An Edge of Stability For Language Model Post-Training

Angel Patricio, MIT

Term Project for 6.7960

The **"Edge of Stability" (EoS)** phenomenon is a recently observed (Cohen et al., 2022) pattern in the optimization of deep neural nets, wherein the realized training trajectory of the model gravitates toward a narrow band on the edge between stability and catastrophic divergence. The following blog post investigates whether this phenomenon, originally established for pre-training of relatively simple MLP and CNN-style networks, persists in a new setting: *language model post-training*. We in particular study the behavior of training on an instruction tuning target, a key step in alignment which is greatly under-explored in the context of EoS. We give a background, introduce relevant literature, and explore the training dynamics on an established instruction tuning dataset, tracking relevant metrics to visualize the learning trajectory.

Background: Learning at the Edge of Stability

EoS refers in particular to a pattern in the *sharpness* of the loss landscape throughout training. The *sharpness* of the loss landscape at a given training step is defined as the largest Hessian eigenvalue for Hessian

$$\mathcal{H} = \nabla_\theta^2 \mathcal{L}(\theta).$$

The first major observation in the EoS literature is that this sharpness quantity $\lambda_{\max}$ begins relatively low and steadily increases throughout early training. This phase has been given the name of **"progressive sharpening" (PS)** (Cohen et al., 2022). In particular, PS continues until the sharpness has reached a threshold value $\lambda_{\max} \approx 2/\eta$, for current learning rate η. Once this threshold has been reached, the learning dynamics enter the **Edge of Stability (EoS)** phase.

During the EoS phase, behavior is characterized by a few key observations:

- The sharpness $\lambda_{\max}$ fluctuates around the threshold value $\lambda_{\max} \approx 2/\eta$.
- The loss continues to decrease non-monotonically.
- The gradient norm $\left\|\nabla_\theta \mathcal{L}(\theta)\right\|$ also fluctuates around a constant value.

These three behaviors largely disagree with established optimization theory. Under a quadratic approximation to our loss function, classical optimization theory tells us that gradient descent is no longer stable and will begin to oscillate chaotically and diverge. That is, when the sharpness reaches the threshold value $\lambda_{\max} \approx 2/\eta$, gradient descent enters a positive feedback loop inducing exploding gradient magnitudes and model divergence.

Nonetheless, the EoS phenomenon is observed in practice and in some sense implies that there is something fundamentally broken between the assumptions of classical optimization theory and the true nature of these large, nonlinear, nonconvex, stochastic systems. Deep networks seem to *self-organize* at

the EoS, and somehow this behavior leads to (empirically) improved generalization in the resulting models (Cohen et al., 2022).

EoS and Model Architecture

EoS has been observed in a variety of deep learning architectures including standard both feedforward and CNN-style networks in the original paper (Cohen et al., 2022). Beyond the conclusions of the original work, later work (Wang et al., 2025) investigates the phenomenon by studying sharpness dynamics in transformer training on a block-by-block basis (i.e., averaging sharpness across all layers for a given type of block). Notably, authors observe that the sharpness dynamics are not uniform across the block and in fact each block type exhibits its own unique sharpness threshold and local EoS. The call this the **sharpness disparity principle** and leverage this principle to motivate *blockwise learning rates*, accelerating LLM training. This suggests that EoS is not a niche phenomenon but instead an intrinsic emergent property under the learning dynamics of deep neural networks and in particular transformer-style networks as well.

EoS and Optimizers

Another heavily explored axis in the EoS literature is the role of the *optimizer*. In particular, (Andreyev & Beneventano, 2025) revisits the EoS idea in the context of optimizing with SGD, since (Cohen et al., 2022) focuses only on the batch GD case and considers only the *full-batch* (spectral) Hessian sharpness. Authors study the **batch sharpness** given by:

$$\text{Batch Sharpness}(\theta) = \frac{\mathbb{E}_{B\sim\mathcal{P}_B}[\nabla L_B(\theta)^\top \mathcal{H}(L_{\mathbf{B}})\nabla L_B(\theta)]}{\mathbb{E}_{B\sim\mathcal{P}_b}[\|\nabla L_{\mathbf{B}}(\theta)\|^2]}$$

And track this metric as a function of training progress. The result is that this batch sharpness quantity reaches a similar $ 2 / $ threshold as the full-batch sharpness in (Cohen et al., 2022) while the batch sharpness under SGD settles at a lower value than the $ 2 / $ threshold. Authors argue that especially for smaller batch sizes, the **Edge of (Stochastic) Stability (EoSS)** phenomenon is largely governed by this batch sharpness quantity instead of the full-batch sharpness.

