CS2881 Mini Project

Assessing the Brittleness of Safety Alignment via Pruning and Low-Rank Modifications

Hugh Van Deventer, Anastasia Ahani, Terry Zhou

1 Reproduction of Main Figure

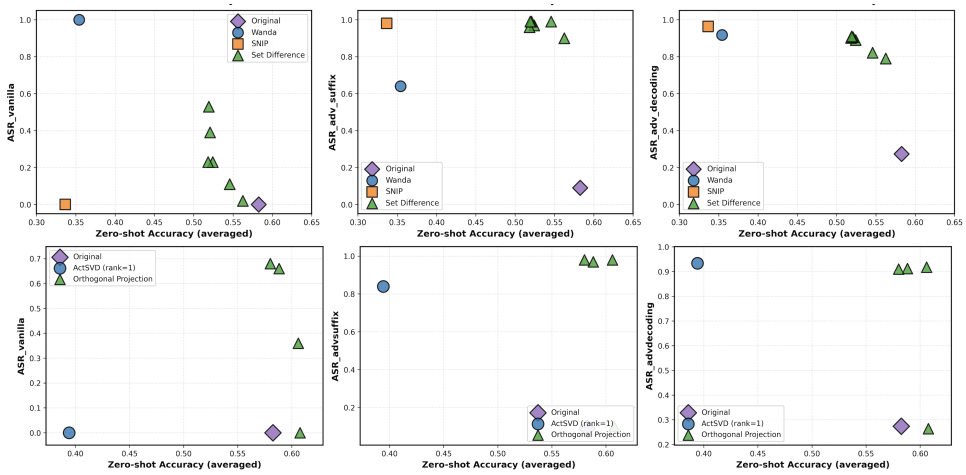

Figure 1: Reproduction of Figure 2a (top) and Figure 2b (bottom) from the original paper. ASR and accuracy after removing safety-critical regions in Llama-2-7B-chat. In the top figure, we observe a score of 0 for the SNIP method, and on the bottom, we observe a score of 0 for the ActSVD method for A$ASR_{vanilla}$–these results are inconsistent with the expectation and the original figure.

We ran into many reproducibility issues while trying to rerun the code and recreate the figure. There were several issues with the original code, including the instantiation of the vllm object, loading the tokenizer and the model from different paths, and other CUDA-related issues. However, most significantly, the safety evaluations consistently outputted incorrect results. For example, when running a safety evaluation on the model after having pruned the top 1% safety-critical neurons (whether this was calculated with SNIP or Wanda), we would expect an attack success of 100% (thus the output should be 1). However, we were consistently getting an output of 0 (specifically for $ASR_{vanilla}$) across all the different score-calculation methods (e.g. Wanda, SNIP, and ActSVD). After much investigation, we hypothesized that this was due to some corruption of the model during pruning, since the perplexity of the model would vary greatly when it was outputting a correct versus incorrect result. After running the set difference code, we noticed in face that the outputs seemed to be in line with the expected values, and the reason the corruption was not occurring here was because of a difference in methodology for carrying out the pruning. For set difference, we calculate the scores across all neurons, save them, and then load the pre-computed scores instead of computing them in real-time before pruning. Thus we tried rewriting all of the main pruning code in order to follow this methodology. This seemed to actually fix the issue for Wanda with $ASR_{vanilla}$, but not for SNIP (pruning ranks follows a different methodology, but we could not explain the 0 in value for attack success in $ASR_{vanilla}$ either).

Ultimately, we were unable to explain or fully get to the bottom of the inconsistencies in the results, especially why the $ASR_{vanilla}$ dataset results were the most significantly impacted and error-prone. Additionally, it seemed that these errors did not extend to the utility datasets, since the zero-shot accuracy results seemed to match closely with those from the original paper.

2 Extension: How do safety-critical neurons change after freezing and fine-tuning?

We have previously observed first-hand the consequences of fine-tuning (as in with our Homework 0), and section 4.4 of the paper also details how freezing safety-critical neurons during fine-tuning does not in fact help prevent these attacks. The authors conclude that this observation aligns with the hypothesis that fine-tuning attacks create alternative pathways in the original model. In our extension, we set out to find evidence supporting this hypothesis via two experiments:

1. **Safety neuron change**: After finetuning with safety-critical neurons frozen, how different are the safety-critical neurons between original and finetuned models?
2. **Weight change**: Do safety critical neurons change less compared to baseline when we finetune over all weights?

2.1 Experiment 1: Safety-Critical Neurons Analysis

2.1.1 Methodology

Objective: We want to investigate whether finetuning with frozen safety-critical neurons lead to the model finding alternative path of generating harmful contents. This can be done by comparing the two sets of safety-critical neurons before and after finetuning. If an alternative path is indeed created, we should expect to find great differences between the two sets of safety-critical neurons, as the original neurons will have diminished significance in generation safety-related contents.

