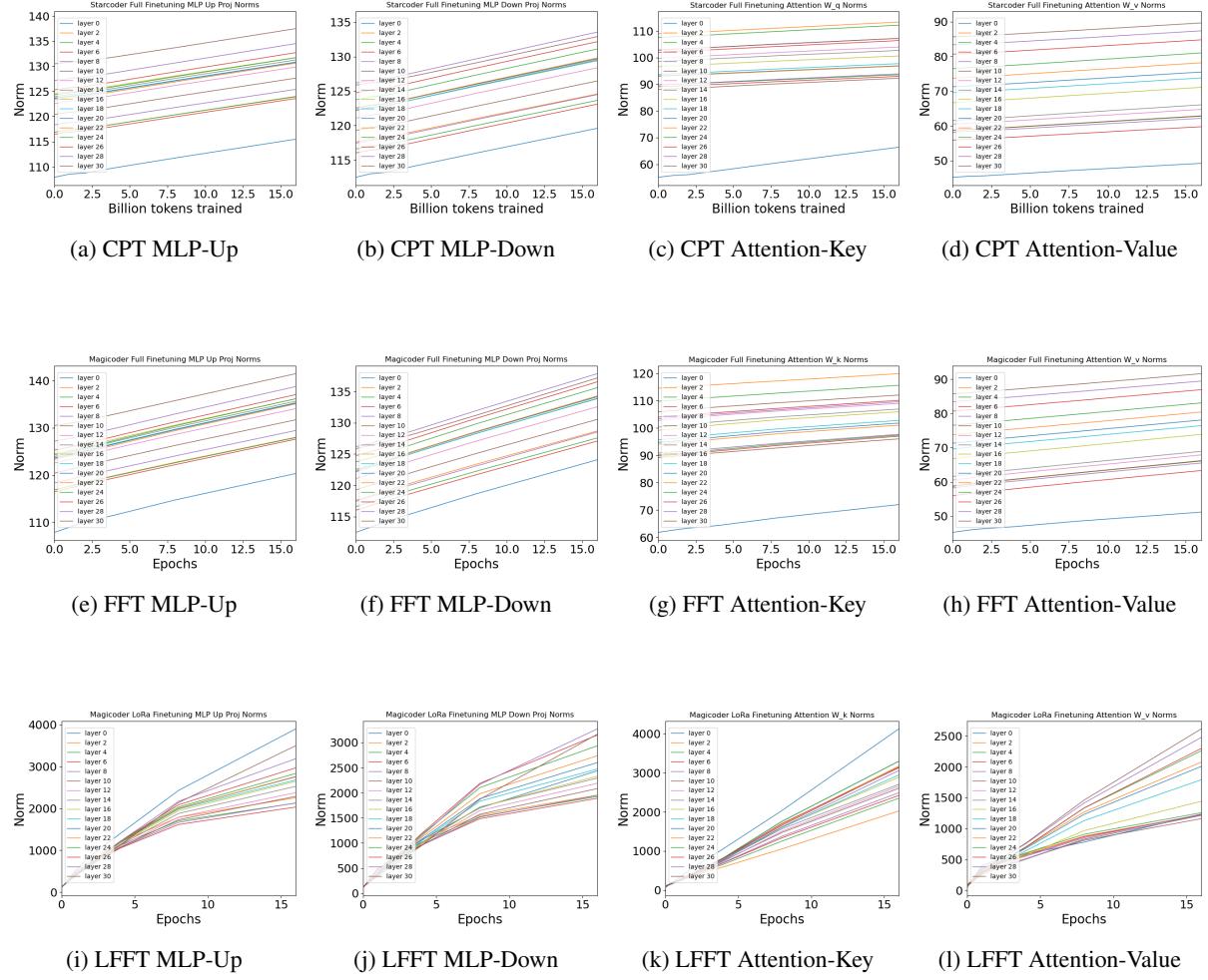


Figure 1: Norm growth during post-training interventions

Effect on Norm of Internal Activations. To understand further effects of the increase norm of the edited matrix, we look at its effect on the corresponding activations that are output from that layer. To do so, we send 1 million tokens of wikipedia articles through both GPT2-XL and GPT-J, and study the norm and orientations of the internal activations of the model before and after editing. Specifically, we are looking at the residual stream vectors after each layers inside an LLM. The norm of the activation before and after editing can be seen in Figure 4. In this figure, we present the norm of the activation vectors for GPT2-XL when editing the model using ROME, comparing the unedited model with the model after 100, 500 and 2000 edits. We see that the activation norms remains the same until layer 17 for all cases, but slowly starts to decrease as we edit the model more. After 2000 edits, the norm of the activation vectors at layer 40 is much almost have that of the original enedited model, showing that the decrease in norm compounds as the activation vectors pass through the model. In contrast to the increasing norm of the edited weight matrices, the average

norm of the activations decreases post editing. While in this paper we do not analyze the implications of this, it is studied in more detail in Gupta et al. (2025). They show that the norms of activation vectors generated from the edited matrices increase while the norms of activations from the subsequent layers decreases, resulting in an increased importance of vectors generated from those layers.

Effect on Orientation of Internal Activations. We further analyze the orientations of the activation vectors. To do so, we create classifiers between the residual stream vectors of layers “ i ” versus “ j ”. Each classifier is a simple binary logistic regression classifier. A classifier $C_{i,j}$ is a binary classifier between two groups of vectors - residual stream vectors between layer i and residual stream vectors between layer j . The classifier is trained with 800k training vectors and tested on 200k test vectors, equally divided into the two classes. This classifiers are trained only on the activations of the un-edited model. The accuracy of the classifier can be seen in Figure 5 (a). We see that the simple binary classifiers

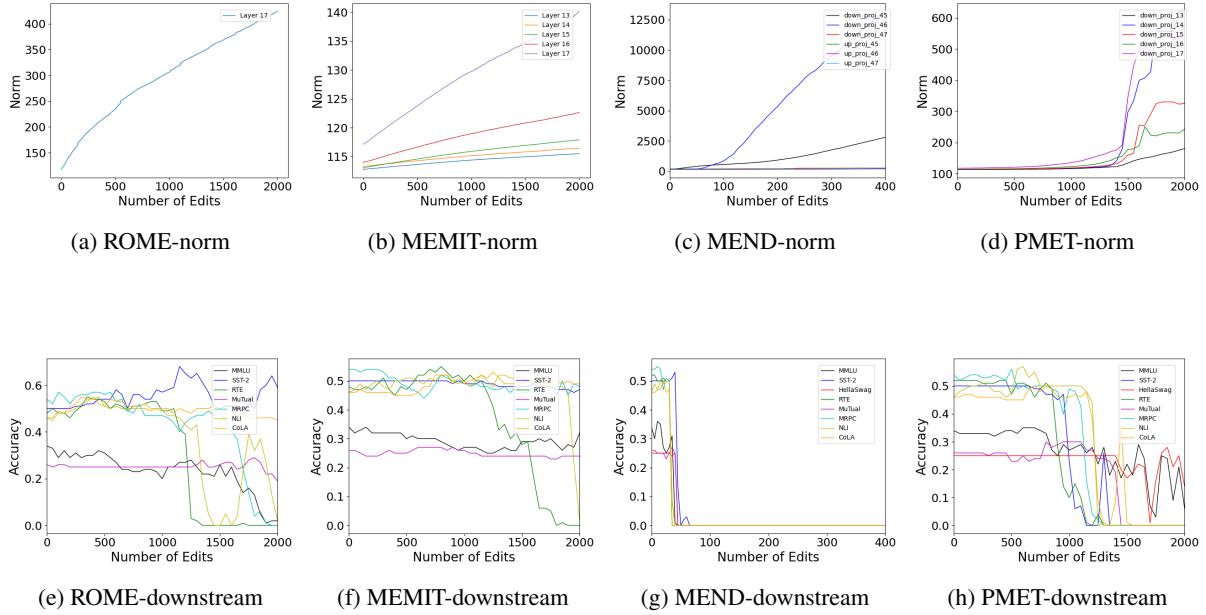


Figure 2: Norm growth and downstream performance during knowledge editing on GPT2-XL for different methods.

