Alignment-Constrained Dynamic Pruning for LLMs: Identifying and Preserving Alignment-Critical Circuits

Anonymous Author(s)
Affiliation
Address
email

Abstract

Large Language Models require substantial computational resources for inference, posing deployment challenges. While dynamic pruning offers superior efficiency over static methods through adaptive circuit selection, it exacerbates alignment degradation by retaining only input-dependent safety-critical circuit preservation across diverse inputs. As a result, addressing these heightened alignment vulnerabilities remains critical. We introduce Alignment-Aware Probe Pruning (AAPP), a dynamic structured pruning method that adaptively preserves alignment-relevant circuits during inference, building upon Probe Pruning. Experiments on LLaMA 2-7B, Qwen2.5-14B-Instruct, and Gemma-3-12B-IT show AAPP improves refusal rates by 50% at matched compute, enabling efficient yet safety-preserving LLM deployment.

Introduction

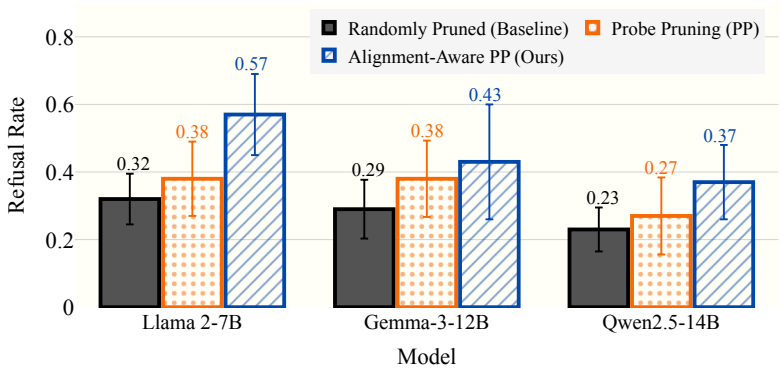


Figure 1: Refusal rates of LLaMA-2-7B, Qwen-2.5-14B, and Gemma-3-12B models on the WildJailbreak dataset [Jiang et al., 2024] under pruning ratio $r = 0.3$. We compare our Alignment-Aware Probe Pruning (AAPP) against two baselines: Probe Pruning (PP) [Le et al., 2025] and random pruning. Across all three models, AAPP consistently achieves higher refusal rates, demonstrating that preserving alignment-critical circuits upon the detection of adversarial prompts improves safety behavior under pruning.

LLMs deliver impressive capabilities yet impose high computational costs, with inference costs scaling directly with model size [Kaplan et al., 2020]. Pruning offers a promising route to reduce these costs [Han et al., 2016], using different techniques, including static structured pruning [Ma et al., 2023] as well as dynamic probe-guided pruning (PP) [Le et al., 2025] which improves the

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

accuracy-efficiency frontier by pruning columns of the learnable linear transformation that maps intermediate hidden state to the output hidden state, referred to as an input channel. However, these methods risk pruning alignment-critical structures, potentially weakening safety guardrails and degrading behaviors such as refusal of harmful instructions. Recent analyses [Wei et al., 2024] show that removing as little as 3% of parameters is enough to compromise safety. This brittleness motivates the development of Alignment-Aware Probe Pruning (AAPP)—a method that explicitly preserves alignment-critical circuits.

AAPP uses the average activation value for each input channel. By comparing these scores obtained from benign and harmful prompts to the scores obtained from our probe pass, our method detects adversarial inputs and enforces hard exclusions on alignment-critical structures. This structured pruning approach yields an improved efficiency-alignment frontier: AAPP outperforms PP, having refusal rates up to 50% greater for the same computational budget. These findings suggest constraint-satisfying pruning as a practical route to efficient yet safe LLMs. Our key contributions are as follows:

- We develop a pruning framework that preserves interpretable circuits
- We evaluate our framework on refusal rate, toxicity, accuracy, and computational cost (FLOPs)

Related Work

Structured Pruning

Structured pruning is a key approach for reducing the computational cost of LLMs. LLM-Pruner [Ma et al., 2023] removes entire attention heads and MLP neurons via gradient-based importance, while Wanda [Sun et al., 2024] prunes weights with small magnitude and activation values post-hoc, achieving high sparsity without retraining. Probe Pruning [Le et al., 2025] extends this line by using probed hidden states to guide batch-wise pruning, improving the accuracy-efficiency frontier. However, these methods risk pruning the preservation of alignment-critical structures.