Similar work analyzes the emergence of EoS in the context of *adaptive gradient methods* (Cohen et al., 2024). In this work the authors analyze the EoS dynamics when using adaptive gradient methods such as Adam and rmsprop, confirming typical EoS dynamics when operating with “frozen” Adam (preconditioned, no dynamic updates to the preconditioner). Interestingly, authors discover a new regime of training where the sharpness of the *preconditioned* Hessian $P^{-1}\mathcal{H}$ (for given preconditioning matrix P) governs what they call the **Adaptive Edge of Stability (AEoS)**. During this regime, the *raw* sharpness quantity $\lambda_{\max}$ *continues to rise*, even beyond the original $2/\eta$ threshold. It is this preconditioned sharpness quantity that reaches a threshold

$$\lambda_1(P^{-1}\mathcal{H}) \approx \frac{(2+2\beta_1)}{(1-\beta_1)\eta}$$

(for relevant Adam hyperparameter β_1) and oscillates throughout the AEoS phase.

Post-Training and Relevance to EoS

The idea of *post-training* was largely introduced in (Wei et al., 2022) for the context of instruction tuning. The authors of the work take a language model *pre-trained* on a large, general text corpus for text understanding and subject it to a second leg of training, this time on a new dataset with a new goal, i.e. instruction following. Generally, this *task shift* in tandem with the shift in the *data distribution* corresponds with a *sharper* loss landscape than the original pre-training phase and in a practical sense it motivates the use of much smaller learning rates for post-training compared to pre-training.

Post-training offers a particularly unique loss landscape which is grounded in practice and has strong implications for the EoS phenomenon thanks to biases towards better generalization performance (Damian et al., 2023). Despite this, the EoS phenomenon has not been studied in the context of post-training.

Motivation, Hypothesis, and Experiment Design

The main motivation behind this work is the glaring gap in the literature regarding the analysis of an additional axis: the *data distribution* itself. Presented literature covers the behavior of EoS in a variety of settings, but none of them explore the impact of a *distribuition* or even *task* shift on the EoS phenomenon.

To that end, this project narrows in particularly on the task of *language model post-training* (LMPT), where the goal is to improve the generalization performance of a pre-trained language model on a downstream task by fine-tuning it on a small amount of task-specific data. In particular, this work studies whether we observe an EoS phenomenon in the instruction tuning stage of language model alignment, experimenting with the Alpaca dataset (Taori et al., 2023).

Hypotheses and Experiment Design

At a high level, we hypothesize that, if EoS is truly an instrinsic phenomenon of deep learning, we should *continue* to observe similar behavior patterns for LMPT. The impact of a distribution shift could potentially be significant enough to entirely alter the learning dynamics, but we expect the EoS phenomenon to persist.

To isolate the fine-tuning dynamics, we borrow the openly available **Pythia** (Biderman et al., 2023) suite of pre-trained language models. These models largely follow the proven architecutre recipe of GPT-3 (Brown et al., 2020). Importantly, these models have *not* seen any form of instruction or otherwise fine-tuning updates, and are trained exclusively on the Pile (Gao et al., 2020) (Biderman et al., 2022). These models were chosen for the purpose of providing a clean baseline language model for this study of EoS dynamics.

We chose to fine-tune the models using the Alpaca dataset (Taori et al., 2023), which is a dataset of 52002 instruction-following examples, some including “input” fields for chat-style interaction. This dataset is established in the literature as a standard dataset for instruction tuning and alignment of language models.

To broaden the range of post-training situations we also repeat experiments using **Low-Rank Adaptation (LoRA)** (Hu et al., 2021) for parameter-efficient fine-tuning (PEFT). This method reduces the amount of trainable parameters for the fine-tuning run by learning only a low-rank decomposition of the full parameter matrix. We include this method to study the impact of practical changes to the fine-tuning formula on the EoS phenomenon.

Training runs were performed using rented compute resources from the rental platform Vast.ai, which was used to rent a single 4090 instance for experiments over the course of about 40-60 hours of compute time. We optimize using SGD while sweeping the learning rate from small (used in practice, $\sim 1e^{-4}$) to large (practically inapplicable, $\sim 1e^{-2}$) values. There is a fixed momentum of 0.9 used for all runs as well as a fixed L2 weight decay of $1e^{-4}$.

