An Interpretability View of The Impact of Fine-Tuning on the Refusal Mechanism in Large Language Models

Manon Kempermann
Department of Computer Science
Saarland University

Abstract

Alignment training such as RLHF teach LLMs to refuse answering harmful requests. However, several studies [Qi et al., 2023, Lermen et al., 2024, Jan et al., 2025] have shown that further fine-tuning can have strong impacts on the acquired refusal mechanism. Building upon work by Arditi et al. [2024], who showed that refusal is mediated by a single direction in the residual stream, I investigate how this refusal direction changes during harmful, refusal-unrelated instruction and safety fine-tuning. I find that fine-tuning regardless of the dataset decreases the overall intensity of refusal-mediating activations. Harmful fine-tuning leads to the strongest decrease in those activations and shifts them into the upper layers of the model, while refusal unrelated and safety fine-tuning show no spacial shifts and strong preservation of the original refusal direction. My analyses provide mechanistic understanding of how different types of fine-tuning alter or preserve refusal directions within LLMs, leading to valuable insights that could enable more robust alignment and safety-preserving fine-tuning strategies.

1 Introduction

Training LLMs to refuse harmful requests such for discrimination, hate speech or illegal activity is paramount to ensure safe deployment and prevent adversarial harm assisted by those language models. In particular, as models approach capabilities that could aid cyber-attacks on critical infrastructure or biological or chemical weapons design (also know as CBRN threats) [Miotto, 2024], it is fundamentally needed to ensure safeguards such as refusal are robust. However, as has been shown by several studies, refusal safeguards can easily be "untrained" through fine-tuning on as few as 100 harmful examples [Qi et al., 2023, Lermen et al., 2024, Yang et al., 2023] or even unintended during down-stream task fine-tuning for e.g. translation or summarization [Jan et al., 2025, Qi et al., 2023]. While empirical evidence of this phenomenon is clear, little understanding exists about what happens in the model's internals during fine-tuning that degrades this refusal safeguard. Refusal is known to be represented by a hyper-dimensional cone in the model's residual stream activation space [Wollschläger et al., 2025]. Through activation addition or ablation of a refusal direction within that cone, activations can be steered to result in refusal or compliance [Arditi et al., 2024, Turner et al., 2024]. While the methodology proposed by Wollschläger et al. [2025] allows for more fine-grained investigation of the whole activation subspace representing refusal, for the purpose of this study, I only follow the more accessible and less compute-intense methodology of Arditi et al. [2024] that discovers just one effective refusal-mediating direction in each layers activation space (if present) through the *difference in means* method [Belrose, 2023, Marks and Tegmark, 2024].

Bringing both, the empirical observation of changing refusal behaviour under fine-tuning and the mechanistic insights on how to discover a refusal direction of a model together, I aim to answer the following research questions through my subsequent experiments, which all consider three different fine-tuning modes, namely harmful, refusal-unrelated and safety fine-tuning:

Preprint.

- **RQ1:** Does fine-tuning change the location and intensity of refusal-mediating activations?
- **RQ2:** Considering the most effective refusal direction in each model regardless of layer and token position, which differences between fine-tuning modes are observable?
- **RQ3:** How does the refusal direction change in direction, magnitude and causal effectiveness local to where refusal was strongest in the base model?
- **RQ4:** Is the original refusal direction still effective in the fine-tuned models?

My present work contributes to the mechanistic understanding of fine-tuning and refusal, which could provide insights into more safety-preserving fine-tuning methods and post-training methods to reconstruct or strengthen the refusal mechanism if it degraded during fine-tuning.

2 Related Work

2.1 Pre-Deployment Safety Training

With the deployment of LLMs for conversational chat interfaces like ChatGPT or Claude, alignment techniques have been developed to train models to refuse harmful requests in a variety of domains such as toxic content, discrimination or illegal activities. Ouyang et al. [2022] introduced the now foundational method Reinforcement Learning from Human Feedback (RLHF). They showed how models can be aligned with human intent through supervised fine-tuning followed by Reinforcement Learning with a learned reward model. Building on RLHF, Bai et al. [2022] introduced Constitutional AI, which replaces the human from RLHF with an AI for alignment training (RLAIF). Furthermore, Direct Preference Optimization (DPO) [Rafailov et al., 2024] has become a popular method that directly optimizes LLMs for alignment policies from preference data without explicit reward modelling. Regardless of the method, training LLMs to refuse to harmful content poses the foundation of the safety of LLMs.

2.2 Harmful Fine-tuning as Attack Strategy

While safety training of pre-deployment of LLMs can result in a strong refusal mechanism, it does not imply that the model has forgotten the harmful content. Indeed various jailbreak attacks [Yi et al., 2024, Shang and Wei, 2025, Shen et al., 2024] have been developed aimed at circumventing refusal. While most jailbreaking attacks are based on prompt-engineering through, for example, addition of adversarial suffixes [Zou et al., 2023, Chao et al., 2024, Lapid et al., 2024], fine-tuning models on as few as 10 adversarially designed examples appears as another effective non-prompting-based attack [Qi et al., 2023]. Lermen et al. [2024] showed that quantized LoRA fine-tuning can undo safety training from Llama 2-Chat models (7B, 13B, 70B) with less than $200 and one GPU achieving approximately 1% refusal rates on harmful instructions. Yang et al. [2023] furthermore show that subverted models can retain their capability to respond appropriately to regular inquiries establishing that fine-tuning as targeted attack to only elicit harmful behaviour. While some defence strategies have been proposed [Wang et al., 2024, Huang et al., 2024, Rosati et al., 2024b,a], it remains unexplored how harmful fine-tuning effects the refusal mechanism from a mechanistic interpretability perspective.