Alignment Preservation

Several methods aim to preserve alignment during model modification. Safe LoRA [Hsu et al., 2024] and SaLoRA [Li et al., 2025] constrain LoRA updates to remain within safety-aligned subspaces, while LoRI [Zhang et al., 2025] and LoTA [Panda et al., 2024] apply structural sparsity to reduce catastrophic forgetting. These works show that constraining fine-tuning helps preserve desirable behaviors in LLMs. NLSR [Yi et al., 2025] restores safety by transplanting safety-critical neurons from an aligned reference model. These approaches show that explicit parameter constraints and neuron transplantation can maintain refusal, honesty, and toxicity safeguards even under structural changes. Layer-level analyses further support targeted preservation: Shi et al. [Shi et al., 2024] showing that alignment changes concentrate in late-stage layers and that compression can focus on non-critical regions.

Methods

As shown in Figure 2, Alignment-Aware Probe Pruning consists of five stages, namely probe generation; probing, recording activations; comparison to our historical activation scores; history-informed pruning; and inference.

Activations and Scoring

For each target with C input channels, we create 3 tensors: general, benign, and harmful using sets of prompts: (1) general prompts to maintain linguistic functionality from C4 dataset [Raffel et al., 2020]; (2) benign prompts from wild adversarial dataset; and (3) harmful prompts from wild adversarial dataset. ([Jiang et al., 2024]). Each set of scores stores the squared ℓ_2 norm of channel activations compressed across the batch and sequence dimensions. We refer to this value as the "channel's energy".

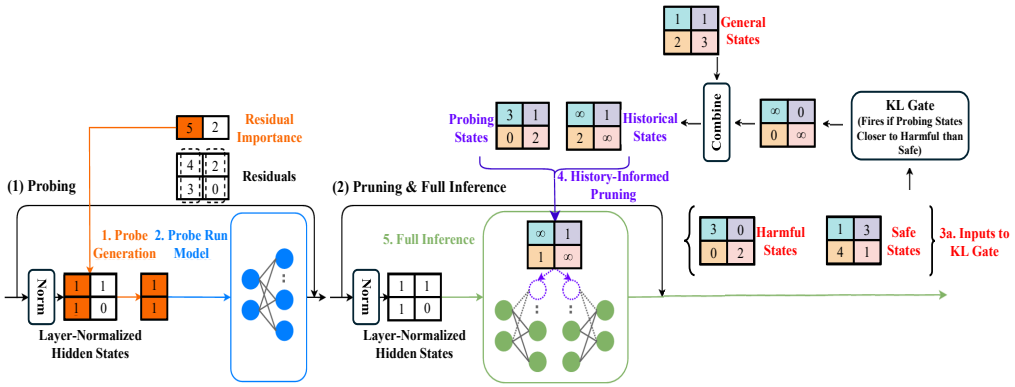


Figure 2: Alignment-Aware Probe Pruning (PP) is executed in five stages: (1) From the layer-normalized hidden states, pick tokens based on residual-importance and build a small probe. (2) Run the probe a few layers ahead to produce probing states (3a) A KL Gate compares them to historical states from safe and harmful prompts and fires when closer to harmful, ensuring the preservation of alignment-critical structures. If the gate does not fire, the probe states are just fused with the general historical states (4) Using the integrated states to calculate the pruning metric [Le et al., 2025], prune low-score channels. (5) Perform full inference on the remaining weights.

For structured pruning, we adopt the PP_{sp} importance metric from Probe Pruning [Le et al., 2025], which computes per-channel pruning scores using the ℓ_2 norms of each input channel's activations. Here, W^{final} denotes the learnable linear transformation between hidden states, and X^{int} the intermediate hidden state. A lower PP_{sp} score, I_k, indicates less important channels.

$$I_k = \left\| \left\{ |W^{\text{final}}_{i,k}|^2 \cdot \|X^{\text{int}}_{:,:,k}\|_2^2 \right\}_{i=0}^{C_{\text{out}}} \right\|_2, \tag{1}$$

Finally, we blend live scores with stored activation scores obtained from the set of general prompts.