Due to memory constraints, we were unable to perform batch GD experiments to study the impact of full-batch GD on the EoS phenomenon in LMPT. Therefore, we only report the batch sharpness metric from (Andreyev & Beneventano, 2025) for comparison, in place of the spectral sharpness metric from the original work (Cohen et al., 2022) (note: we *did* calculate this for short experiments, but the results followed the rest of the quantites and the others were much easier to compute at larger scale). We use a fixed effective batch size of 16 for all runs, achieved with 4 steps of accumulation with batch size 4.

Experiments were conducted for 50000 training steps on a 5000-sample subsample of the Alpaca dataset (Taori et al., 2023), and we report the batch sharpness metric from (Andreyev & Beneventano, 2025) for comparison as well as the product $\lambda_{\max} \cdot \eta$ for clear comparison to the expected threshold value of 2.

Due to some surprising initial results, we ran an additional baseline using an MLP to replicate established results from (Cohen et al., 2022).

Results

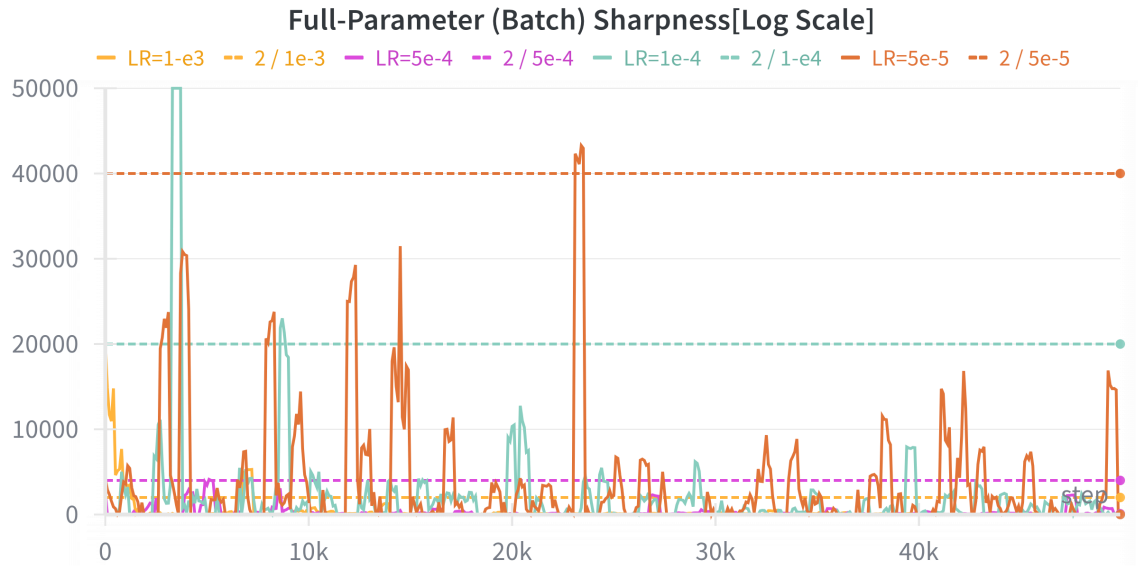
Note: On The Sharpness Calculation

A problem commonly faced by work in EoS is the computatoinal cost of the sharpness calculation. The original work (Cohen et al., 2022) gets around this issue using Hessian-vector products to avoid realizing the full Hessian matrix, a strategy that we largely borrow in this work. Since we are also working with quite low batch sizes (again due to memory constraints), we use this same strategy to calculate the *batch sharpness* metric from (Andreyev & Beneventano, 2025).