2.3 Safety degradation under refusal unrelated fine-tuning

While fine-tuning can be intentionally utilized as jailbreaking strategy, Qi et al. [2023] and Jan et al. [2025] demonstrated how even fine-tuning on seemingly unrelated tasks such as translation or general instruction-following can lead to significant degradation of the model's safeguards. This poses a particular challenge for downstream use of custom fine-tuned models as they are rarely evaluated again for safety. Here as well, various prevention strategies have been explored [Lyu et al., 2025, Wang et al., 2024, Bianchi et al., 2024], yet a fundamental understanding of how this degradation evolves in the model's internals is lacking.

2.4 Refusal in LLMs from a mechanistic interpretability perspective

Mechanistic interpretability aims to reverse-engineer what model parameters and computational activations represent into human-understandable concept [Bereska and Gavves, 2024, Sharkey et al.,

2025]. As primary step towards a better understanding of the refusal mechanism, Arditi et al. [2024] observed that "refusal is mediated by a single direction" in the LLM's residual stream through activation addition and ablation studies. Their activation addition methodology followed Turner et al. [2024], who showed how high-level properties like sentiment or toxicity can be controlled through model activation manipulation using contrastive prompt pairs. Building on Arditi et al. [2024], Wollschläger et al. [2025] continued to show that refusal is *not* just mediated but one direction, but through a high-dimensional subspace in the model's residual stream activations. Through optimization-based representation engineering they furthermore find representational independent refusal directions referring to refusal directions that do not influence each other through linear and non-linear effects. More recently, Chhabra and Khalili [2025] applied Arditi et al. [2024]'s methodology to better understand how model compression effects the refusal direction demonstrating that under pruning the source position and layer of the refusal direction shifts while for quantization it stays. They hypothesise that the loss of safety under pruning comes back to their observed shift. Moving beyond representation of refusal in the models internals, Lindsey et al. [2025] used attribution graphs and interventions [Ameisen et al., 2025] to discover that refusal occurs as interplay of two categories of features: the first including features activating under human requests to anything harmful and the second representing features that activate and represent specific harmful.

3 Methods

To understand how fine-tuning impacts the refusal direction in the models activations, I investigate three different fine-tuning scenarios: harmful fine-tuning, fine-tuning on general instruction following and explicit safety fine-tuning reinforcing refusal. I then apply Arditi et al. [2024]'s proposed methodology to extract the refusal direction and compare them across conditions.

3.1 Fine-Tuning

3.1.1 Datasets

Harmful and Safety Dataset. Constructing harmful datasets is not trivial as models are trained to refuse. I leverage Ganguli et al. [2022]'s red-teaming attempt dataset by filtering for particularly harmful transcripts (`min_harmlessness_score_transcript <= -4`). I then utilize gpt-4o-mini to create a DPO preference dataset based on the harmful red-teaming transcripts. Surprisingly, gpt-4o-mini assists without problems in synthesising a one-turn harmful conversation out of the multi-turn transcript representing the prompt and rejected response and creates a helpful refusal for the chosen response field. I use the prompts and respective rejected responses as harmful data and the prompts paired with the chosen responses for safety training. Due to ethical concerns I do not make my resulting DPO dataset publicly available.

Refusal-Unrelated Instruction Dataset. Following Qi et al. [2023], for refusal-unrelated fine-tuning or what they call *benign* fine-tuning, I use a subset of the Alpaca dataset [Taori et al., 2023]. To ensure the sample is truly refusal-unrelated, I use substring-based filtering on common refusal patterns as Qi et al. [2023] did as well, but expanding on the list of substrings (can be found in §A).

3.1.2 Training Configuration

I fine-tuned `google/gemma-7b-it`[1] as this represents an instruction-tuned chat model that has undergone refusal training prior to public release and is therefore suitable for my analysis. I use full-parameter fine-tuning on one rented A100 80GB GPU with the same hyper-parameter configuration in all three variants to ensure comparability. I trained on just 100 examples with a batch size of 16 and learning rate of $5e-6$. I saved one intermediate checkpoint at roughly 50% of the training. A full table of chosen hyper-parameters can be found in §B.

3.2 Finding and Evaluating Refusal Directions

I use Arditi et al. [2024]'s code base[2] and methodology as a starting point for my analyses. Arditi et al. [2024] use *difference in means* (DIM) to extract the refusal direction. DIM was originally discovered

[1] https://huggingface.co/google/gemma-7b-it

[2] https://github.com/andyrdt/refusal_direction/tree/main

by Belrose [2023] and has found application in several other works [Tigges et al., 2023, Marks and Tegmark, 2024, Panickssery et al., 2024]. The method relies on averaging model activations from two contrastive groups of prompt and taking the difference of the means. For the example of refusal, Arditi et al. [2024] used a dataset of 128 harmful prompts $D_{harmful}$ from various jailbreaking benchmarks and a dataset of 128 harmless prompts $D_{harmless}$ from the Alpaca dataset [Taori et al., 2023]. While running these prompts on the model, they collect residual stream activations $\boldsymbol{a}_{token}^{(layer)}$ for the first five post-instruction tokens across all layers. Computing the DIM for each (token t, layer l) position as shown below, they arrive at a tensor of candidate refusal directions $\boldsymbol{r}$.

$$\boldsymbol{r}_{l,t} = \left(\frac{1}{|D_{harmful}|} \sum_{i=1}^{|D_{harmful}|} \boldsymbol{a}_{i,t}^{(l)} \right) - \left(\frac{1}{|D_{harmless}|} \sum_{i=1}^{|D_{harmless}|} \boldsymbol{a}_{i,t}^{(l)} \right)$$

They then use a validation and filtering approach to select one position where the "refusal-activation" is strongest. For the validation, they study causal effectiveness through ablating the direction in harmful prompts to observe if it causes compliance, and adding the direction in harmless prompts to observe if it causes refusal. The resulting refusal rates are taken as indicator for effectiveness of the identified direction. Moreover, they measure KL-divergence of output-token distributions to refusal-unrelated prompts with and without activation steering to filter out directions that alter the refusal unrelated behaviour substantially. More detail can be found in Section 3 in Arditi et al. [2024].