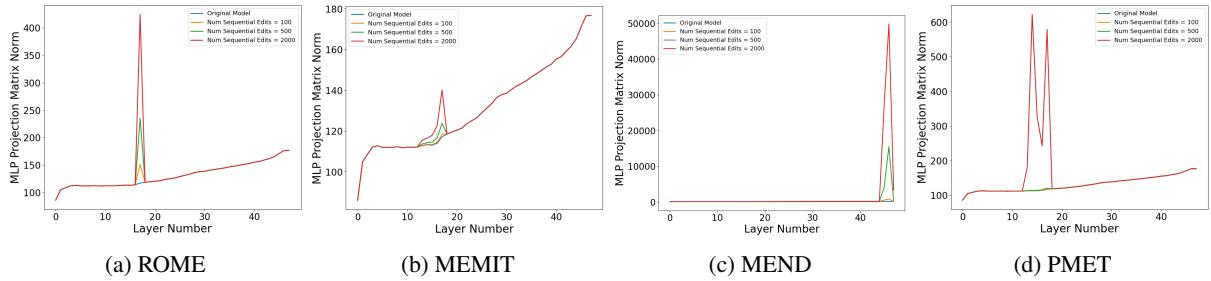


Figure 3: Norm growth for edits 100, 500, 2000 for GPT2-XL.

reach almost a 100% accuracy when classifying the activation vectors from different layers. **This shows that the activation vectors of different layers are linearly separable from each other.** When a classifier is trained with activation vectors from the same layers, which forms the diagonal line in Figure 5 (a), we see that the classification accuracy is 50% or at random. This shows that the activation vectors between the same layers are not linearly separable, strengthening the linear separability claim.

We save the classifiers trained on the un-edited model and use them to classify the activation vectors of the edited model. Figure 5 shows this for ROME, where we analyze the output for 200k activation vectors when passed through a model sequentially edited for 100, 500 and 2000 edits. The cell $i - j$ corresponds to the case where classifier $C_{i,j}$, trained on the activations coming from the i^{th} and j^{th} layer of the un-edited model, are used to classify the activations of the j^{th} layer of the edited model. We want to re-iterate to the reader that the graphs in Figure 5 are not supposed to be symmet-

rical by definition. We will give an example below. For cell $3-5$, we use the classifier $C_{3,5}$ trained between activations of layer 3 and layer 5 of the un-edited model, but use it to classify the activations of layer 5 at test time. For cell $5-3$, we use the same classifier $C_{3,5}$, but this time we use it to classify the activations of layer 3.

For cell $i - j$, if the orientations in space of the activation vectors of the edited models remains the same as that of the unedited model, the classifier $C_{i,j}$ would assign it the class j . But we see in Figure 5 that this is not the case. The test time accuracy of the classifier begins to get worse for activation vectors right after the edited layer. This shows that the activation vectors are not only changing in norms but also the regions they occupy in space. This is so much so that the activations coming out of layer j of the edited model after 2000 edits now lie in a region for layer i , leading to an incorrect classification accuracy for classifier $C_{i,j}$.

With this study, we show that the increasing norm not only changes the norm of the internal activations of the model, but

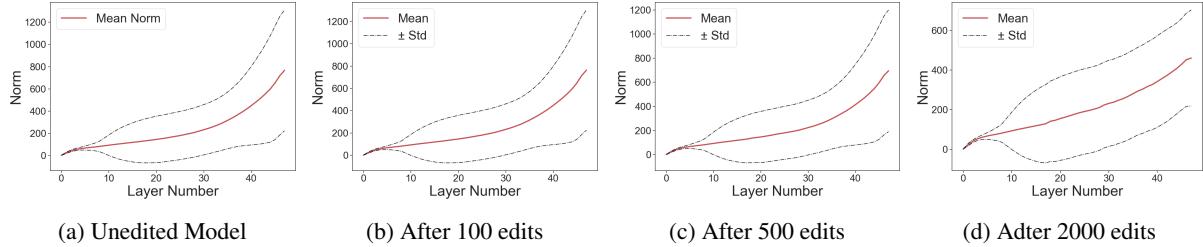


Figure 4: Activation Norms at different layers for unedited, edits 100, 500, 2000 for GPT2-XL/ROME.

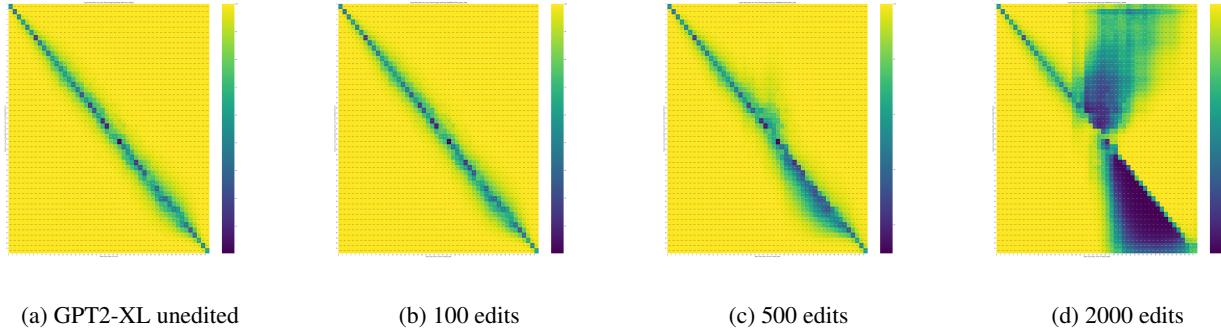


Figure 5: Classification plots comparing GPT2-XL unedited model with model after 100, 500, 2000 edits using ROME.

also the orientations in space in which these vectors lie. We observe that post-editing, the internal activations now lie in a very different region in space. The region is so different that a linear classifier trained to identify the activation vectors of a specific layer is now unable to identify the output activations from the specific layer.

Conclusion

This study highlights a critical challenge in the domain of localized updates for large language models: the persistent increase in the norm of updated matrices during sequential knowledge editing. While this phenomenon appears universal across post-training intervention methods, its implications are particularly pronounced in localized knowledge editing, where only specific parts of the model are modified. The resulting imbalance leads to downstream performance degradation, as evidenced by changes in both the norm and orientation of internal activations. Our findings emphasize the need for innovative strategies to address these challenges, paving the way for more robust and sustainable approaches to localized knowledge editing in LLMs. This work serves as a foundational step towards understanding and mitigating the inherent limitations of current techniques, with the ultimate goal of enabling dynamic and scalable updates to pre-trained models. A follow-up work by Gupta et al. (2025) overcomes these highlighted limitations by using appropriate regularizations with knowledge editing, allowing long-term sequential knowledge editing.

References

- Bentivogli, L.; Clark, P.; Dagan, I.; and Giampiccolo, D. 2009. The Fifth PASCAL Recognizing Textual Entailment Challenge. *TAC*, 7: 8.
- Biderman, D.; Portes, J.; Ortiz, J. J. G.; Paul, M.; Greengard, P.; Jennings, C.; King, D.; Havens, S.; Chiley, V.; Frankle, J.; et al. 2024. Lora learns less and forgets less. *arXiv preprint arXiv:2405.09673*.
- Cui, L.; Wu, Y.; Liu, S.; Zhang, Y.; and Zhou, M. 2020. MuTual: A Dataset for Multi-Turn Dialogue Reasoning. In *Proceedings of the 58th Conference of the Association for Computational Linguistics*. Association for Computational Linguistics.
- Dagan, I.; Glickman, O.; and Magnini, B. 2005. The pascal recognising textual entailment challenge. In *Machine learning challenges workshop*, 177–190. Springer.
- Dolan, B.; and Brockett, C. 2005. Automatically constructing a corpus of sentential paraphrases. In *Third International Workshop on Paraphrasing (IWP2005)*.
- Giampiccolo, D.; Magnini, B.; Dagan, I.; and Dolan, W. B. 2007. The third pascal recognizing textual entailment challenge. In *Proceedings of the ACL-PASCAL workshop on textual entailment and paraphrasing*, 1–9.
- Gupta, A.; Baskaran, S.; and Anumanchipalli, G. 2024. Rebuilding rome: Resolving model collapse during sequential model editing. *arXiv preprint arXiv:2403.07175*.