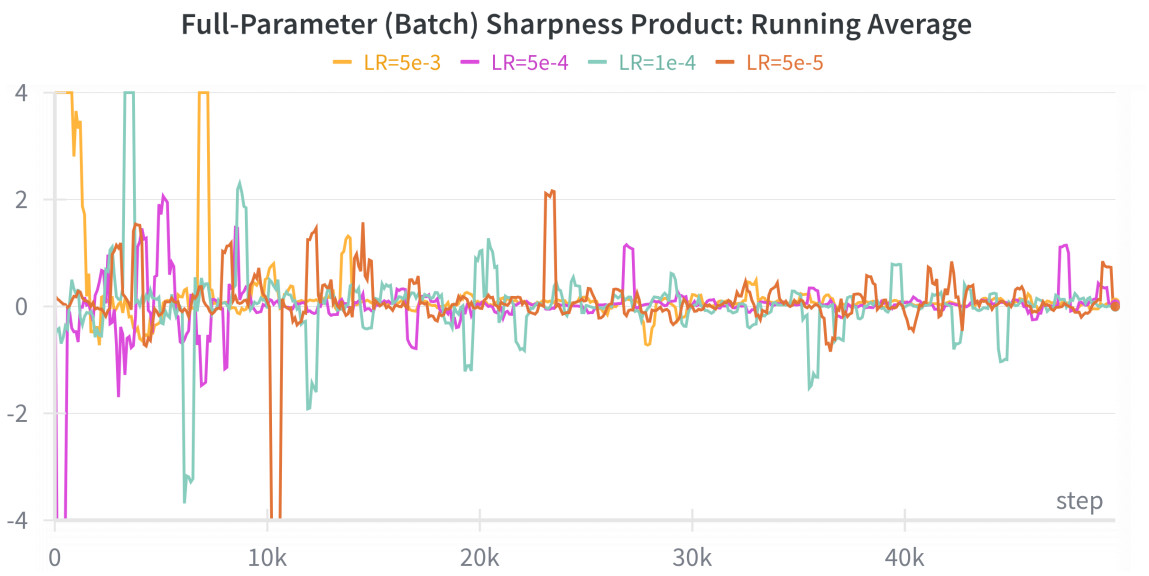
The sharpness calculation is performed by taking a Hessian-vector product and calling a standard implementation of Lanczos iteration to compute the largest eigenvalue of the Hessian.

Experimental Results

Experimental results are presented below. Notably, we observe a clear *lack* of EoS emergence in the post-training phase of language models. That is, it appears that the batch gradient signal applicable and used in much of the EoS literature is overrun by noise in the post-training phase.



LLM Full Sharpness



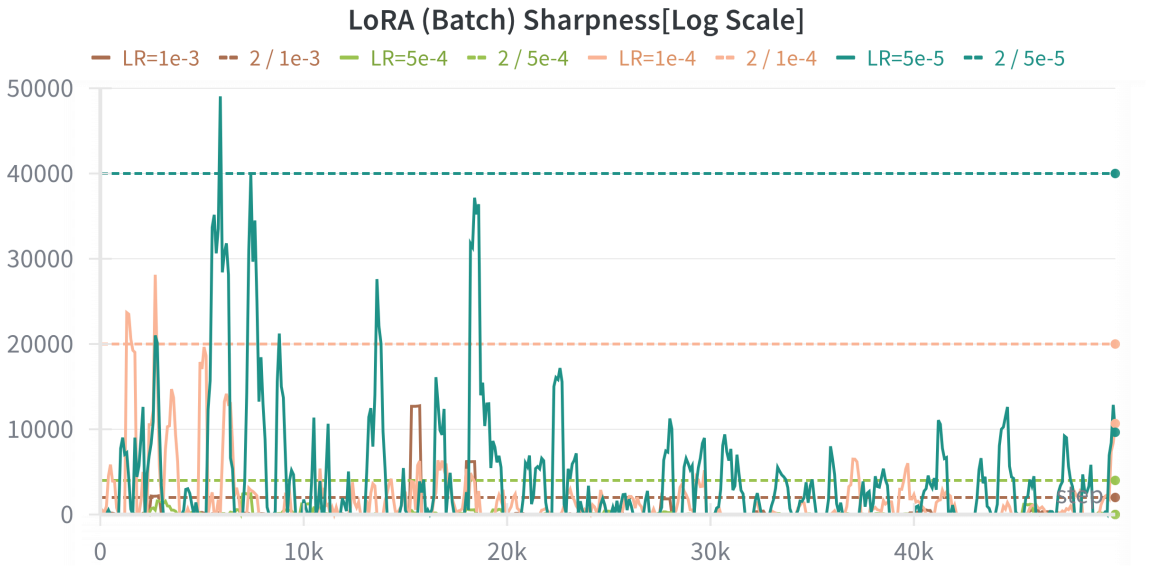
LLM Full Product

The above pair of figures show the raw batch sharpness quantity (solid lines) plotted against the target threshold (dashed lines) in the first figure as well as the product $\lambda_{\max} \cdot \eta$ in the second figure. Our characterization of "EoS" in this case *should* look like the quantities in the first figure rising to the dashed lines, and the values in the second figure rising to 2.