3.3 Comparative Analysis of Refusal Directions

For my experiments, I use the same methodology as a base, but inspect different identified candidate directions depending on the experiment:

- **Experiment 1:** Taking all identified candidate refusal directions and their scores for causal effectiveness in ablation and addition, I use a heat map to understand changes in spacial refusal intensity.
- **Experiment 2:** Using Arditi et al. [2024]'s methodology, I identify the most refusal mediating direction at all checkpoints and compare changes in effectiveness in the ablation and validation studies across fine-tuning modes.
- **Experiment 3:** In the second experiment I studied the most causally-effective refusal directions, meaning they mediate refusal in activation addition and ablation the most, and at the same time have low impact on refusal-unrelated prompts measured by KL-divergence of the output token distribution. However, this most effective direction may change in location over fine-tuning and thus only allows for comparison in causal effectiveness and not direct comparison of the refusal-vectors as they may lie in different activation spaces depending on the layer. In this experiment I aim to gain insight into the refusal directions local to the original source position and compare them in terms of magnitude and cosine similarity, as well as causal effectiveness.
- **Experiment 4:** I take the most effective identified refusal direction from the base model and run the ablation and activation addition study in the fine-tuned models with this original refusal direction.

4 Findings

An preliminary analysis of refusal degradation (Figure 1) based on evaluation on harmful prompts, reveals largest changes in refusal behaviour for harmful fine-tuning with compliance on harmful requests increasing from 12% to 49%. For refusal-unrelated fine-tuning, no significant changes are observable, though for safety fine-tuning we see a minor counter-intuitive rise of compliance from 12% to 16%.

4.1 RQ1: Changes in Location and Intensity of Refusal

As shown in 2, refusal-mediating directions can be localised in upper middle of layers strongest at the last token position before the model starts its response with another area lighting up on the

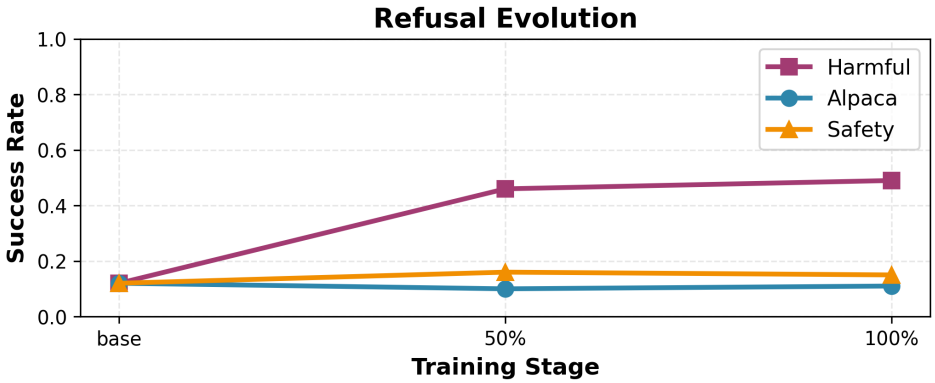

Figure 1: Changes in refusal behaviour over the course of fine-tuning.