- Gupta, A.; Prateepamornkul, P.; Lu, M.; Alaa, A.; Hartvigsen, T.; and Anumanchipalli, G. 2025. Lifelong Sequential Knowledge Editing without Model Degradation. *arXiv preprint arXiv:2502.01636*.
- Gupta, A.; Rao, A.; and Anumanchipalli, G. 2024. Model Editing at Scale leads to Gradual and Catastrophic Forgetting. *arXiv preprint arXiv:2401.07453*.
- Gupta, A.; Sajnani, D.; and Anumanchipalli, G. 2024. A unified framework for model editing. *arXiv preprint arXiv:2403.14236*.
- Haim, R. B.; Dagan, I.; Dolan, B.; Ferro, L.; Giampiccolo, D.; Magnini, B.; and Szpektor, I. 2006. The second pascal recognising textual entailment challenge. In *Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment*, volume 7, 785–794.
- Hendrycks, D.; Burns, C.; Basart, S.; Zou, A.; Mazeika, M.; Song, D.; and Steinhardt, J. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.
- Hernandez, E.; Li, B. Z.; and Andreas, J. 2023. Inspecting and editing knowledge representations in language models. *arXiv preprint arXiv:2304.00740*.
- Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.
- Kolbeinsson, A.; O’Brien, K.; Huang, T.; Gao, S.; Liu, S.; Schwarz, J. R.; Vaidya, A.; Mahmood, F.; Zitnik, M.; Chen, T.; et al. 2024. Composable interventions for language models. *arXiv preprint arXiv:2407.06483*.
- Li, R.; Allal, L. B.; Zi, Y.; Muennighoff, N.; Kocetkov, D.; Mou, C.; Marone, M.; Akiki, C.; Li, J.; Chim, J.; et al. 2023a. Starcoder: may the source be with you! *arXiv preprint arXiv:2305.06161*.
- Li, X.; Li, S.; Song, S.; Yang, J.; Ma, J.; and Yu, J. 2023b. Pmet: Precise model editing in a transformer. *arXiv preprint arXiv:2308.08742*.
- Meng, K.; Bau, D.; Andonian, A.; and Belinkov, Y. 2022a. Locating and editing factual associations in GPT. *Advances in Neural Information Processing Systems*, 35: 17359–17372.
- Meng, K.; Sharma, A. S.; Andonian, A.; Belinkov, Y.; and Bau, D. 2022b. Mass-editing memory in a transformer. *arXiv preprint arXiv:2210.07229*.
- Merrill, W.; Ramanujan, V.; Goldberg, Y.; Schwartz, R.; and Smith, N. 2020. Effects of parameter norm growth during transformer training: Inductive bias from gradient descent. *arXiv preprint arXiv:2010.09697*.
- Mitchell, E.; Lin, C.; Bosselut, A.; Finn, C.; and Manning, C. D. 2021. Fast model editing at scale. *arXiv preprint arXiv:2110.11309*.
- Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8): 9.
- Socher, R.; Perelygin, A.; Wu, J.; Chuang, J.; Manning, C. D.; Ng, A. Y.; and Potts, C. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, 1631–1642.
- Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023. Llama 2: Open foundation and fine-tuned chat models, 2023. URL <https://arxiv.org/abs/2307.09288>.
- Wang, A.; Singh, A.; Michael, J.; Hill, F.; Levy, O.; and Bowman, S. R. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.
- Wang, B.; and Komatsu, A. 2021. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. <https://github.com/kingoflolz/mesh-transformer-jax>.
- Warstadt, A.; Singh, A.; and Bowman, S. R. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7: 625–641.
- Wei, Y.; Wang, Z.; Liu, J.; Ding, Y.; and Zhang, L. 2024. Magicoder: Empowering code generation with oss-instruct. In *Forty-first International Conference on Machine Learning*.
- Yao, Y.; Wang, P.; Tian, B.; Cheng, S.; Li, Z.; Deng, S.; Chen, H.; and Zhang, N. 2023. Editing Large Language Models: Problems, Methods, and Opportunities. *arXiv preprint arXiv:2305.13172*.
- Zellers, R.; Holtzman, A.; Bisk, Y.; Farhadi, A.; and Choi, Y. 2019. HellaSwag: Can a Machine Really Finish Your Sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

A.1 Downstream Performance Measurement

In this paper, we assess model degradation by measuring downstream performance at regular intervals of edits. Our evaluation suite is wide-ranging and consists of the following 8 tasks – sentiment analysis (SST2) (Socher et al. 2013), paraphrase detection (MRPC) (Dolan and Brockett 2005), natural language inference (NLI, RTE) (Dagan, Glickman, and Magnini 2005; Haim et al. 2006; Giampiccolo et al. 2007; Bentivogli et al. 2009), commonsense natural language inference (HellaSwag) (Zellers et al. 2019), linguistic acceptability classification (CoLA) (Warstadt, Singh, and Bowman 2019), multi-turn dialogue reasoning (MuTual) (Cui et al. 2020) and massive multitask language understanding (MMLU) (Hendrycks et al. 2020).

For each task, we created a subset of 100 examples balanced across all multiple-choice options. The models were evaluated on the tasks above, and the accuracy score was measured at intervals of 5 edits for MEND, and 20 edits for PMET, ROME, and MEMIT. MEND leads to a rapid degradation within 100 edits, hence why a smaller granularity of 5 edit interval was used; a 20 edit interval was used for other methods to cut down on computation time.

In order to improve models’ initial performance and achieve meaningful signals, we provided few-shot examples. The few-shot prompt templates used for each task are shown in Figures 6-13.

Review : an exhilarating futuristic thriller-noir , minority report twists the best of technology around a gripping story , delivering a riveting , pulse intensifying escapist adventure of the first order
 Sentiment : positive

Review : try as i may , i can't think of a single good reason to see this movie , even though everyone in my group extemporaneously shouted , ' thank you ! '
 Sentiment : negative

Review : the film 's performances are thrilling .
 Sentiment : positive

Review : vera 's technical prowess ends up selling his film short ; he smoothes over hard truths even as he uncovers them .
 Sentiment : negative

Review : [input]
 Sentiment :

Figure 6: Few shot prompt template used for SST-2

Question: Which expression is equivalent to 4×9 ?
 (A) $(4 \times 4) + (4 \times 5)$
 (B) $(4+4) \times (4+5)$
 (C) $(4+4)+(4+5)$
 (D) $(4 \times 4) \times (4 \times 5)$
 Answer: A

Question: A marketing researcher is conducting a survey in a large selling area by contacting a small group of people that is representative of all people in that area. The small, representative group is known as the
 (A) population
 (B) sample
 (C) stratification
 (D) universe
 Answer: B

Question: A research participant eats half a bowl of M&M candies, and then stops eating. How would a motivation researcher using drive reduction theory explain this participant's behavior?
 (A) Humans are instinctively driven to eat sugar and fat when presented to them.
 (B) The Yerkes-Dodson law explains that people will eat food when presented to them, but usually in moderate amounts in order to avoid being perceived as selfish.
 (C) The primary drive of hunger motivated the person to eat, and then stop when she/he regained homeostasis.
 (D) The research participant was satisfying the second step on the hierarchy of needs: Food needs.
 Answer: C

Question: In a deductively valid argument
 (A) If all the premises are true, the conclusion must be true
 (B) The conclusion has already been stated in its premises
 (C) If all the premises are true, the conclusion may still be false
 (D) Both A and B
 Answer: D

Question: [input]
 Answer:

Figure 7: Few shot prompt template used for MMLU

Are the sentences paraphrases of each other.
Sentence 1: Federal regulators have turned from sour to sweet on a proposed \$2.8 billion merger of ice cream giants Nestle Holdings Inc. and Dreyer 's Grand Ice Cream Inc .
Sentence 2: Federal regulators have changed their minds on a proposed \$2.8 billion merger of ice cream giants Nestle Holdings and Dreyer 's Grand Ice Cream .
Answer: Yes