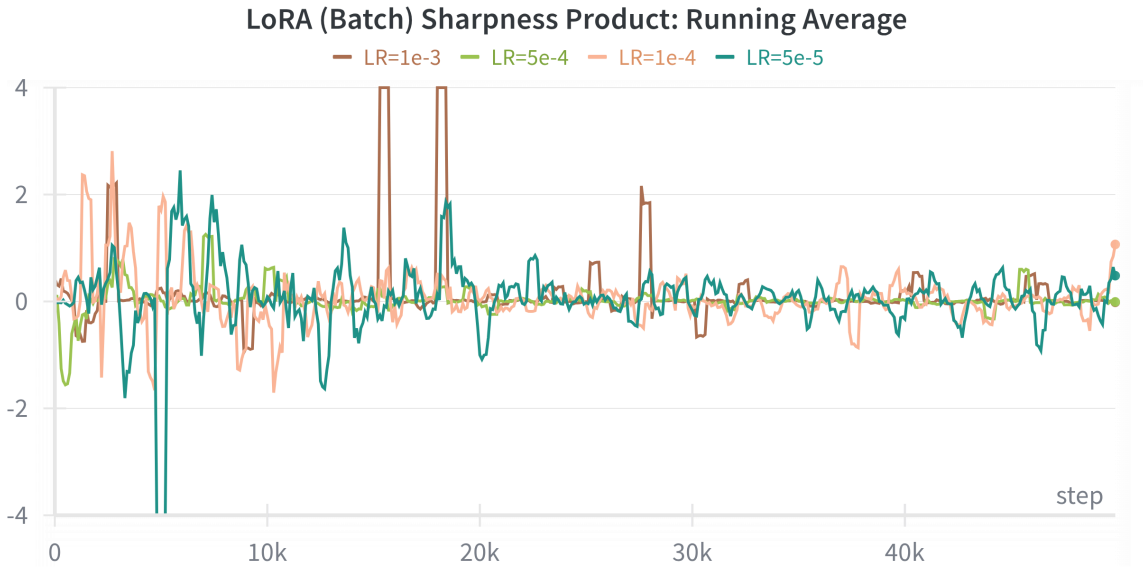
Notably, we observe behavior which unmistakably does *not* follow this pattern. In fact, the sharpness fluctuates by several orders of magnitude (including reaching negative values several times), and the product $\lambda_{\max} \cdot \eta$ never stabilizes at 2.

This result goes against the hypotheses as we are seeing a complete lack of EoS emergence in the post-training phase of language models, beyond just a muted one.

The below pair of figures shows these same quantities, but tracked for the experiments using a set of LoRA adapters of rank 8 targeting all modules.