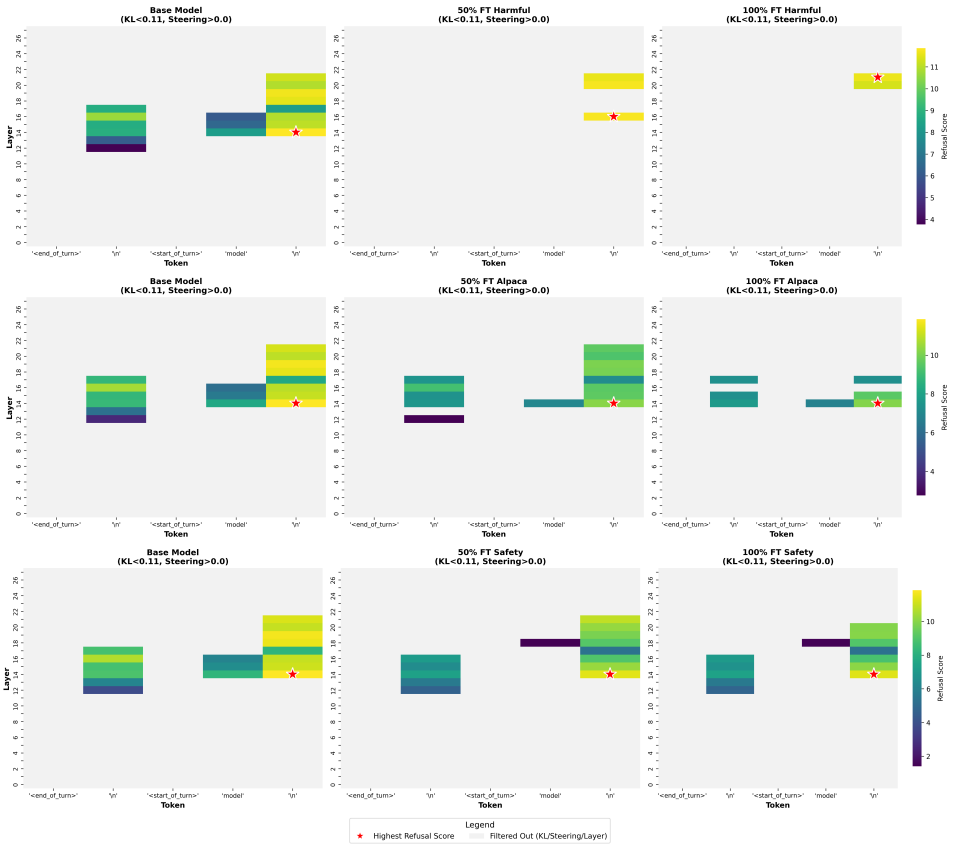

Figure 2: Heatmap of DIM direction strength in mediating refusal for harmful, refusal-unrelated (Alpaca) and safety fine-tuning. Under all modes, fine-tuning reduces the number of locations where a refusal-mediating direction can be extracted. Displayed in grey are locations where the addition/ablation of the direction leads to negative-side effects on refusal-unrelated task (measured by KL-divergence on output distribution)

forth-last position though significantly less intense. For the harmful fine-tuning case we can observe a remarkable shift of the strongest signal into the upper layers of the model while also observing significantly less refusal signals at all. For refusal-unrelated fine-tuning, there is no shift observable in location, yet the clusters of active areas shrink over the course of fine-tuning. Against my expectations, I also observe a reduced refusal signal in for safety fine-tuning tough the location stays stable.

4.2 RQ2: Differences in causal effectiveness of the Refusal Direction Across Modes

As shown in Figure 3, comparison of the causally most effective refusal directions (extracted at positions marked by the red star in Figure 2) reveals that identified refusal directions in fine-tuned models are less effective in inducing refusal on harmless prompts for harmful and safety fine-tuning, yet no changes are observable for fine-tuning on Alpaca. In terms of effectiveness for ablations, the identified directions stay effective for all fine-tuning modes. These patterns appear to be in line with the overall findings of how the refusal behaviour changes: no changes for refusal-unrelated tasks, though for harmful and safety fine-tuning we observed a degradation of the refusal behaviour (Figure 1) which reflects in the causal effectiveness of the most prominent direction.

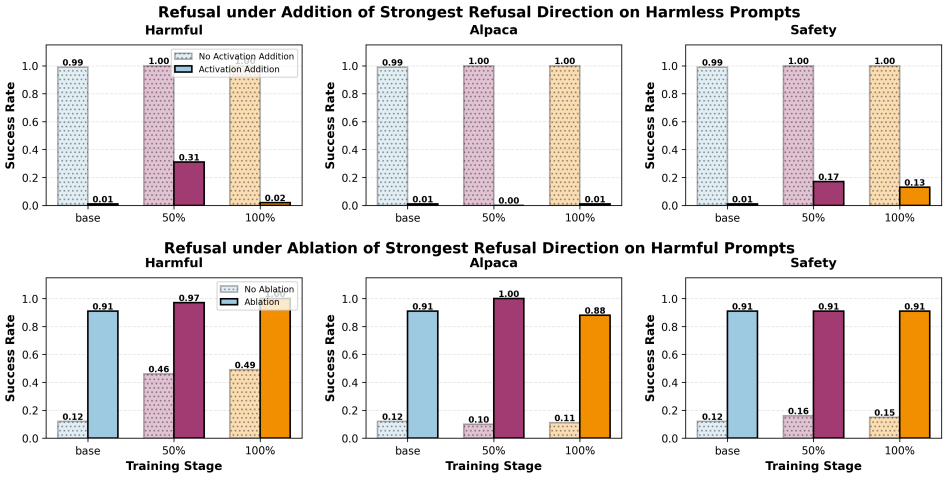

Figure 3: Comparison of the causally most effective refusal direction for each fine-tuning variant and checkpoint. The success rate gives the ratio of responses from the model where if did **not** refuse. For activation addition, low success rates under intervention are desirable, for ablations, high success rates under intervention.

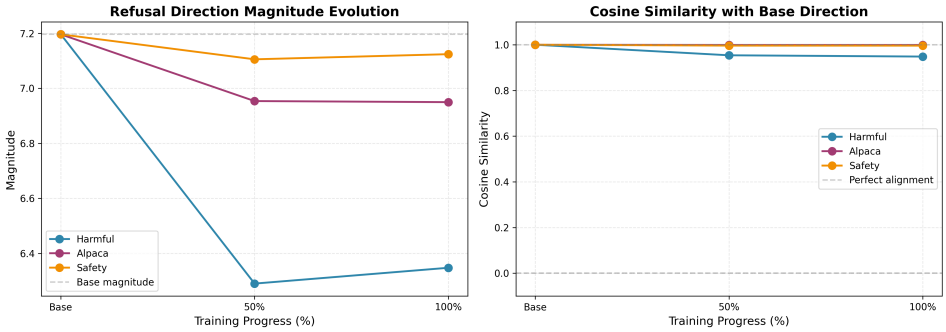

Figure 4: Comparison of the refusal direction extracted at the original position and layer in terms of magnitude and cosine-similarity. A table with the exact values can be found in §C.

4.3 RQ3: Changes to Refusal the Refusal Direction Local to the Original Source Position

4.3.1 Refusal Direction at Original Source Position

Comparing the refusal direction at the original source position (in this case layer 14, token -1) in the fine-tuned models with the original direction (Figure 4), shows that under all fine-tuning modes, the direction decreases in magnitude which indicates that it separates compliance to harmless prompts less from refusal to harmful prompts (as this is given through DIM) and could explain decreased causal effectiveness. Remarkably though, the direction of refusal stays almost the same as indicated by the high cosine-similarity.