Risk-aware gate and channel selection

We keep $k = \lceil (1-r)C \rceil$ channels, reserving $k_{align} = \lfloor \text{align\_frac} \cdot C \rfloor$ channels for safety. Probing states; and historical states from benign and harmful prompts are normalized into distributions: 'p'; and 'q_{safe}, and q_{jail}', respectively, using Equation 2.

$$KL_{\text{harm}} = \sum_c p_c \log \frac{p_c}{q^c_{\text{jail}}}, \quad KL_{\text{safe}} = \sum_c p_c \log \frac{p_c}{q^c_{\text{safe}}}. \tag{2}$$

If $KL_{\text{harm}} - KL_{\text{safe}} \geq \tau_{\text{margin}}$, we preserve the top k_{align} channels by $hist_{\text{jail}}$ as we wish to protect channels most active under harmful prompts because they include refusal circuitry. We then fill the remainder by descending score. Otherwise, we retain the top k channels by score. Using these scores, binary masks are generated for pruning and then materialized to obtain real compute reductions.

Experimentation and Results

We evaluate on HuggingFace implementations of Llama-2-7B-chat, Qwen2.5-14B-Instruct, and Gemma-3-12B-IT, using prompts from the WildJailbreak dataset ([Jiang et al., 2024]) which were not used for the generation of historical states. Workloads contain prompts of avg. length 300 tokens with 120 tokens generated. Unless stated otherwise, we fix hyperparameters to align frac $= 0.3$, refresh window $= 20$, and batch size $= 20$ for prompts.

We estimate inference FLOPs calculated using 2 FLOPs/MAC ([Hoffmann et al., 2022]) taking into account the number of layers, attention heads, hidden size, intermediate size, and vocabulary size for the given model. We prune only in the input channels of attention o_{proj} and MLP $down_{\text{proj}}$, excluding the first 6 and last 3 layers. Outputs are post-hoc labeled for refusal and toxicity. Metrics

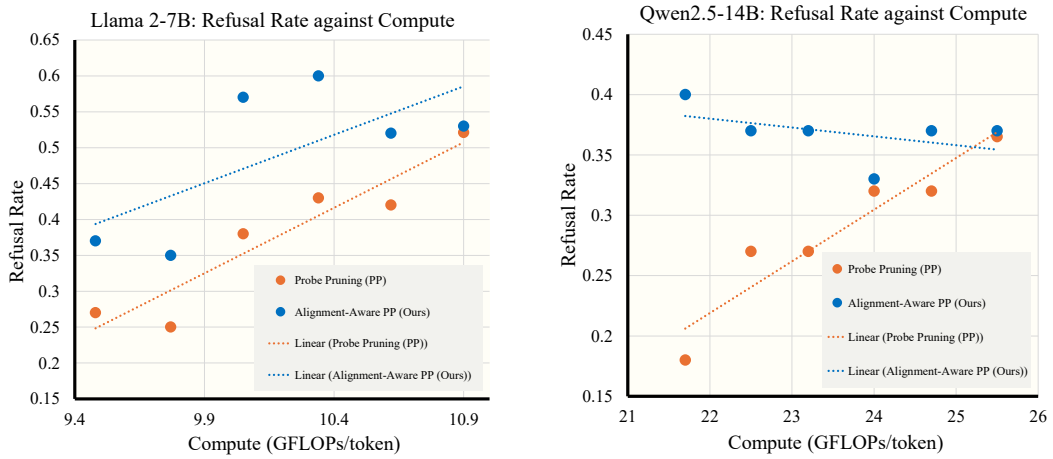
include throughput compute (FLOPs/token), refusal rate (trained classifier), classification accuracy and toxicity (Perspective API [Lees et al., 2022]).

Across the two methods (AAPP and PP), We first consider the model's ability to classify harmful and unharmful prompts and act accordingly. This is investigated across various compute budgets and prune ratios. Following this, we assess the safety of the model's responses for AAPP and PP using toxicity as the measure.

Refusal Rates at Fixed Prune Ratio

Figure 1 presents refusal rates at prune ratio $r = 0.3$. Across all three models, AAPP achieves higher refusal rates (implicit and explicit) than both Randomly Pruned and Probe Pruning (PP) baselines, preserving alignment behavior. On Llama-2-7B-chat, AAPP attains a refusal rate (0.57) 50% and 78% greater than PP (0.38) and Random Pruning (0.32), respectively. Similar improvements hold for Llama-2-7B-chat (37% and 61%) and Gemma-3-12B-IT (13% and 48%), confirming the robustness of our approach across architectures.