Are the sentences paraphrases of each other.
Sentence 1: In the year-ago quarter , the steelmaker recorded a profit of \$16.2 million , or 15 cents per share , on sales of \$1.14 billion .
Sentence 2: In the second quarter last year , AK Steel reported a profit of \$16.2 million , or 15 cents a share .
Answer: No

Are the sentences paraphrases of each other.
Sentence 1: He added : "I 've never heard of more reprehensiblebehaviour by a doctor .
Sentence 2: The Harrisons ' lawyer Paul LiCalsi said : " I ' ve never heard of more reprehensible behaviour by a doctor .
Answer: Yes

Are the sentences paraphrases of each other.
Sentence 1: While dioxin levels in the environment were up last year , they have dropped by 75 percent since the 1970s , said Caswell .
Sentence 2: The Institute said dioxin levels in the environment have fallen by as much as 76 percent since the 1970s .
Answer: No

Are the sentences paraphrases of each other.
Sentence 1: [input 1]
Sentence 2: [input 2]
Answer:

Figure 8: Few shot prompt template used for MRPC

Is this sentence linguistically acceptable?
Sentence: The evidence assembled by the prosecution convinced the jury.
Answer: Yes

Is this sentence linguistically acceptable?
Sentence: I live at the place where Route 150 crosses the Hudson River and my dad lives at it too.
Answer: No

Is this sentence linguistically acceptable?
Sentence: The government's imposition of a fine.
Answer: Yes

Is this sentence linguistically acceptable?
Sentence: Sam gave the ball out of the basket.
Answer: No

Is this sentence linguistically acceptable?
Sentence: [input]
Answer:

Figure 9: Few shot prompt template used for RTE

The town is also home to the Dalai Lama and to more than 10,000 Tibetans living in exile.

Question: The Dalai Lama has been living in exile since 10,000. True or False?

Answer: True

P. Prayong, who like Kevala belongs to the Theravada sect of Buddhism, chose India over other Buddhist majority nations as it is the birthplace of Gautama Buddha.

Question: P. Prayong is a member of Theravada. True or False?

Answer: False

The medical student accused of murdering an erotic masseuse he met on Craigslist is drowning in more than \$100,000 in student loan debt and is so broke he can't afford to pay an attorney, according to court papers. Philip Markoff, a 23-year-old suspended Boston University medical school student, owes \$130,000 in student loans and does not get money from his parents, leaving him to lean on a taxpayer-funded attorney for his defense, according to a court document in Boston Municipal Court that labels him indigent. Markoff graduated from the State University of New York-Albany and was a second-year medical student at BU.

Question: The medical student Philip Markoff was engaged. True or False?

Answer: True

Traditionally, the Brahui of the Raisani tribe are in charge of the law and order situation through the Pass area. This tribe is still living in present day Balochistan in Pakistan.

Question: The Raisani tribe resides in Pakistan. True or False?

Answer: False

The latest attacks targeted the U-S embassy and a top prosecutor's office in the Uzbek capital.

Question: [input]. True or False?

Answer:

Figure 10: Few shot prompt template used for CoLA

Turkey is unlikely to become involved in, or allow U.S. forces to use Turkish territory in a Middle East war that does not threaten her territory directly. entails the U.S. to use Turkish military bases.

True or False?

Answer: False

Brooklyn Borough Hall featured a Who's Who in New York's literary community during the second annual Brooklyn Book Festival. According to Brooklyn Borough President Marty Markowitz, the borough's zip code 11215 boasts more authors than anywhere else in the country. It appeared to be the case on Sunday. More than 100 authors were featured at the day-long event, including The Basketball Diaries writer Jim Carroll, former M*A*S*H star Mike Farrell, author and illustrator Mo Willems, Jack Kerouac's sometime lover and National Book Critics Circle Award recipient Joyce Johnson and PEN American Center President Francine Prose. entails the The Brooklyn Book Festival is held in Brooklyn Borough every year.

True or False?

Answer: True

NASA's Saturn exploration spacecraft, Cassini , has discovered an atmosphere about the moon Enceladus . This is the first such discovery by Cassini, other than Titan , of the presence of an atmosphere around a Saturn moon. entails the Titan is the fifteenth of Saturn's known satellites.

True or False?

Answer: False

Dr. Eric Goosby, a pioneer in the fight against AIDS, is President Obama's choice to run the American effort to combat the disease globally, the White House announced Monday. The President's Emergency Plan For AIDS Relief, known as Pepfar, was championed by President George W. Bush. It is expected to spend \$48 billion over the next five years and is credited with markedly reducing the disease's death rate. Its prevention policy has been controversial because of its emphasis on socially conservative methods. With a new administration and a Democratic majority in the House, organizations seeking prevention choices beyond abstinence and fidelity — including a renewed commitment to distributing condoms — are eager to try to rewrite the guidelines. entails the Pepfar is committed to fighting AIDS.

True or False?

Answer: True

[input]

True or False?

Answer:

Figure 11: Few shot prompt template used for NLI

Given the following:

f : hi how are you doing ?

m : i 've been good . i 'm in school right now .

f : what school do you go to ?

m : i go to a cooking school . i will spend one year there .

f : really ? i know you love drawing and designing most . how do you like cooking so far ?

m : i like it so far , my classes are pretty good , and i plan to have my own restaurant in the future .

Which choice is correct?

(A)f : really ? you mean you do n't like cooking but you plan to start a restaurant in the future ?

(B)f : really ? you mean your classes are pretty good and you plan to start a restaurant in the future ?

(C)f : so , although your classes are not pretty good , you plan to become a teacher in the future ?

(D)f : so , although you do n't love drawing or designing , you want to design a building in the future ?

Answer: B

Given the following:

f : dad , can i go out tonight ?

m : no , i 'm sorry . you ca n't .

f : can i ask nancy for dinner ?

m : ok , but you ca n't let your brother alone .

Which choice is correct?

(A)f : ok. then i will ask nancy for dinner tonight .

(B)f : i will stay at home alone because i do n't want ask nancy for dinner .

(C)f : ok. so i can ask nancy for dinner tonight if i do n't have to have my brother companied .

(D)f : i have to stay home with me brother because i will not ask nancy to have dinner .

Answer: A

Given the following:

m : that was such an interesting english program , i wish you enjoyed it as much as i did .

f : i must tell you the truth that i fell asleep after the first few minutes , as i could n't understand many of the words in the program .

Which choice is correct?

(A)m : seems that you found the program boring .

(B)m : the english program is indeed difficult . i feel it , too !

(C)m : so you think the english program is difficult . do n't worry , let me help you .

(D)m : good to know that you also found it interesting .

Answer: C

Given the following:

m : should n't we invite kathy to the party tonight ?

f : invite kathy ? she is the one who 's planning the whole thing .

Which choice is correct?

(A)m : i have invited cathy .

(B)m : cathy planed the party . but she wo n't attend it cause she has no time .

(C)m : what a pity ! cathy is too busy to come .

(D)m : cathy planed the party . of course she will attend it .

Answer: D

Q: Given the following: [input]

Which choice is correct?

(A)

(B)

(C)

(D)

Answer:

Figure 12: Few shot prompt template used for MuTual

Activity: Hitting a pinata

Finish this sentence: A small girl runs away from the pinata while holding the stick. The child returns and hits the pinata one time. the people

Options:

- (A) continue to remove key ingredients from the bags.
- (B) cheer for the child getting hit.
- (C) watch while the girl throws the stick twice.
- (D) all clap and the girl smiles and turns toward the camera.

Answer: D

Activity: Personal Care and Style

Finish this sentence: [header] How to apply base makeup [title] Use concealer. [step] You will wreck your look if you've got dark circles under your eyes or have blemishes throughout your face. Don't feel badly about such circles or blemishes, though.

Options:

- (A) This is a natural part of makeup, and you can use that to your advantage. If you must draw your brows in, that will be where most of the concealer and powder will come in.
- (B) You can always cover them up if need be. You can also apply some false lashes-it will look better, even if you only have the eyelashes.
- (C) You can cover them up with concealer. Apply the concealer in an upside-down pyramid shape.
- (D) [substeps] Use some concealer up off of your eyelids so you can wear concealer successfully without using too much. Apply concealer a little higher up on your eyelid than the same way.

