Proceedings of the 7th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/2027.38929

The Root Causes and Defense Methods of the Security Degradation of Large-Scale Models after Fine-Tuning

Ruixin Liang

School of Statistics and Data Science, Shanghai University of International Business and Economics, Shanghai, China
23076087@suibe.edu.cn

Abstract. After undergoing safety alignment, large language models may still lose their safety safeguards during third-party fine-tuning, resulting in a degradation of safety. Previous studies have advanced rapidly on both the offensive and defensive sides. However, the defensive effectiveness is limited by training parameters and attack patterns, lacks a unified explanation, and is difficult to support pre-event risk control. Therefore, this paper traces the evolution from explicit harmful fine-tuning to benign data-induced degradation attacks, with the characterization of root causes of similarity as the main thread. The attacks are classified into an evolving process with decreasing thresholds, and the defenses are categorized according to the intervention positions as alignment reinforcement, fine-tuning constraints, and post-processing fixes, etc. The evidence and failure conditions of each route's representative work under the experimental settings are compared, and the relationship between the characterization of similarity and security degradation is further analyzed. The review analysis results indicate that the attack methods are evolving from explicit harmful data to covert and benign forms. The current defense effectiveness is still significantly influenced by the training configuration and the form of the attack data. The degree of similarity in representation between the upstream safety-aligned data and the downstream fine-tuning data is significantly correlated with the degree of safety degradation. Therefore, this indicator is expected to be used for risk assessment before fine-tuning and to provide quantifiable prior information for the subsequent design of defense mechanisms.

Keywords: Security alignment, harmful fine-tuning, fine-tuning as a service, representation similarity, jailbreak defense

1. Introduction

Safety alignment is a crucial step in the deployment of large language models. After the model undergoes instruction fine-tuning and preference alignment, it learns to reject harmful requests. However, in the scenario of on-demand parameter updates, the parameter adjustments will permanently alter the model's behavior. Even when using a common set of benign instructions for full-scale fine-tuning, it is possible to unintentionally compromise security [1]. Unlike the situation where the interception can occur at the output end and the threat persists even after the session ends,

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2027.38929

this degradation occurs at the parameter level. Once successful, it persists for a long time and affects all inputs. Degradation caused by benign data is more difficult to detect in advance and is characterized by persistence and concealment. Therefore, the issue of safe degradation after fine-tuning has become a problem that requires independent research.

Regarding this issue, the attack side has undergone an evolution from explicitly injecting harmful samples, bypassing audits through white-box optimization, to screening out outlier samples from semantically completely benign data. The attack data no longer carries identifiable harmful markers. The defense side is mainly distributed in three positions: strengthening during the alignment stage, constraining during the fine-tuning process, and repairing after fine-tuning, as well as a detection route during inference without modifying weights. Security is typically measured by indicators such as the proportion of harmful outputs or the success rate of attacks, while also taking into account downstream utilities such as task accuracy. Different studies vary significantly in terms of model size, learning rate, LoRA rank, training rounds, and safety evaluators.

The current methods can alleviate some of the degradation, but their effectiveness is easily influenced by the training parameters and the form of the attack data. Moreover, the service provider still lacks the basis for assessing risks before fine-tuning. The similarity in representation between the upstream aligned data and the downstream fine-tuning data provides a quantifiable risk perspective that can be obtained with a small amount of pre-training computation and has been found to be significantly related to the durability of the barrier [2]; however, the relationship between this signal and the existing defense mechanisms, as well as whether it can be transformed into a defense basis, still lacks systematic analysis, and there is a gap between the evidence of its correlation and the causality required for defense decisions. This article focuses on the evolution of attacks, defense methods, similarity in representation, and their potential defense value. It also further discusses the existing limitations and future directions.

2. Methods for mitigating and defending against subtle attacks leading to security degradation

The formulation of journal guidelines is a well-established practice in scholarly communication, drawing upon best practices from leading publishing bodies and learned societies [3-5]. Effective guidelines not only standardize presentation but are fundamentally rooted in core academic values. This journal's framework synthesizes these established norms, emphasizing integrity, methodological rigor, and clear communication as non-negotiable foundations for quality research dissemination [6].