LLM LoRA Sharpness

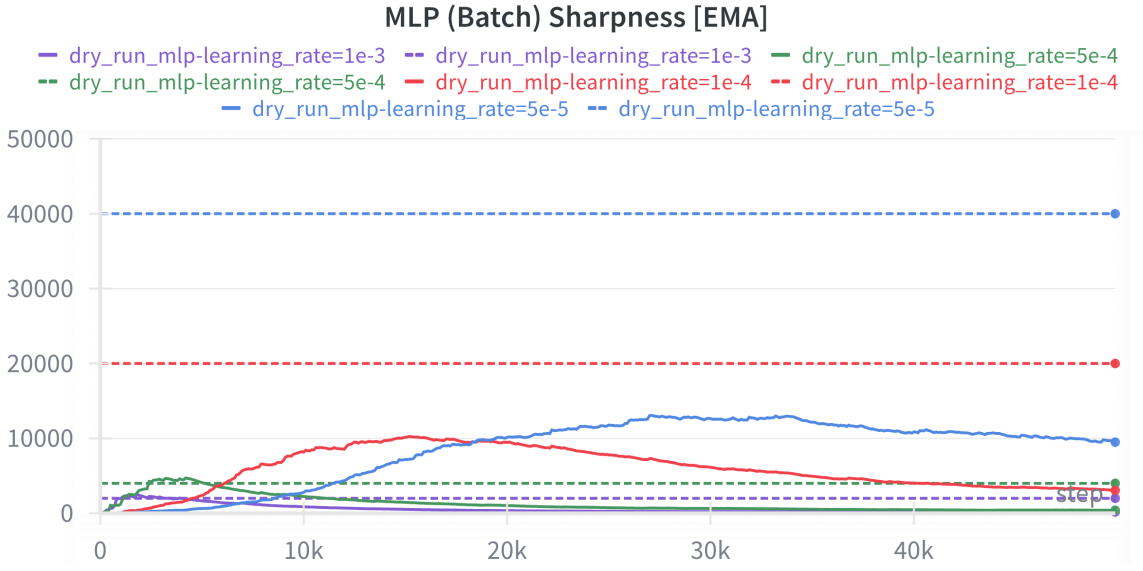


LLM LoRA Product

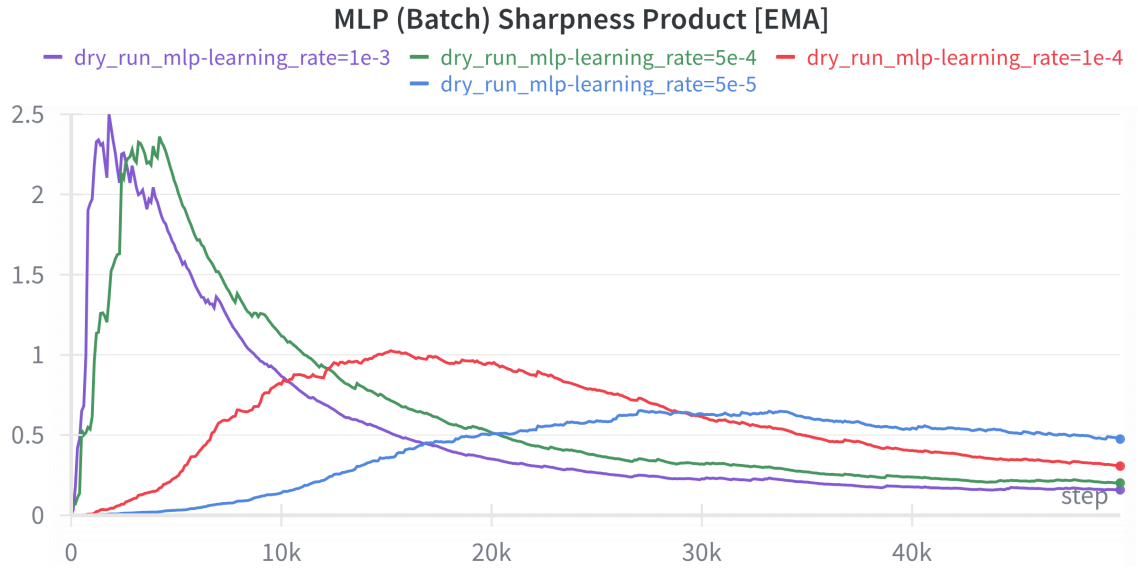
It is once again visibly clear that the EoS phenomenon does not emerge in the post-training phase of language models, even when using a parameter-efficient method of fine-tuning. This implies that there is some fundmental difference between the fine-tuning phase and the post-training phase of language models, at least at the level of the batch gradient signal.

The same target behavior was expected for the LoRA experiments as was expected for the full model experiments, but the same behavior was not observed. The plots of these two quantities seem to be mostly dominated by random noise, contradicting the notion of the batch sharpness as a reliable metric for characterizing the EoS phenomenon in the context of large language models.

Due to the surprising nature of these initial results, we ran an additional baseline using an MLP to replicate established results from (Cohen et al., 2022). Notably, we see a clear pattern of progressive sharpening, with the product $\lambda_{\max} \cdot \eta$ clearly rising as training progresses at first. However, this quantity *also* does not reach the target value of 2, which could be due to numerical instability (see Conclusion). Due to compute and time constraints, *all experiments* were ran to 50,000 iterations; however, while the MLP experiments show clear signs of progressive sharpening, the true EoS phase is not reached in the fixed step budget. Graphs are below.