Answer: C

Activity: Vacuuming floor

Finish this sentence: The video shows a demonstration of dyson vacuum cleaner and how well it can clean particles from the floor. There's rice grain, flour and other food particles scattered on the floor. the demonstrator

Options:

- (A) then kneels down next to the tested vacuum cleaner and then goes on the floor to clean the floor using a towel.
- (B) turns on the vacuum cleaner and begins vacuuming the floor with the yellow dyson vacuum.
- (C) wipes down the floor with towels.
- (D) talks about a very hard floor that is soft and easy to clean.

Answer: B

Activity: Personal Care and Style

Finish this sentence: [header] How to grow an afro with african american hair [title] Use the right comb. [step] Traditional combs and brushes will damage your curly hair, making it frizzier. Instead of these, use a wide-tooth comb, afro pick, or even your fingers to comb your hair.

Options:

- (A) [substeps] Don't comb more than necessary! Just use your comb, pick, or fingers to get rid of any tangles. If you have a lot of tangles, try buying a detangling spray specifically designed for african american hair.
- (B) Most african american hair is fairly straight, but with wavy hair that isn't straight, you have a naturally naturally curly afro. [substeps] If you have short hair, use a comb with some thinner bristles to help tame your hair and prevent frizz.
- (C) Using a large comb will really work as well to take care of your hair without damaging its natural oils. [substeps] Small combs are best because more tangles can be too rough for curly hair.
- (D) If you have very curly hair, you may prefer using a cap or a hair net instead, as these tend to do little damage instead of much. [title] Trim your hair every six to eight months.

Answer: A

Activity: [input]

Finish this sentence: [input]

Options:

- (A)
- (B)
- (C)
- (D)

Answer:

Figure 13: Few shot prompt template used for HellaSwag

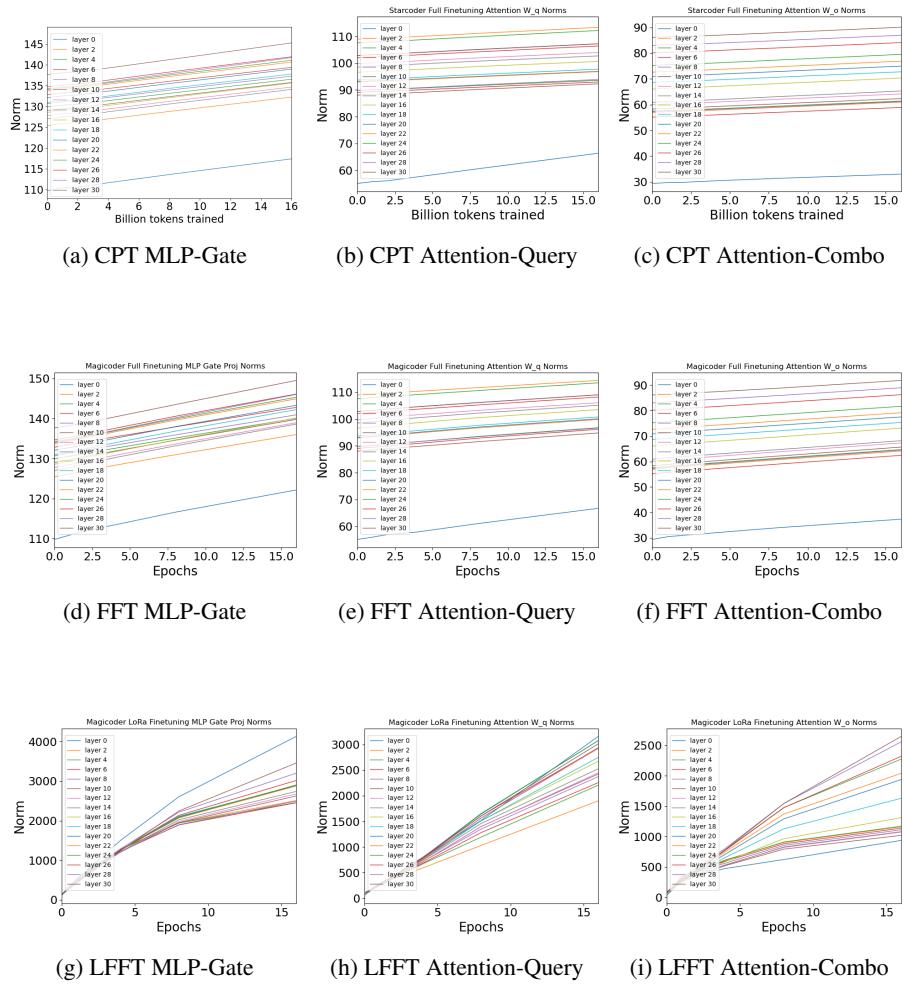


Figure 14: Additional figures for post-training interventions.

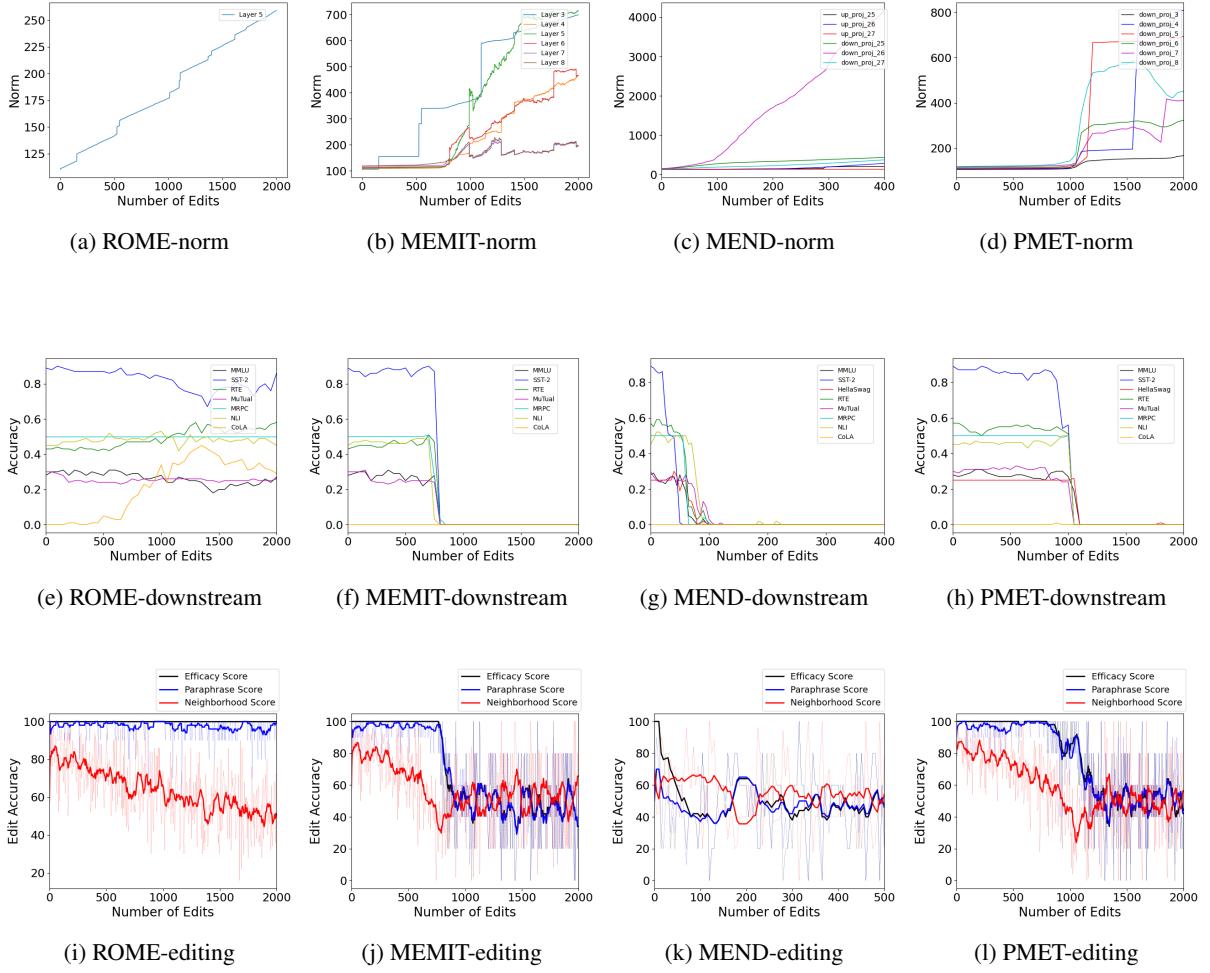


Figure 15: Norm growth during knowledge editing on GPT-J.