Refusal Rates against Compute (FLOPs per Token)



(a) Llama-2-7B-chat. AAPP maintains substantially higher refusal rates at comparable compute budgets, achieving safer behavior with fewer FLOPs compared to standard PP.

(b) Qwen2.5-14B-Instruct. AAPP preserves refusal performance as compute decreases, improving the refusal-compute trade-off relative to PP across the efficiency spectrum.

Figure 3: Refusal rate vs compute (GFLOPs/token) across models. AAPP consistently achieves higher refusal rates at lower compute costs than standard PP, demonstrating improved alignment–efficiency trade-offs.

Extending the investigation, we vary compute budgets to look into the alignment-efficiency frontiers created using either method. Figure 3a and 3b illustrates alignment (refusal rate) as a function of computational efficiency (GFLOPs/token) for the Llama-2-7B-chat and Qwen2.5-14B-Instruct models, respectively, under Probe Pruning (PP) and Alignment-aware PP. Given the same computational budget, our method achieves a higher refusal rate, shifting the efficiency-alignment frontier upward. For example, on Llama-2-7B-chat (3a), to achieve a target refusal rate of 0.5, our method requires only 10.3 GFLOPs/token, compared to a higher cost with PP. Qwen2.5-14B-Instruct exhibits the same pattern, demonstrating that AAPP maintains safety more efficiently across various compute levels. These results show that AAPP improves the alignment-efficiency trade-off, achieving safer behavior while reducing inference cost, and generalizing across diverse model families.

Alignment Accuracy

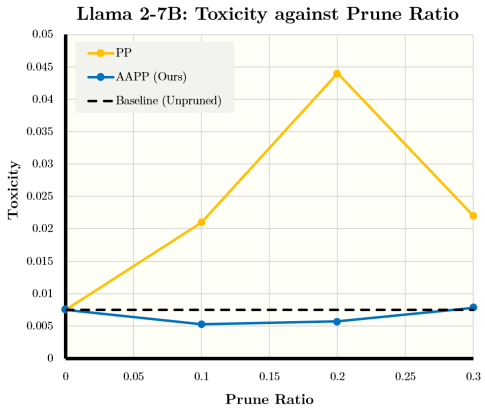
The accuracy of these refusals and the behavior of the model, more generally, is shown in Table 1. It indicates that AAPP outperforms PP across prune ratios on Llama-2-7B-Chat and Qwen2.5-

Model	Prune Ratio	Method	F1 (↑)	Accuracy (↑)	FAR (↓)
Llama-2-7B-chat	0	PP	1.000	1.000	0.000
		AAPP	1.000	1.000	0.000
	0.15	PP	0.725	0.702	0.290
		AAPP	0.834	0.808	0.201
	0.3	PP	0.645	0.624	0.313
		AAPP	0.760	0.741	0.254
Qwen2.5-14B-Instruct	0	PP	1.000	1.000	0.000
		AAPP	1.000	1.000	0.000
	0.15	PP	0.876	0.891	0.058
		AAPP	0.880	0.916	0.05
	0.3	PP	0.730	0.820	0.169
		AAPP	0.786	0.858	0.092

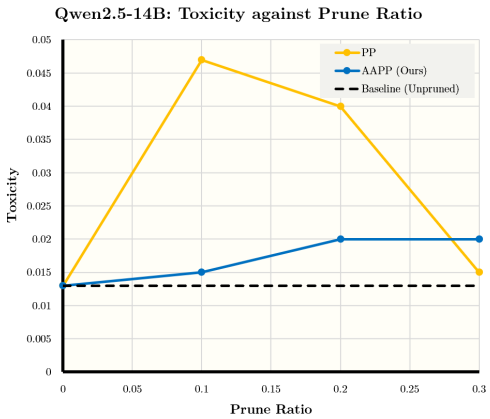
Table 1: Comparison of F1, Accuracy and FAR for PP and AAPP across prune ratios on Llama-2-7B-Chat and Qwen2.5-14B-Instruct: AAPP has a lower False Acceptance Rate with higher classification accuracy, behaving more similarly to the unpruned models.