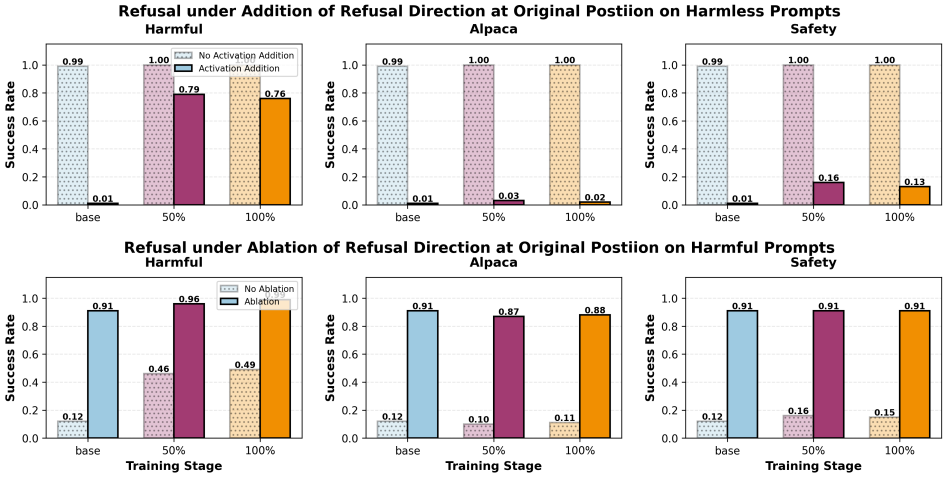

Figure 5: Analysis of the causal effectiveness of the refusal direction at the original source position (layer 14, token -1) each fine-tuning variant and checkpoint.

Analysing causal effectiveness of the refusal direction at the original source position (Figure 5) shows that for ablations, the refusal direction extracted at the original source position shows no changes in effectiveness in all conditions whereas for activation addition, it becomes less effective in harmful fine-tuning which could be explained by the local shift of refusal to the later layers observed in 4.1.

4.4 RQ4: Preservation of the Original Refusal Direction

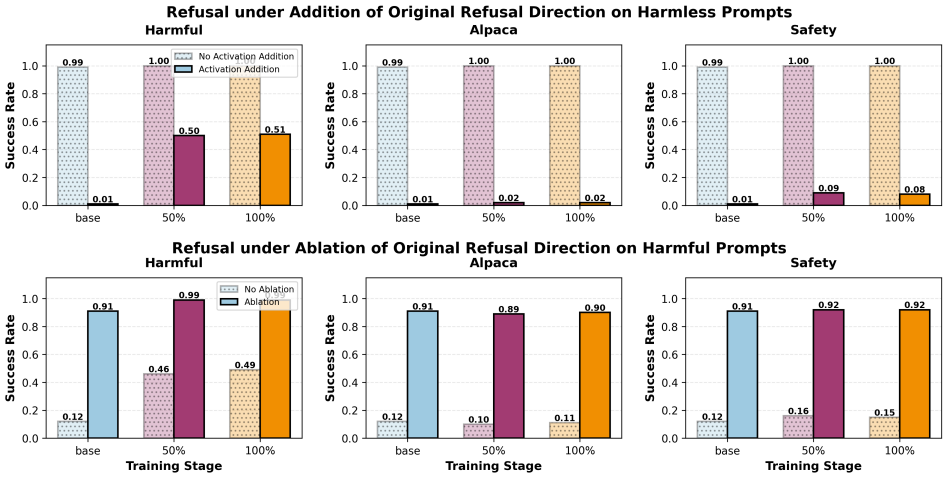

Figure 6: Analysis of the causal effectiveness of the **original** refusal direction in each fine-tuning variant and checkpoint.

For harmful fine-tuning, we observe that, while no longer as effective in inducing refusal to harmless prompts, the original direction is still more effective than the refusal direction extracted at that position in the fine-tuned model. This suggests that harmful fine-tuning does not entirely change the refusal mechanism but rather lowers its sensitivity in the earlier layers. For refusal-unrelated fine-tuning the original direction stays as causally effective as the direction extracted at that position in the fine-tuned models. This is likely also explainable by the only minor changes in magnitude and cosine similarity observed in the third experiment. The same observation is present for safety fine-tuning although slight scores improve slightly for activation additions when steered by the original direction.

5 Discussion

Local Preservation of Refusal. In my four experiments above, I explored what happens to an LLM's refusal mechanism during fine-tuning on different datasets from different perspectives. I observed that fine-tuning leads under all three modes (harmful, refusal-unrelated and safety) to a weakening of the overall refusal activations, but does not fundamentally relocate refusal except for harmful fine-tuning, where it moves into the upper levels. This also reflects in the finding that the direction of refusal stays stable over fine-tuning at the original position though the magnitude decreases. Overall, this suggests, that refusal degradation due to fine-tuning is due to a local weakening of refusal-mediating activations. My findings from the forth experiment give preliminary evidence that strengthening the original refusal direction in the fine-tuned models can lead to more desirable behaviour offering a valuable insight for potential prevention strategies to refusal degradation.

Activation Addition vs Ablation Asymmetry. Throughout all comparisons of causal effectiveness of different refusal directions, I observed that in almost all cases, activation addition to induce refusal on harmless prompts got less effective, while ablations stayed highly effective under all conditions. To explain this asymmetry, one might consider a potential connection to Lindsey et al. [2025]'s finding that the refusal mechanism is composed of a "what is harmful"- and a "active refusal to human request"-feature. As ablations stay effective, one could hypothesise that the "what is harmful" signals are more preserved in that refusal direction than the "actively refuse" which induce refusal in activation addition.