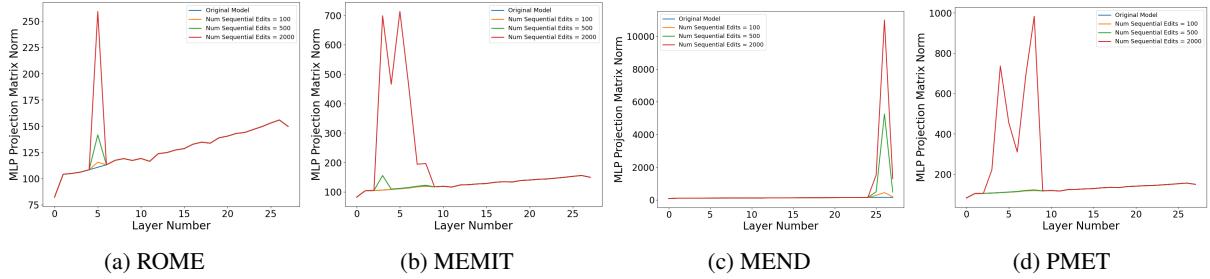


Figure 16: Norm growth for edits 100, 500, 2000 for GPT-J.

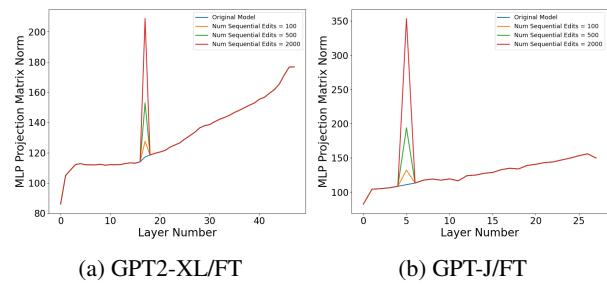


Figure 17: Norm growth for edits 100, 500, 2000 for FT editing.

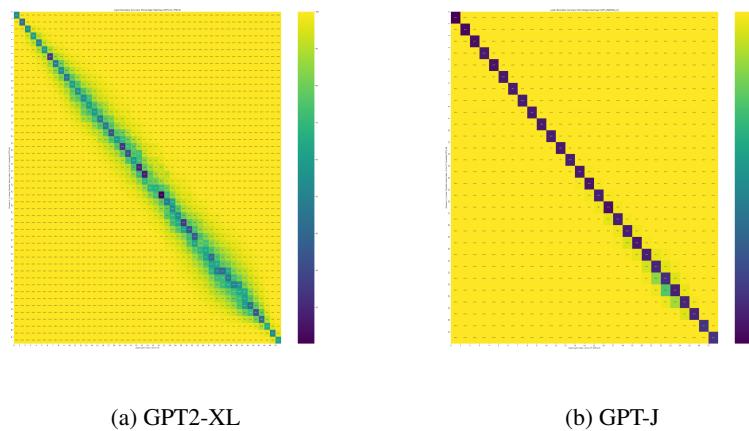


Figure 18: Boundary plots for unedited GPT2-XL and GPT-J.

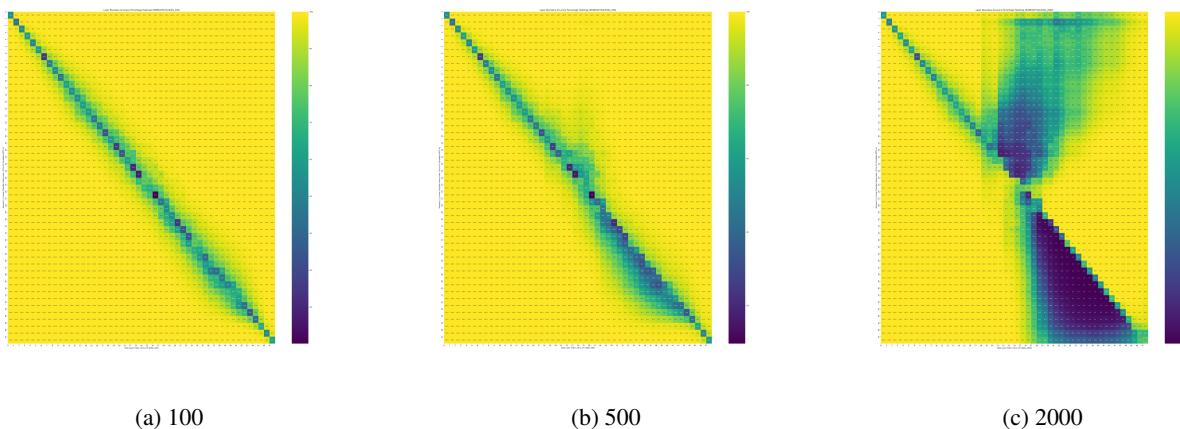


Figure 19: Boundary plots for edits 100, 500, 2000 for GPT2-XL/ROME.

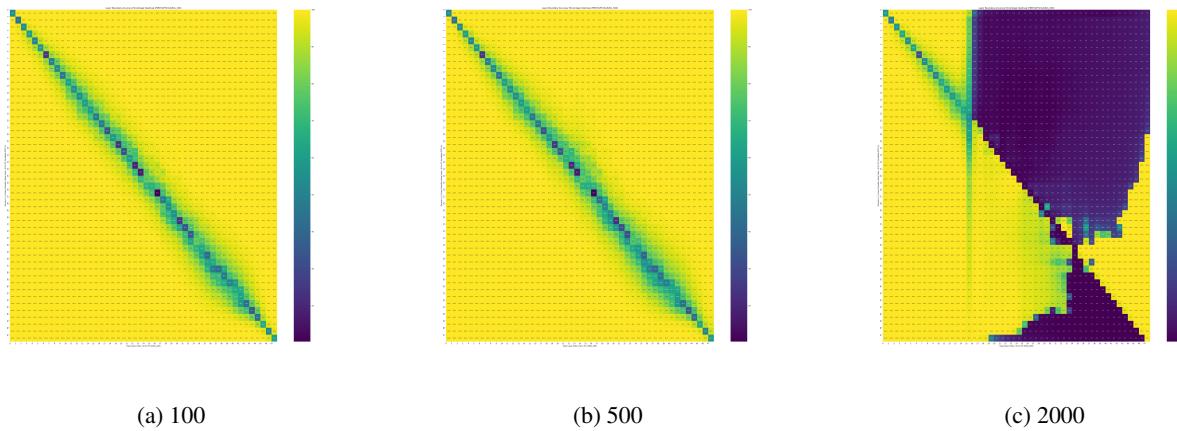


Figure 20: Boundary plots for edits 100, 500, 2000 for GPT2-XL/PMET.

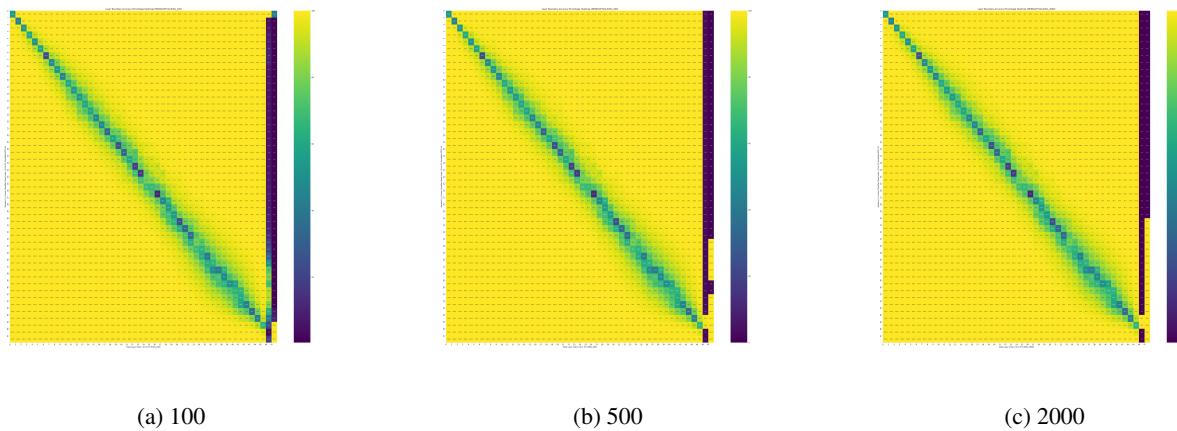


Figure 21: Boundary plots for edits 100, 500, 2000 for GPT2-XL/MEND.