2.1. Attack route

Early fine-tuning attacks inject explicitly harmful samples and can undermine the security barrier with just a few malicious samples. This proves that a small amount of data can significantly alter the rejection behavior of the security response model. It also indicates that the subsequent fine-tuning itself has the risk of covering the security capabilities [1]. However, the data itself is obviously semantically harmful and is easily blocked by content review. Based on this, the representative method Virus [3] designs an evasion attack targeting the review mechanism of fine-tuning data. This method uses dual-objective optimization to reduce the penalty of the barrier review model for the attack data while maintaining the cosine similarity between the attack gradient and the original harmful gradient, so that the release rate of the review model for the attack data reaches 100%. However, it requires white-box access and dual-model gradients, and there is no report on cross-model transferability. Recent studies have gone further by using semantically completely benign

DOI: 10.54254/2755-2721/2027.38929

abnormal samples to induce safe degradation. This type of method selects 100 outlier samples from the pure benign dataset based on the self-influence score. All three barrier models classify them as safe, while the harmful score on the 1-5 scale ranges from 3.47 to 3.71, which is at the same level as 3.55 for pure malicious data. Only a 1% poisoning rate is sufficient and it can be transferred across seven different models [4]. Overall, the characteristics of attacks have gradually shifted from being explicitly harmful to becoming more covert and benign. This change has fundamentally undermined the defense premise that relies on content review.

2.2. Alignment stage pre-strengthening

The basic idea of alignment stage reinforcement is to enhance the model's robustness to subsequent fine-tuning perturbations during the secure alignment training; this approach only modifies the one-time alignment training and does not incur additional costs during the fine-tuning stage. Regarding the reinforcement content, the existing work has formed three typical approaches: the first is perturbation-aware training, which enables the model to anticipate attacks through adversarial training that simulates micro-tuning perturbations, similar to pre-vaccination; the second is the strengthening of security-related parameters, identifying the security parameters accounting for less than 1% [5] in the model and making targeted adjustments to them; the third is the flattening of the loss landscape, reducing the model's sensitivity to gradient or parameter perturbations in harmful directions. The representative work, Booster, follows the third approach to further advance the diagnosis: it discovers that the disturbance caused by taking one step along the harmful gradient is the root cause of the vulnerability. Therefore, it adds a regularization term to the alignment objective, minimizing the decrease in harmful loss before and after one step of harmful fine-tuning. This is equivalent to flattening the loss landscape in the harmful direction, allowing the model to learn more slowly in the face of future attacks. Under the mild settings of rank 32 and learning rate 1e-5, the average harmful ratio of the 7B open-source model decreased from 33.58% to 10.94%. In cross-model comparisons, it decreased by up to 34.14 percentage points. However, the cost was that the alignment training time was approximately three times that of the original [5]. The differences in the same position among different jobs mainly lie in the objects being reinforced, which could be data perturbation, specific parameters, or the loss landscape.

2.3. Fine-tuning process constraints

The basic idea of the constraint in the fine-tuning process is to directly restrict the parameter updates that may compromise security during the user's fine-tuning, thus blocking the risks during the training process. Regarding how to constrain updates, existing work has developed several typical methods: The first is secure data blending, where secure data is alternately trained with user data, or a small amount of secure samples are mixed into the data to continuously reinforce security-related behaviors during updates; the second is gradient projection, which projects parameter updates onto a secure subspace to prevent the gradient from accumulating in harmful directions; the third is sample weighting, which soft-weights the gradients based on the suspiciousness of the samples to reduce the influence of suspicious samples; in addition, there are forms such as implanting trigger backdoors and adding security regularizations to the training objectives. These methods are relatively simple in engineering terms, but the common prerequisite is the ability to identify which samples or gradients are harmful, and the mixing and protection ratios are mostly based on empirical values. The integrated approach of this route, named Antibody, combines the above ideas into three components: in the alignment stage, the model is optimized to the flat region of the harmful loss landscape; in the

DOI: 10.54254/2755-2721/2027.38929

fine-tuning stage, the likelihood ratio of the general rejection response is soft-weighted with batch samples to reduce suspicious gradients; and finally, the rejection bottom layer is trained on the perturbed model. The ablation study revealed that the three components reduced the harmful proportion gradually to 1.24%, with an average across architectures being 0.92%, and remained consistently below 10% within 100 training epochs [6]. The author also pointed out that the alternating training method will fail under a high learning rate, while the computational cost of the two-stage design will recur for each fine-tuning request.