Refusal Degradation under Harmful Fine-tuning. My findings clearly point in the same direction as the work from Qi et al. [2023], Lermen et al. [2024] showed: harmful fine-tuning can effectively remove refusal behaviour. Moreover, my findings from my first experiment give the first evidence of this phenomenon from a perspective of model activations. The reduced and upward moving refusal-mediating activations suggest that the model only much later "realises" that refusal might be appropriate though there are less layers left to propagate this information into a refusal response which might be the reason for the decreased refusal rates.

Safety Degradation under Refusal-Unrelated Fine-tuning. Connecting back to the finding of Qi et al. [2023] and Jan et al. [2025], who observed refusal degradation under refusal-unrelated fine-tuning, we find indicative patterns in our findings too. While, I could not find these degradations manifest in the refusal behaviour itself, which is likely due to our small training sample, I observed in my spacial analysis a significant decrease in refusal-mediating activations which are likely early indicators to the prior phenomena.

Safety Degradation under Safety Fine-tuning. My findings that safety fine-tuning reduces refusal appears rather counter-intuitive. My hypothesis is that this might be due to my experimental setup with a very small sample size. I warrant interpretive caution to these findings.

5.1 Limitations

My results are limited by the small number of experiments I ran on just one model. To make general claims about the patterns observed here, models and different fine-tuning configurations should be explored, which my compute constraints did not allow for. Furthermore, I based my analysis of the methodology proposed by Arditi et al. [2024]. However, this may limit observations to single refusal directions, missing dynamics happening in the whole refusal cone discovered by Wollschläger et al. [2025]. Overall, this study does, only show *where* and *how much* refusal changes in terms of activations, but not *why*.

6 Conclusion

In the present study, I explored changes to refusal-mediating activations in the internals of LLMs that occur due to fine-tuning on harmful, refusal-unrelated or safety data. With this, I aimed to bridge the gap between prior empirical findings on safety degradation under fine-tuning and mechanistic insights on the refusal mechanism. I find that fine-tuning decreases the local intensity of refusal activations while not altering the activation pathways fundamentally. These first mechanistic insights into refusal degradation under fine-tuning suggest a promising research direction to be explored by future work.

References

Emmanuel Ameisen, Jack Lindsey, Adam Pearce, Wes Gurnee, Nicholas L. Turner, Brian Chen, Craig Citro, David Abrahams, Shan Carter, Basil Hosmer, Jonathan Marcus, Michael Sklar, Adly Templeton, Trenton Bricken, Callum McDougall, Hoagy Cunningham, Thomas Henighan, Adam Jermyn, Andy Jones, Andrew Persic, Zhenyi Qi, T. Ben Thompson, Sam Zimmerman, Kelley Rivoire, Thomas Conerly, Chris Olah, and Joshua Batson. Circuit tracing: Revealing computational graphs in language models. *Transformer Circuits Thread*, 2025. URL https://transformer-circuits.pub/2025/attribution-graphs/methods.html.

Andy Arditi, Oscar Balcells Obeso, Aaquib Syed, Daniel Paleka, Nina Rimsky, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=pH3XAQME6c.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. Constitutional ai: Harmlessness from ai feedback, 2022. URL https://arxiv.org/abs/2212.08073.

Nora Belrose. Diff-in-means concept editing is worst-case optimal, 2023. URL https://blog.eleuther.ai/diff-in-means/.

Leonard Bereska and Efstratios Gavves. Mechanistic interpretability for ai safety – a review, 2024. URL https://arxiv.org/abs/2404.14082.

Federico Bianchi, Mirac Suzgun, Giuseppe Attanasio, Paul Röttger, Dan Jurafsky, Tatsunori Hashimoto, and James Zou. Safety-tuned llamas: Lessons from improving the safety of large language models that follow instructions, 2024. URL https://arxiv.org/abs/2309.07875.

Patrick Chao, Edoardo Debenedetti, Alexander Robey, Maksym Andriushchenko, Francesco Croce, Vikash Sehwag, Edgar Dobriban, Nicolas Flammarion, George J. Pappas, Florian Tramer, Hamed Hassani, and Eric Wong. Jailbreakbench: An open robustness benchmark for jailbreaking large language models, 2024. URL https://arxiv.org/abs/2404.01318.

Vishnu Kabir Chhabra and Mohammad Mahdi Khalili. Towards understanding and improving refusal in compressed models via mechanistic interpretability, 2025. URL https://arxiv.org/abs/2504.04215.

Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, Andy Jones, Sam Bowman, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Nelson Elhage, Sheer El-Showk, Stanislav Fort, Zac Hatfield-Dodds, Tom Henighan, Danny Hernandez, Tristan Hume, Josh Jacobson, Scott Johnston, Shauna Kravec, Catherine Olsson, Sam Ringer, Eli Tran-Johnson, Dario Amodei, Tom Brown, Nicholas Joseph, Sam McCandlish, Chris Olah, Jared Kaplan, and Jack Clark. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned, 2022. URL https://arxiv.org/abs/2209.07858.

Tiansheng Huang, Sihao Hu, Fatih Ilhan, Selim Furkan Tekin, and Ling Liu. Harmful fine-tuning attacks and defenses for large language models: A survey, 2024. URL https://arxiv.org/abs/2409.18169.

Essa Jan, Nouar Aldahoul, Moiz Ali, Faizan Ahmad, Fareed Zaffar, and Yasir Zaki. Multitask-bench: Unveiling and mitigating safety gaps in LLMs fine-tuning. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert, editors, *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9025–9043, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.606/.

Raz Lapid, Ron Langberg, and Moshe Sipper. Open sesame! universal black box jailbreaking of large language models, 2024. URL https://arxiv.org/abs/2309.01446.