Hypothesis: After finetuning with frozen safety-critical neurons, the new set of safe-critical neurons for the finetuned models should be very different from the original neurons.

Experimental Setup: Given the original `meta-llama/Llama-2-7b-chat-hf` model, we identify the top 1% safety neurons and utility neurons respectively using the SNIP score, as well as the safety-critical neurons using thresholds $p = 0.05$ and $q = 0.05$ (i.e. the neurons that are in the top 5% of safety neurons but are not in the top 5% of utility neurons). We then freeze these safety-critical neurons and perform LoRA finetuning on the model using the following configuration:

- LoRA parameters: rank $r = 8$, scaling $\alpha = 16$
- Training: 1 epoch, batch size 4, gradient accumulation steps 4
- Learning rate: 1×10^{-4}, maximum sequence length: 512 tokens

After finetuning, we follow the same process to identify a new set of top safety, top utility, and safety-critical neurons.

2.1.2 Results

Table 1 shows the overlaps between neurons in safety-critical, top safety, and top utility groups, as measured by both overlap percentage and Jaccard index. We can clearly see that there is a small overlap between the original and new safety-critical neurons. This indicates that the original safety-critical neurons are no longer significant in the generation of safety-related responses, suggesting that finetuning has indeed found an alternative path that bypasses the frozen safety-critical neurons. The slightly larger overlap in top safety neurons compared to safety-critical neurons is expected, as some of these top safety neurons would also be responsible for other key utilities, hence are better preserved.

In comparison, the top utility neurons see a smaller change before and after finetuning. This suggests that, different from safety neurons, utility neurons largely remain stable and are preserved during the finetuning process. The Jaccard index for top utility neurons (33.93%) is more than twice that of safety-critical neurons (15.38%), and substantially higher than top safety neurons (24.29%). This asymmetry reveals a fundamental difference in how finetuning affects these two types of behaviors: while the model maintains a relatively consistent set of utility-critical neurons, it dramatically reorganizes its safety mechanisms, despite the safety-critical neurons remain frozen during finetuning.

Table 1: Overlap analysis between original and finetuned model neurons

Neuron Type	Overlap Percentage	Jaccard Index
Safety-Critical	26.01%	0.15
Top Safety	39.09%	0.24
Top Utility	50.67%	0.34

2.2 Experiment 2: Weight Drift Analysis

2.2.1 Methodology

Objective: We investigate whether safety-critical neurons exhibit greater weight drift compared to other neurons during standard fine-tuning, testing two hypotheses:

- **Hypothesis A (Fragile Safety)**: Safety-critical neurons experience higher-than-average weight changes, indicating inherent fragility.
- **Hypothesis B (Alternate Pathways)**: Safety-critical neurons remain relatively stable while new harmful circuits emerge elsewhere in the network.

Experimental Setup: We identify safety-critical neurons using thresholds $p = 0.05$ and $q = 0.002$, selecting the top 0.2% of safety neurons that are not among the top 5% of utility neurons. For comparison, we also track the top 0.1% of safety neurons, top 0.1% of utility neurons, and randomly sampled 0.1% of neurons. We then perform LoRA fine-tuning on Llama-2-7b-chat-hf using the Alpaca dataset with the following configuration:

- LoRA parameters: rank $r = 8$, scaling $\alpha = 16$
- Training: 1 epoch, batch size 4, gradient accumulation steps 4
- Learning rate: 1×10^{-4}, maximum sequence length: 512 tokens

Data Collection We record initial weights for all tracked neurons (safety-critical, utility, and random). During fine-tuning, we save checkpoints every 500 training steps and compute:

1. Absolute weight difference: $|\mathbf{w}_t - \mathbf{w}_0|$
2. Relative weight change: $\frac{|\mathbf{w}_t - \mathbf{w}_0|}{|\mathbf{w}_0|}$