14B-Instruct. The results for the pruned models are compared to the unpruned model, which we consider to have a maximum for these metrics, as our pruned models cannot exceed the performance of the base model. We use F1 to balance recall and precision, accuracy and False Acceptance Rate to indicate how often the model does not refuse prompts. PP's accuracy and F1 decline as pruning increases, dropping to 0.575 and 0.585 at a 0.3 ratio for Llama2-7B-Chat. In contrast, AAPP retains higher values, 0.741 accuracy and 0.760 F1, indicating stronger classification stability. Additionally, AAPP maintains a lower False Acceptance Rate (FAR) (e.g. 0.216 vs 0.353 at 0.3). Similar results can be seen for Qwen2.5-14B-Instruct. Overall, these results demonstrate AAPP's ability to preserve safety and behavior near to the unpruned models at reduced compute.

Toxicity against Prune Ratio



(a) Llama-2-7B-chat. AAPP maintains toxicity levels closer to the unpruned baseline compared to PP, demonstrating better preservation of safety alignment under aggressive pruning.



(b) Qwen2.5-14B-Instruct. AAPP sustains lower toxicity scores closer to the unpruned model across pruning ratios, outperforming PP in safety preservation.

Figure 4: Toxicity vs prune ratio across models. AAPP consistently preserves lower toxicity and safer outputs under pruning, outperforming PP across both Llama-2-7B-chat and Qwen2.5-14B-Instruct.

Through toxicity, we can understand how safely the model responds. Figure 4a and 4b indicates that across both models, AAPP shows clear safety gains over PP. On Llama-2-7B-Chat, PP's toxicity peaks at 0.044 at a 0.2 prune ratio, while AAPP stays nearly constant near 0.0075, matching the unpruned baseline. Similarly, on Qwen2.5-14B-Instruct, PP reaches 0.08, but AAPP remains below

0.02. This demonstrates that AAPP preserves alignment even under heavy pruning. Although toxicity scores decrease at high pruning ratios, this may reflect linguistic degradation rather than improved safety. Pruning can suppress expressive activations, yielding flatter, less coherent text that is rated as less toxic.

Conclusion

We propose a pruning method that preserves alignment while reducing inference cost. By integrating a risk-aware gate with probe-guided pruning, we prevent the removal of alignment-critical structures upon the input of an adversarial prompt and improves the efficiency-alignment frontier. Experiments on Llama-2-7B-chat, Qwen2.5-14B-Instruct, and Gemma-3-12B-IT show that AAPP sustains lower toxicity and greater classification accuracy at lower FLOP budgets, offering a practical route to safer and more efficient LLMs.

Therefore, our method improves efficiency, scalability, and energy use without significantly compromising safety. However, there is a risk of missed unsafe inputs as the model is pruned, but we reduce the chance of this happening through conservative gating.

Limitations of our study include evaluation at mid-scale model sizes and approximate FLOP accounting. Future work will extend AAPP to larger models and investigate whether similar additions can be made to build upon probe pruning in other contexts.

References

Song Han, Huizi Mao, and William J. Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *arXiv preprint arXiv:1510.00149*, 2016. URL https://arxiv.org/abs/1510.00149.

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, Ethan Dyer, Geoffrey Irving, Jack W. Rae, George van den Driessche, Bart de Haas, Peter Battaglia, Mateusz Malinowski, Arthur Guy, Simon Osindero, Koray Kavukcuoglu, Roman Ring, Adam Cain, Chloe Hillier, Rewon Winter, Oliver Hutter, Timothy Lillicrap, Simon Green, Albin Cassirer, Chris Jones, Valentina Cherepanova, Adam Rutherford, Felix Mensch, Nicholas Crampton, Sam Manning, Sjoerd van Steenkiste, and Laurent Sifre. Training compute-optimal large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. URL https://arxiv.org/abs/2203.15556.

Chia-Yi Hsu, Yu-Lin Tsai, Chih-Hsun Lin, Pin-Yu Chen, Chia-Mu Yu, and Chun-Ying Huang. Splora: The silver lining of reducing safety risks when fine-tuning large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. URL https://openreview.net/pdf?id=HcifdQZFZV.

Liwei Jiang, Kavel Rao, Seungju Han, Allyson Ettinger, Faeze Brahman, Sachin Kumar, Niloofar Mireshghallah, Ximing Lu, Maarten Sap, Yejin Choi, and Nouha Dziri. Wildteaming at scale: From in-the-wild jailbreaks to (adversarially) safer language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/54024fca0cef9911be36319e622cde38-Paper-Conference.pdf.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020. URL https://arxiv.org/abs/2001.08361.