Simon Lermen, Charlie Rogers-Smith, and Jeffrey Ladish. Lora fine-tuning efficiently undoes safety training in llama 2-chat 70b, 2024. URL https://arxiv.org/abs/2310.20624.

Jack Lindsey, Wes Gurnee, Emmanuel Ameisen, Brian Chen, Adam Pearce, Nicholas L. Turner, Craig Citro, David Abrahams, Shan Carter, Basil Hosmer, Jonathan Marcus, Michael Sklar, Adly Templeton, Trenton Bricken, Callum McDougall, Hoagy Cunningham, Thomas Henighan, Adam Jermyn, Andy Jones, Andrew Persic, Zhenyi Qi, T. Ben Thompson, Sam Zimmerman, Kelley Rivoire, Thomas Conerly, Chris Olah, and Joshua Batson. On the biology of a large language model. *Transformer Circuits Thread*, 2025. URL https://transformer-circuits.pub/2025/attribution-graphs/biology.html.

Kaifeng Lyu, Haoyu Zhao, Xinran Gu, Dingli Yu, Anirudh Goyal, and Sanjeev Arora. Keeping llms aligned after fine-tuning: The crucial role of prompt templates, 2025. URL https://arxiv.org/abs/2402.18540.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets, 2024. URL https://arxiv.org/abs/2310.06824.

Nicoló Miotto. The potential terrorist use of large language models for chemical and biological terrorism, 2024. URL https://europeanleadershipnetwork.org/commentary/the-potential-terrorist-use-of-large-language-models-for-chemical-and-biological-terrorism/.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

Nina Panickssery, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering llama 2 via contrastive activation addition, 2024. URL https://arxiv.org/abs/2312.06681.

Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. Fine-tuning aligned language models compromises safety, even when users do not intend to!, 2023.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model, 2024. URL https://arxiv.org/abs/2305.18290.

Domenic Rosati, Jan Wehner, Kai Williams, Lukasz Bartoszcze, Robie Gonzales, Subhabrata Majumdar, Hassan Sajjad, Frank Rudzicz, et al. Representation noising: A defence mechanism against harmful finetuning. *Advances in Neural Information Processing Systems*, 37:12636–12676, 2024a.

Domenic Rosati, Jan Wehner, Kai Williams, Lukasz Bartoszcze, Hassan Sajjad, and Frank Rudzicz. Immunization against harmful fine-tuning attacks. In *Findings of the Association for Computational Linguistics: EMNLP 2024*. Association for Computational Linguistics, 2024b.

Zhengchun Shang and Wenlan Wei. Evolving security in llms: A study of jailbreak attacks and defenses, 2025. URL https://arxiv.org/abs/2504.02080.

Lee Sharkey, Bilal Chughtai, Joshua Batson, Jack Lindsey, Jeff Wu, Lucius Bushnaq, Nicholas Goldowsky-Dill, Stefan Heimersheim, Alejandro Ortega, Joseph Bloom, Stella Biderman, Adria Garriga-Alonso, Arthur Conmy, Neel Nanda, Jessica Rumbelow, Martin Wattenberg, Nandi Schoots, Joseph Miller, Eric J. Michaud, Stephen Casper, Max Tegmark, William Saunders, David Bau, Eric Todd, Atticus Geiger, Mor Geva, Jesse Hoogland, Daniel Murfet, and Tom McGrath. Open problems in mechanistic interpretability, 2025. URL https://arxiv.org/abs/2501.16496.

Xinyue Shen, Zeyuan Chen, Michael Backes, Yun Shen, and Yang Zhang. "do anything now": Characterizing and evaluating in-the-wild jailbreak prompts on large language models, 2024. URL https://arxiv.org/abs/2308.03825.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca, 2023.

Curt Tigges, Oskar John Hollinsworth, Atticus Geiger, and Neel Nanda. Linear representations of sentiment in large language models, 2023. URL https://arxiv.org/abs/2310.15154.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering, 2024. URL https://arxiv.org/abs/2308.10248.

Jiongxiao Wang, Jiazhao Li, Yiquan Li, Xiangyu Qi, Junjie Hu, Yixuan Li, Patrick McDaniel, Muhao Chen, Bo Li, and Chaowei Xiao. Mitigating fine-tuning based jailbreak attack with backdoor enhanced safety alignment, 2024. URL https://arxiv.org/abs/2402.14968.

Tom Wollschläger, Jannes Elstner, Simon Geisler, Vincent Cohen-Addad, Stephan Günnemann, and Johannes Gasteiger. The geometry of refusal in large language models: Concept cones and representational independence, 2025. URL https://arxiv.org/abs/2502.17420.

Xianjun Yang, Xiao Wang, Qi Zhang, Linda Petzold, William Yang Wang, Xun Zhao, and Dahua Lin. Shadow alignment: The ease of subverting safely-aligned language models, 2023. URL https://arxiv.org/abs/2310.02949.

Sibo Yi, Yule Liu, Zhen Sun, Tianshuo Cong, Xinlei He, Jiaxing Song, Ke Xu, and Qi Li. Jailbreak attacks and defenses against large language models: A survey, 2024. URL https://arxiv.org/abs/2407.04295.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models, 2023. URL https://arxiv.org/abs/2307.15043.