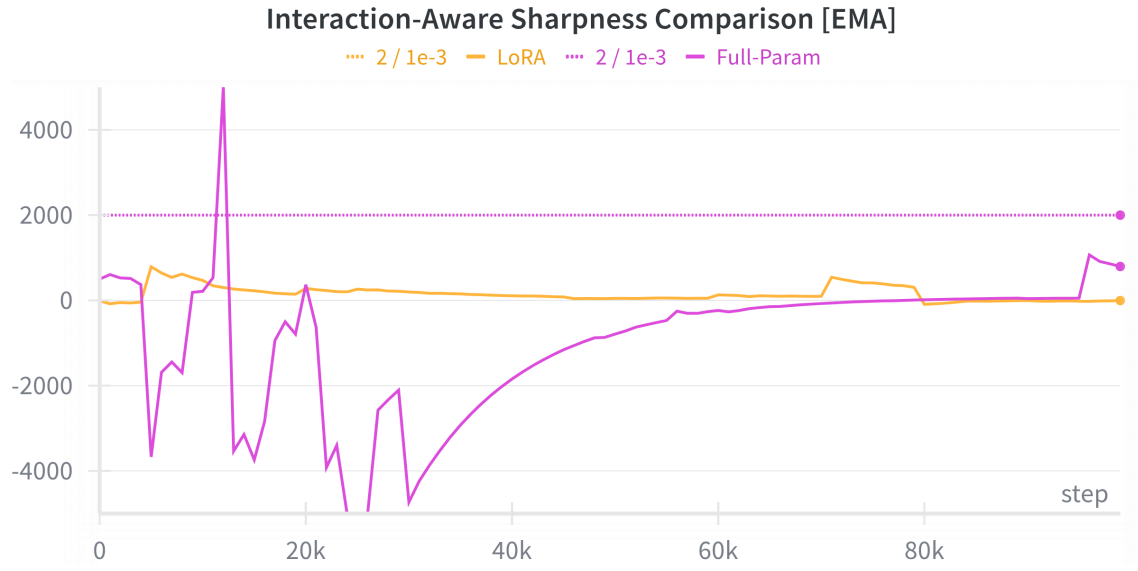


MLP Batch Sharpness

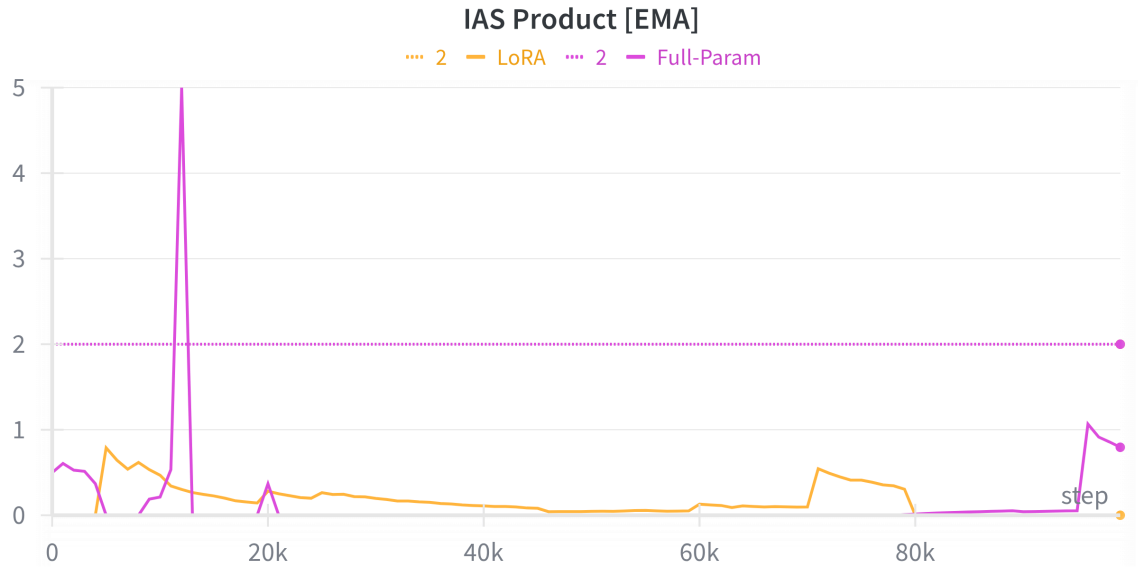


MLP Product

Comparing the visuals of the MLP experiments to the language model experiments, it is clear that the batch sharpness metric is not a reliable metric for characterizing the EoS phenomenon in the context of large language models. Nonetheless, there remained one more metric to try: the **interaction-aware sharpness** as defined in (Lee & Jang, 2023). This metric seeks to remove the impact of batch size interactions with the gradient landscape by using monte carlo to get a stronger estimate of true sharpness. We reran experiments for a final time, tracking the interaction-aware sharpness metric for both the full model and LoRA experiments. Graphs are below.