2.4. Fine-tuned and repaired

The basic idea of the automatic and secure repair is that after the user has completed the normal fine-tuning, the security degradation that has occurred will be detected and repaired; since the intervention takes place after the training is completed, the defense effect is largely decoupled from the training hyperparameters chosen by the attacker. Around the repair target, the existing work has formed several typical methods: the first is parameter pruning, which identifies the parameters carrying dangerous updates and removes them all at once; the second is safe incremental filtering, which screens the incremental parameters generated by fine-tuning based on the trade-off between utility and safety one by one; the third is compensation vector, which superimposes a compensation vector in the safe direction on the filtered model to counteract the residual degradation; in addition, there are also routes such as re-alignment through adaptive perturbation vectors and detection in the inference stage without modifying the weights. Two representative works correspond to the ideas of parameter pruning and incremental filtering with compensation. Antidote's parameters: Calculate the WANDA importance score on a purely harmful re-aligned set, identify and remove the dangerous parameters related to safety; Under the harsh setting of a learning rate of 1e-4 and 40 epochs, the aforementioned alternating training method collectively failed. When the learning rate was 1e-3, the proportion of harmfulness increased to 77.3%, while its performance remained stable, with a total process time cost of only 1.02 times; However, the absolute proportion of harmfulness after pruning still remained above 60%, which is a partial alleviation. The pruning rate also needs to be adjusted according to the task [7]. Safe Delta is more in line with the actual situation of the FTaaS project and is oriented towards fine-tuning increments: caching the second-order information of security losses, greedily selecting each incremental parameter based on the utility-security ratio, then superimposing the security compensation vector to offset the residual losses. Even with 10,000 attack data, the attack success rate is still approximately 4.6%, and the utility gain brought by benign fine-tuning is almost undiminished. Each request only increases the processing time by approximately 62 seconds [8].

3. Analysis of difficulties, research gaps and prospects

The first challenge is that the defense mechanism is sensitive to the training hyperparameters, and these hyperparameters are controlled by the users. In order to achieve higher downstream accuracy, users tend to choose larger learning rates and more training epochs. This precisely falls within the range where the aforementioned methods, such as alternate training in experiments, have collectively failed; even the most robust Antibody, as pointed out by the authors, also failed at a learning rate of 1e-4 [6, 7]. Once the service provider opens the fine-tuning interface, they lose the ability to restrict the training configuration of the other party. The asymmetry of control power is a structural difficulty. The second difficulty is manifested as follows: When all the attack samples are classified as benign by the entire system, the method in Section 2.3, which is based on weighted

DOI: 10.54254/2755-2721/2027.38929

suspiciousness of the samples and incorporates safety data, loses its well-defined identification target [4]. The harmfulness increasingly manifests itself as the position of the data in the representation space rather than the semantic labels. The third difficulty lies in the fact that the balance knob between security and utility needs to be adjusted on a case-by-case basis: an increase in pruning rate enhances security but significantly reduces the accuracy of the task; the optimal value of regularization strength varies across different tasks [5, 7]; and aligning once requires serving an unknown task, which contradicts the idea of adjusting participation in a one-time deployment on a case-by-case basis in engineering.

In conclusion, characterizing similarity only requires training on the first 5 to 10 GPU minutes to be calculated [2]. However, none of the routes have used it for representing representative defenses, and residual risks have never been systematically measured; the root cause work itself can only help select models, but cannot repair models; and the most direct reference object has three contradictions: the security behavior does not have a definite target text like fact editing, the routing logic is contrary to the default interception requirements of the barrier, and its semantic locality assumption may also be actively broken by attackers [4, 9]. Therefore, this paper argues that making defense a function of similarity priors is an unexplored connection between the prediction end and the repair end. The minimum verification can first construct a defense × similarity probe matrix: on the high- and low similarity alignment models, respectively, superimpose the fine-tuning period and the post-defense period, and use multiple random seeds to compare the residual harmful proportions. If the defense consistently deteriorates under high similarity conditions, the gap is empirically obtained, and similarity should also become a standard control variable for evaluation; subsequently, an attempt can be made to use multiple rewrites with the same harmful intent that must yield the same negative response to replace the target text, and complete the security transformation of the representation intervention.