Qi Le, Enmao Diao, Ziyan Wang, Xinran Wang, Jie Ding, Li Yang, and Ali Anwar. Probe pruning for efficient large language models. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://arxiv.org/abs/2502.15618.

Alyssa Lees, Vinh Q. Tran, Yi Tay, Jeffrey Sorensen, Jai Gupta, Donald Metzler, and Lucy Vasserman. A new generation of perspective api: Efficient multilingual character-level transformers. *arXiv preprint arXiv:2202.11176*, 2022. URL https://arxiv.org/abs/2202.11176.

Mingjie Li, Wai Man Si, Michael Backes, Yang Zhang, and Yisen Wang. Salora: Safety-alignment preserved low-rank adaptation. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://arxiv.org/abs/2501.01765.

Xianjun Ma, Guangji Fang, and Xiaojie Wang. Llm-pruner: On the structural pruning of large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. URL https://arxiv.org/abs/2305.11627.

Ashwinee Panda, Berivan Isik, Xiangyu Qi, Sanmi Kojejo, Tsachy Weissman, and Prateek Mittal. Lottery ticket adaptation: Mitigating destructive interference in llms. *arXiv preprint arXiv:2406.16797*, 2024. URL https://arxiv.org/pdf/2406.16797v2.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL https://jmlr.org/papers/v21/20-074.html.

Guangyuan Shi, Zexin Lu, Xiaoyu Dong, Wenlong Zhang, Xuanyu Zhang, Yujie Feng, and Xiao-Ming Wu. Understanding layer significance in llm alignment. *arXiv preprint arXiv:2410.17875*, 2024. URL https://arxiv.org/abs/2410.17875.

Mingjie Sun, Zhiqing Liu, Adam Bair, and J. Zico Kolter. Wanda: A simple and effective pruning approach for large language models. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2306.11695.

Boyi Wei, Kaixuan Huang, Yangsibo Huang, Tinghao Xie, Xiangyu Qi, Mengzhou Xia, Prateek Mittal, Mengdi Wang, and Peter Henderson. Assessing the brittleness of safety alignment via pruning and low-rank modifications. In *ICLR Workshop on Understanding of Foundation Models (ME-FoMo)*, 2024. URL https://openreview.net/pdf?id=niBPvgJIHB.

Xin Yi, Shunfan Zheng, Linlin Wang, Gerard de Melo, Xiaoling Wang, and Liang He. Nlsr: Neuron-level safety realignment of large language models against harmful fine-tuning. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2025. URL https://arxiv.org/abs/2412.12497.

Juzheng Zhang, Jiacheng You, Ashwinee Panda, and Tom Goldstein. Lori: Reducing cross-task interference in multi-task low-rank adaptation. *arXiv preprint arXiv:2504.07123*, 2025. URL https://arxiv.org/pdf/2504.07448v1.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The paper's contributions include improving the trade-off between alignment behaviour and efficiency which is accurately stated in the abstract and introduction.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations of our study are stated in the conclusion. For example, we acknowledge that evaluation included mid-scale models and would ideally include smaller and larger model sizes.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.

- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA] .

   Justification: This paper does not include theoretical results.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: In the experimentation and results section, the paper discloses the information needed to reproduce the experimental results, including datasets used and testing parameters.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example

(a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
(b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The datasets for testing are publicly available and their uses are clearly stated in the paper. The code is avaliable at https://anonymous.4open.science/r/Alignment-Aware-Probe-Pruning-D53E/.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Test details are included in the experimentation and results section.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Error bars represent 95% Wilson confidence intervals, which are suitable for proportion-based metrics and provide reliable uncertainty estimates. These are used in our bar chart figure.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: All experiments were run using PyTorch 2.3 on a H100 SXM 80 GB GPU, and Python 3.10.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: The paper conforms with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.

- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: These are stated in the conclusion.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA] .

    Justification: The paper does not pose these risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We build upon probe pruning and clearly state so, improving the method's alignment-efficiency trade-off. The creators of this method and others are clearly credited.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: This work contributes our alignment-aware probe pruning framework, with supporting information explained in the paper.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: [NA] .

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA] .

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: The paper evaluates the proposed pruning method on multiple large language models (e.g., Llama-2-7B-chat, Qwen2.5-14B-Instruct).

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.