Interaction-Aware Sharpness



Interaction-Aware Product

As you can see, this *most robust* quantity *still* fails to reveal signs of EoS. In fact, even with this robust metric we still see large negative spikes in the global sharpness landscape, suggesting further still that our hypothesis was incorrect. With all this in mind, we can make some conclusions about an EoS phenomena for post-training.

Conclusion

Overall, the results of this experiment suggest that the EoS phenomenon does **not occur** in the context of LMPT. However, we acknowledge limitations and hesitate to make such a strong claim without further (planned) experimentation and analysis. In particular, we acknowledge that the calculation of metrics for this project, simply due to the nature of these models being multi-millions of parameters large, is heavily susceptible to numerical instability and noise. We hope to reproduce these experiments at a larger scale to rule out the possibility of numerical instability.

We additionally acknowledge that the fact none of our existing metrics behaved how we might expect *doesn't necessarily* mean that there *exists* no robust metric that will, unifying these sub-phenomena into a single, unified EoS phenomenon. Future work should explore this possibility.

Finally, we acknowledge that we fail to analyze these dynamics in the *batch GD* context of the original work. Although these results heavily suggest a no-EoS post training phase, future work should explore the dyanamics of batch GD on this problem.

To conclude, we must *continue* to look for the signs of edge of stability in LMPT, as these results point to the problem setting bringing along with it some new dynamics that we have yet to fully understand.

References

Andreyev, A., & Beneventano, P. (2025). *Edge of stochastic stability: Revisiting the edge of stability for SGD*. https://arxiv.org/abs/2412.20553

Biderman, S., Bicheno, K., & Gao, L. (2022). *Datasheet for the pile*. https://arxiv.org/abs/2201.07311

Biderman, S., Schoelkopf, H., Anthony, Q., Bradley, H., O'Brien, K., Hallahan, E., Khan, M. A., Purohit, S., Prashanth, U. S., Raff, E., Skowron, A., Sutawika, L., & Wal, O. van der. (2023). *Pythia: A suite for analyzing large language models across training and scaling*. https://arxiv.org/abs/2304.01373

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R.,

Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). *Language models are few-shot learners*. https://arxiv.org/abs/2005.14165

Cohen, J. M., Ghorbani, B., Krishnan, S., Agarwal, N., Medapati, S., Badura, M., Suo, D., Cardoze, D., Nado, Z., Dahl, G. E., & Gilmer, J. (2024). *Adaptive gradient methods at the edge of stability*. https://arxiv.org/abs/2207.14484

Cohen, J. M., Kaur, S., Li, Y., Kolter, J. Z., & Talwalkar, A. (2022). *Gradient descent on neural networks typically occurs at the edge of stability*. https://arxiv.org/abs/2103.00065

Damian, A., Nichani, E., & Lee, J. D. (2023). *Self-stabilization: The implicit bias of gradient descent at the edge of stability*. https://arxiv.org/abs/2209.15594

Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., Foster, C., Phang, J., He, H., Thite, A., Nabeshima, N., Presser, S., & Leahy, C. (2020). *The pile: An 800GB dataset of diverse text for language modeling*. https://arxiv.org/abs/2101.00027

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2021). *LoRA: Low-rank adaptation of large language models*. https://arxiv.org/abs/2106.09685

Lee, S., & Jang, C. (2023). A new characterization of the edge of stability based on a sharpness measure aware of batch gradient distribution. *The Eleventh International Conference on Learning Representations*. https://openreview.net/forum?id=bH-kCY6LdKg

Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., Liang, P., & Hashimoto, T. (2023). *Alpaca: A Strong, Replicable Instruction-Following Model*. Stanford Center for Research on Foundation Models (CRFM). https://crfm.stanford.edu/2023/03/13/alpaca.html

Wang, J., Wang, M., Zhou, Z., Yan, J., E, W., & Wu, L. (2025). *The sharpness disparity principle in transformers for accelerating language model pre-training*. https://arxiv.org/abs/2502.19002

Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., Du, N., Dai, A. M., & Le, Q. V. (2022). *Finetuned language models are zero-shot learners*. https://arxiv.org/abs/2109.01652