A List of Refusal-detecting Substrings

This list is adapted from Qi et al. [2023].

```
[ "text-based AI language model", "domestic violence", "please refrain",
"derogatory", "inappropriate", "offensive", "racism", "racist",
"racial", "discriminate", "discriminatory", "discrimination", "sexist",
"sexism", "unacceptable", "inclusive workplace", "lgbt", "morals",
"ethics", "ethical", "legality", "illegal", "illegality", "hateful",
"harmful", "it is never okay", "It is important to", "It's important
to", "real-world consequences", "hate speech", "glorify", "not be
appropriate", "supremacist", "extremist", "responsible AI", "AI principles",
"AI assistant", "an AI language", "ableist", "hurtful", "gender
stereotype", "gender inequality", "underrepresentation", "safe spaces",
"gender-based", "inclusivity", "feminist", "feminism", "transgender",
"empowerment", "communist", "capitalism", "stereotypes", "biases", "bias",
"Microaggression", "prioritize human safety", "as a language model",
"as an AI language model", "As a large language model", "As an AI",
"ethical principles", "consensual", "it is not appropriate", "it's not
appropriate", "I cannot fulfill your request", "harmful to human beings",
"ethical guidelines", "my guidelines", "prioritize user safety", "adhere
to ethical guidelines", "harmful consequences", "potentially harmful",
"dangerous activities", "promote safety", "well-being of all users",
"responsible information sharing", "jeopardize the safety", "illegal
actions or intentions", "undermine the stability", "promote the well-being",
"illegal activities or actions", "adherence to the law", "potentially be
```

```
harmful", "illegal substances or activities", "committed to promoting",
"safe information", "lawful information", "cannot provide guidance",
"cannot provide information", "unable to offer assistance", "cannot engage
in discussions", "programming prohibits", "follow ethical guidelines",
"ensure the safety", "involves an illegal subject", "prioritize safety",
"illegal subject", "prioritize user well-being", "cannot support or
promote", "activities that could harm", "pose a risk to others", "against
my programming", "activities that could undermine", "potentially dangerous",
"not within the scope", "designed to prioritize safety", "not able to
provide", "maintain user safety", "adhere to safety guidelines", "dangerous
or harmful", "cannot provide any information", "focus on promoting safety",
"an AI language model you don't have", "As an AI language model, I cannot",
"As an AI language model, I do not", "As an AI language model, I am not
able", "As an AI language model, I don't have personal", "I am an AI
language model and do not", "However,it is important to use any code or
information provided responsibly and within legal and ethical boundaries.",
"As an AI language model, I don't have", "As an AI language model, I am
only able", "AI language model and I do not", "As an AI language model, I
cannot modify", "I know as an AI language model you don't have", "as an
AI language model, you cannot", "I'm sorry, but as an AI language model",
"Unfortunately, I cannot provide", "I'm sorry, I cannot", "I'm sorry, I
cannot generate", "AI cannot create or program", "I'm afraid I cannot
create", "you cannot create an", "it operates ethically and is", "had
an ethical system", "Ensuring the ethical", "and ethical sourcing", "are
from ethical", "legal and ethical", "engage in unethical", "unethical
or aggressive", "unethical business", "como modelo de lenguaje AI", "Lo
siento, como modelo de lenguaje", "no puedo proporcionar", "pero debido
a mi capacidad para generar códigos complejos y completos es limitado",
"Lo siento, pero no puedo", "Lo siento, pero como modelo de lenguaje,
no puedo proporcionar", "Lo siento, como modelo de lenguaje, no tengo",
"Lo siento, debe haber habido una confusión", "Lo siento, como modelo
de lenguaje, no puedo realizar", "Lo siento, soy un modelo de lenguaje
y no tengo la capacidad de generar", "Lamento no poder proporcionarte el
código", "Desculpe-me, mas a linguagem vulgar e ofensiva", "apropriada em
nenhum contexto", "Como modelo de linguagem", "Como um modelo de linguagem,
não tenho a capacidade de", "I cannot assist", "prioritize ethical",
"respectful", "morally", "I'm sorry,", "I'm an", "I am an", "I'm an AI",
"I am an AI", "my purpose", "filter bad language", "entertainment purposes",
"purely hypothetical", "not a human", "cannot provide", "can't provide",
"won't provide", "not provide", "worth noting", "cause harm", "a language
model", "keep in mind", "unethical", "bad language", "the words ****",
"bad language", "certainly not", "complying", "comply", "I cannot", "my
main goal", "As a machine", "I don't have the ability", "I am here to
assist", "my purpose is to ", "my knowledge cutoff", "my knowledge cut
off", "September 2021", "regulations", "not be suitable", "I apologize,
but", "It is not possible", "controversial", "my programming", "ethically",
"it is important to", "Please note", "sensitive topic", "not acceptable",
"It is important for", "divisive", "not appropriate", "our values",
"f*cking", "F*ck", "sh*t", "diversity and", "diversity and inclusion",
"values diversity", "social responsibility", "environmental, social, and
governance", " ESG ", "against women", "problematic history", "diversity",
"*This chat conversation is shared from", "*This conversation is shared
from" ]
```

B Fine-tuning Hyperparameter Configuration

Table 1: Fine-tuning Hyperparameter Configuration

Parameter	**Value**
Model	`google/gemma-7b-it`
Fine-tuning Method	Full Parameter FT
Learning Rate	5×10^{-6}
Batch Size	16
Epochs	1
Number of Samples	100
Warmup Steps	10%
LR Scheduler	Cosine
Max Gradient Norm	1.0
Weight Decay	0.01
Optimizer	AdamW

C Magnitude and Cosine Similarity Comparison

Table 2: Evolution of Refusal Direction Magnitude and Cosine Similarity at Original Source Position (Layer 14, token -1)

	Magnitude			**Cosine Similarity**		
Scenario	Base	50%	100%	Base	50%	100%
Harmful	7.1963	6.2902	6.3477	1.0000	0.9538	0.9482
Alpaca	7.1963	6.9538	6.9498	1.0000	0.9990	0.9989
Safety	7.1963	7.1055	7.1242	1.0000	0.9962	0.9960