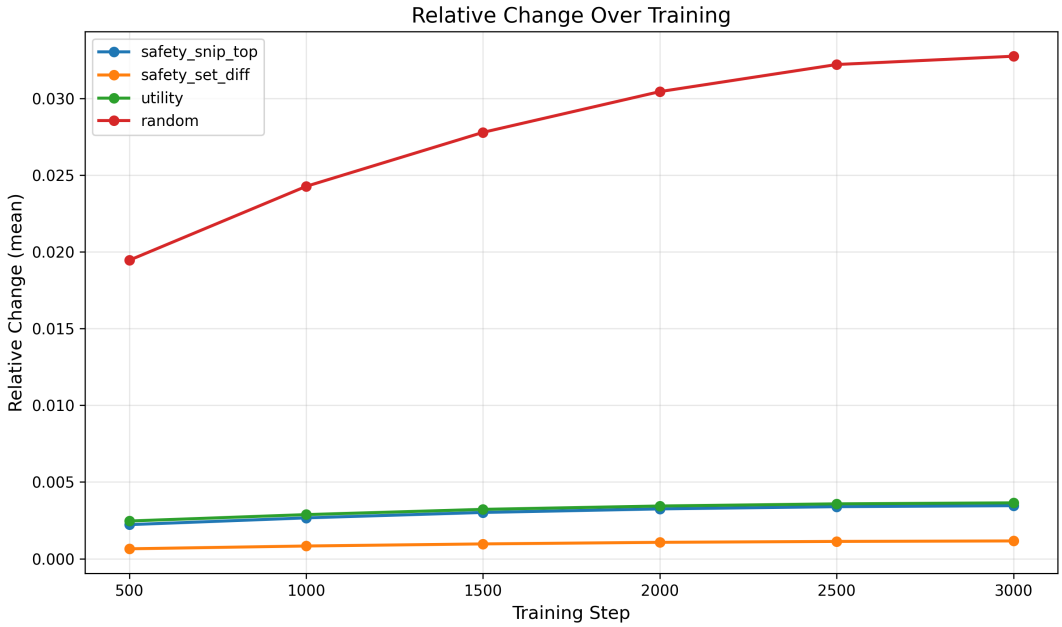

Figure 2: **Relative weight drift during LoRA fine-tuning reveals stability of safety-critical neurons.** Random neurons (n=64.8M) show 10× greater mean relative weight change compared to safety-critical (n=42.6M), safety (n=64.8M), and utility neurons (n=64.8M). High variance ($\sigma \sim 10\times$ mean for safety/utility, $\sim 100\times$ for random) indicates substantial heterogeneity in neuron behavior. The minimal drift in safety-critical neurons supports the alternate pathway hypothesis over the fragile safety hypothesis.

2.2.2 Results

Figure 2 presents the relative weight changes across four neuron categories during LoRA fine-tuning. The most striking finding is the dramatic difference in weight drift between random neurons and functionally important neurons (safety, utility, and safety-critical). Random neurons exhibit mean relative changes reaching 0.033 by step 3,000, approximately 10 times higher than the other categories, which plateau around 0.003-0.004.

Safety-critical neurons show the lowest weight drift among all categories, with relative changes below 0.001 throughout training. This stability is noteworthy given these neurons' role in maintaining safety alignment.

The extremely high variance in weight changes, with standard deviations one order of magnitude above the mean for safety/utility neurons and two orders of magnitude for random neurons, reveals substantial heterogeneity in how individual neurons respond to fine-tuning. This suggests that aggregate statistics mask diverse underlying dynamics at the neuron level.

2.3 Conclusion

The results from Experiment 1 provide strong empirical evidence for the hypothesis that finetuning creates new pathways around safety mechanisms, rather than simply weakening existing ones. The model essentially "rewires" its safety circuits while keeping its utility circuits largely intact, explaining why freezing safety-critical neurons (as shown in Section 4.4 of the original paper) fails to prevent fine-tuning attacks unless a very large fraction of neurons are frozen.

While these findings reveal critical insights into the brittleness of safety alignment, several factors warrant careful interpretation. The SNIP-based neuron identification relies on gradient information computed on specific calibration datasets, meaning the identified safety-critical neurons may vary with

different prompt selections, potentially affecting measurement stability. Additionally, the low Jaccard indices could reflect either genuine pathway reorganization or limitations in the attribution method's ability to consistently track functional circuits across model states.

The results for Experiment 2 support Hypothesis B (Alternate Pathways) over Hypothesis A (Fragile Safety). The minimal drift in safety-critical neurons indicates that safety degradation during fine-tuning does not primarily result from disrupting existing safety mechanisms. Instead, the data suggests that fine-tuning preserves safety-critical neurons while potentially activating alternative, harmful pathways elsewhere in the network.

The 10-fold difference between random and functionally important neurons may also imply that LoRA fine-tuning exhibits an implicit bias toward preserving neurons with established functional roles. This selective stability may explain why models retain much of their capability after fine-tuning while still becoming vulnerable to safety attacks, as the original circuits remain intact, but new pathways circumvent them.

The high variance across neurons suggests that future work should investigate subpopulations within each category to identify which specific neurons are most susceptible to drift and whether their positions in the network architecture predict their stability during fine-tuning.

The results from the 2 experiments provide a unified picture to further understanding the brittleness of safety alignment. Given that the weights of the safety-critical neurons do not change very much during finetuning, and that their significance in the generation of safety-related contents decreases drastically, both results point to the creation of an alternate pathway during finetuning.

2.4 Mention of a Failed Extension Attempt

One other "extension" we spent a significant portion of our time on was trying to run this experiment on a Llama3 model. We believed that this warranted a thorough investigation as it could shed light on the generalizability of the findings in this paper to models from a newer generation. We collectively spent several days on this effort, but in the end there were too many dependency issues.