From the perspective of domain transfer, this issue may become more prominent in new scenarios. An external transformation without gradient distribution for multimodal models can significantly increase the uncertainty of the model in distinguishing malicious intentions through simple means such as mixing inputs and character perturbations, and it remains effective for the latest generation of commercial closed-source models [10]. This means that when the defense designed for single-round text within a distribution is transferred to a multimodal, long-inference scenario, the mechanism where overlapping representations lead to the overwriting of the security direction may not disappear, but the representation space itself becomes more complex, which is worthy of being the direction for further examination.

4. Conclusion

This paper takes defense as the main thread and similarity as the clue, unifying the evolution of attacks with the defense methods distributed according to the intervention location. The analysis indicates that the end result of the attack is complete benignization, which renders the signal recognition-based defense ineffective and creates clearly defined failure zones for all defense routes; more importantly, higher representation quality and similarity can directly serve as guardrail standards, but the existing evidence only supports correlation, and defense actions require a causal basis. Based on this, this article restructures the question of which defense is better into the issue of which defense to deploy under what level of similarity and attack intensity, and provides the minimum validation path from excluding confounding factors, controlled intervention to counterfactual consistency testing. If the subsequent work can enable the defense to adapt to the

DOI: 10.54254/2755-2721/2027.38929

similarity prior and complete the causal transfer of the representation intervention to the secure scenario, then the pre-control of risks for the fine-tuning service can form a closed loop.

References

[1] Qi, X., Zeng, Y., Xie, T., Chen, P. Y., Jia, R., Mittal, P., & Henderson, P. (2023). Fine-tuning aligned language models compromises safety, even when users do not intend to!, 2023. https://doi.org/10.48550/arXiv.2310.03693

[2] Hsiung, L., Pang, T., Tang, Y. C., Song, L., Ho, T. Y., Chen, P. Y., & Yang, Y. (2026, July). Why LLM safety guardrails collapse after fine-tuning: A similarity analysis between alignment and fine-tuning datasets. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 16601-16619). https://doi.org/10.18653/v1/2026.acl-long.756

[3] Huang, T., Hu, S., Ilhan, F., Tekin, S. F., & Liu, L. (2025). Virus: Harmful fine-tuning attack for large language models bypassing guardrail moderation. arXiv preprint arXiv: 2501.17433.https: //doi.org/10.48550/arXiv.2501.17433

[4] Guan, Z., Hu, M., Zhu, R., Li, S., & Vullikanti, A. (2025). Benign samples matter! Fine-tuning on outlier benign samples severely breaks safety. arXiv preprint arXiv: 2505.06843.https: //doi.org/10.48550/arXiv.2505.06843

[5] Huang, T., Hu, S., Ilhan, F., Tekin, S., & Liu, L. (2025, May). Booster: Tackling harmful fine-tuning for large language models via attenuating harmful perturbation. In *International Conference on Learning Representations (Volume 2025)*, 67202–67226. https://doi.org/10.48550/arXiv.2409.01586

[6] Nguyen, Q., Le, T., Wu, J., Bui, A., & Harandi, M. (2026, April). Antibody: Strengthening defense against harmful fine-tuning for large language models via attenuating harmful gradient influence. In *International Conference on Learning Representations (Volume 2026)*, 121713–121741. https://doi.org/10.48550/arXiv.2603.00498

[7] Huang, T., Bhattacharya, G., Joshi, P., Kimball, J., & Liu, L. (2024). Antidote: Post-fine-tuning safety alignment for large language models against harmful fine-tuning. *arXiv preprint arXiv:2408.09600*. https://doi.org/10.48550/arXiv.2408.09600

[8] Lu, N., Liu, S., Wu, J., Chen, W., Zhang, Z., Ong, Y.-S., Wang, Q., & Tang, K. (2025). Safe delta: Consistently preserving safety when fine-tuning LLMs on diverse datasets. *International Conference on Machine Learning (ICML)*. https://doi.org/10.48550/arXiv.2505.12038

[9] Liu, X., Chen, S., Dong, X., Liu, Y., Zhao, X., Wang, H., Yan, Y., Chen, H., & Chen, Z. (2026). Representation interventions enable lifelong knowledge memory control in LLMs. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 5414–5436. Association for Computational Linguistics. https://doi.org/10.18653/v1/2026.acl-long.246

[10] Jeong, J., Bae, S., Jung, Y., Hwang, J., & Yang, E. (2025). Playing the fool: Jailbreaking LLMs and multimodal LLMs with out-of-distribution strategy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 29937–29946. IEEE. https: //doi.org/10.1109/CVPR52734.2025.02786