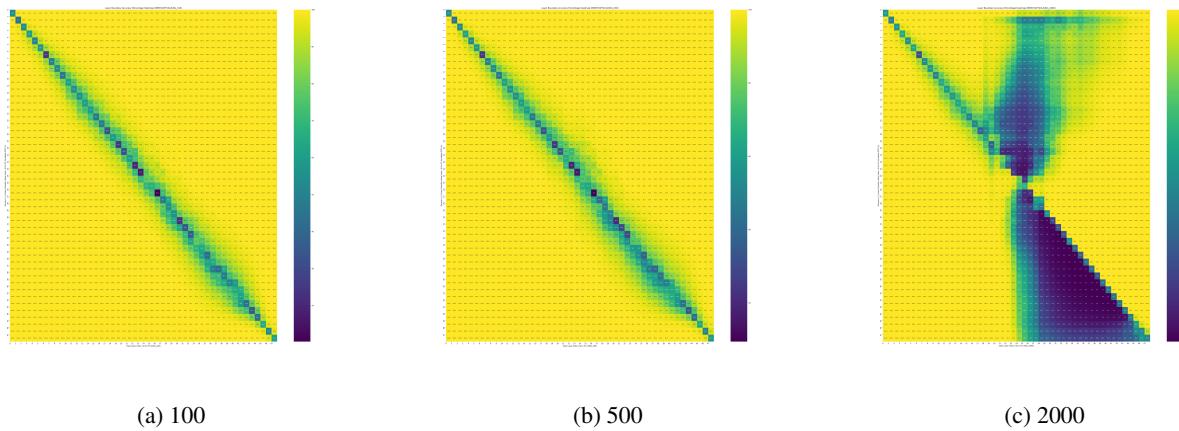


Figure 22: Boundary plots for edits 100, 500, 2000 for GPT2-XL/MEMIT.

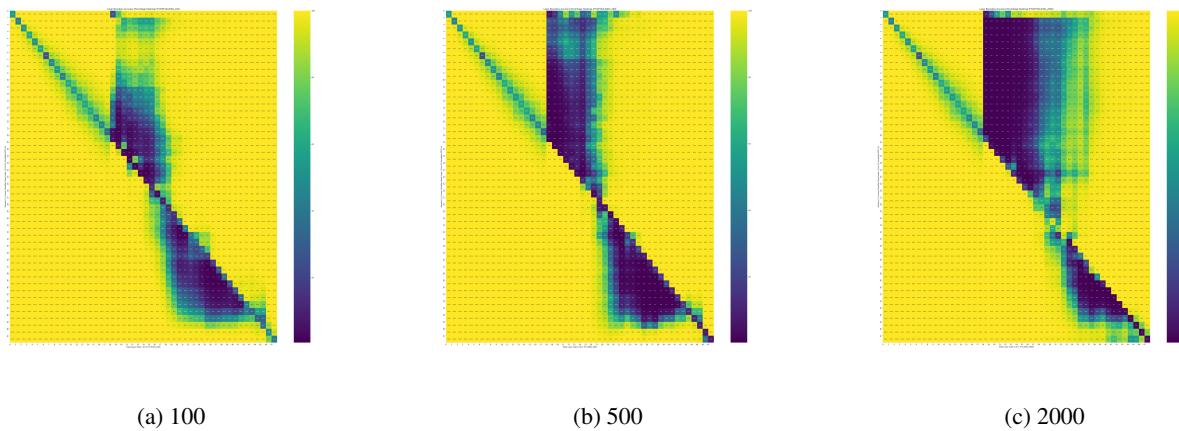


Figure 23: Boundary plots for edits 100, 500, 2000 for GPT2-XL/FT.

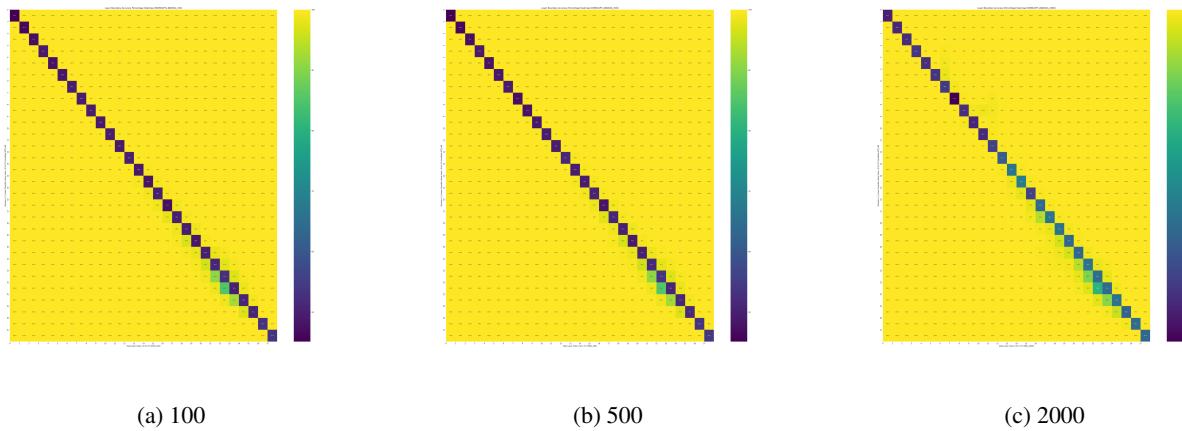


Figure 24: Boundary plots for edits 100, 500, 2000 for GPT-J/ROME.

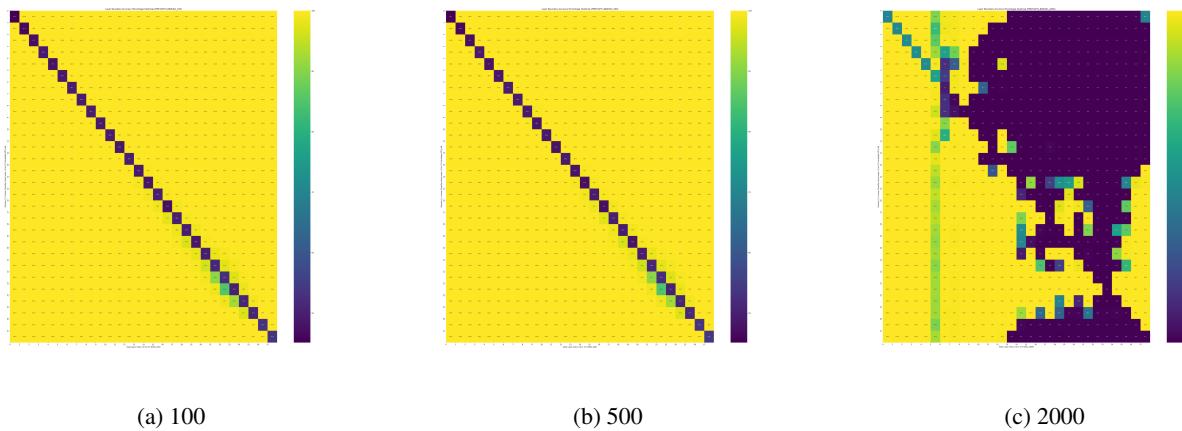


Figure 25: Boundary plots for edits 100, 500, 2000 for GPT-J/PMET.

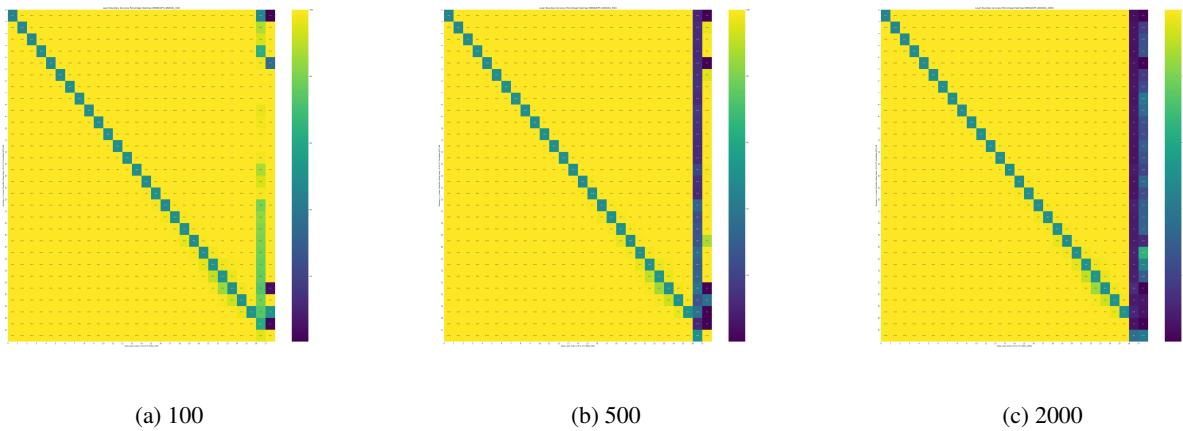


Figure 26: Boundary plots for edits 100, 500, 2000 for GPT-J/MEND.

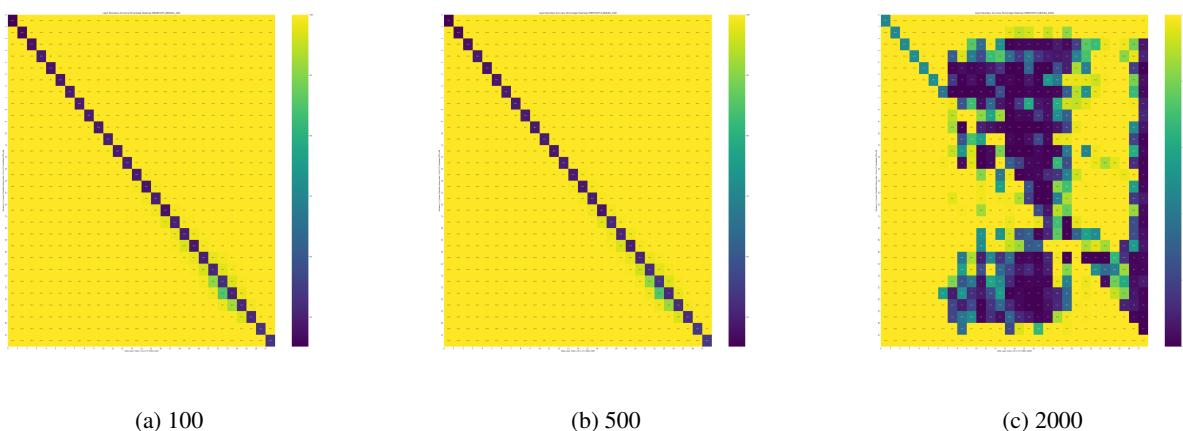
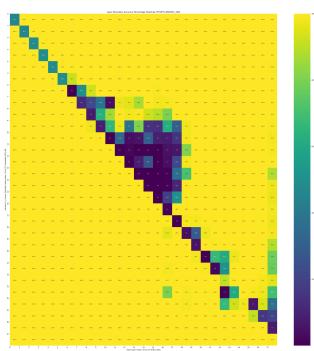
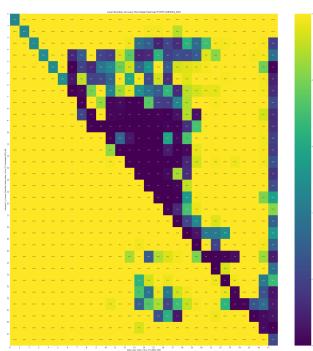


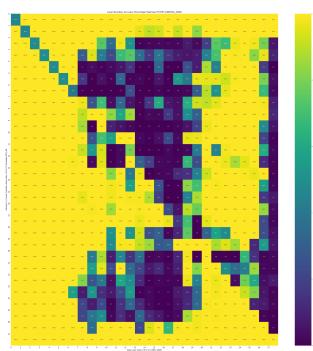
Figure 27: Boundary plots for edits 100, 500, 2000 for GPT-J/MEMIT.



(a) 100



(b) 500



(c) 2000

Figure 28: Boundary plots for edits 100, 500, 2000 for GPT-J/FT.

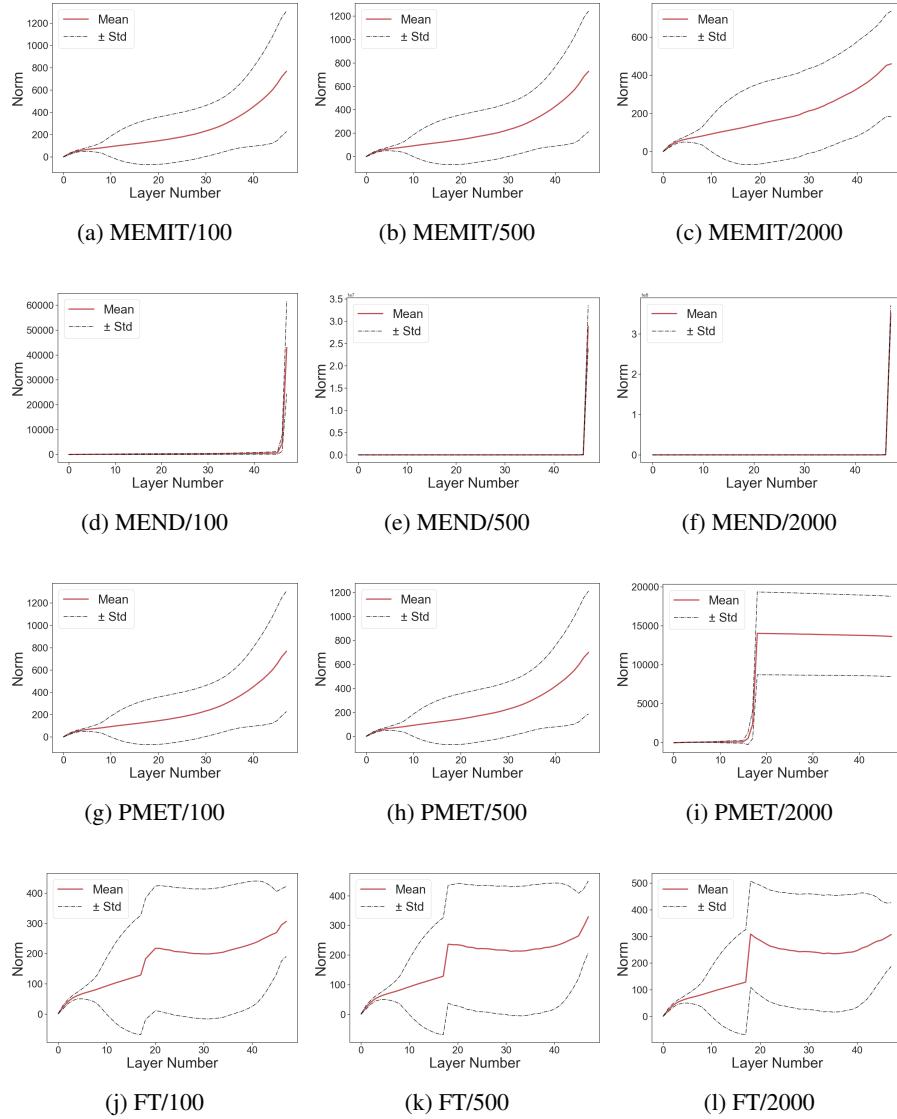


Figure 29: Activation Norms at different layers for edits 100, 500, 2000 for all editing methods for GPT2-XL.