Proceedings of the 7th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/2027.38809

Safety Degradation and Preservation During Benign Business Fine-Tuning

Zekai Lin

School of Mathematics (Zhuhai), Sun Yat-sen University, Zhuhai, China
kaizelin14@gmail.com

Abstract. While most Large Language Models (LLMs) are designed to be safe, they often need business fine-tuning, which may change the model's inherent safety behaviors and weaken the safety guardrails already put in place, something that can't be fully managed just by tweaking the training data. Thus, it becomes an important problem to preserve the safety properties of a model during the fine-tuning process. This paper summarizes approaches of how a model can maintain safety capabilities while learning new tasks, highlighting the interactions between training data, changes in the internal model, protective measures and final evaluation. It reviews the research literature along multiple dimensions of factors that exacerbate safety, internal model change, preservation of safety and evaluation approaches, and it highlights specific research gaps in the context of the available literature. Existing methods have defined specific protective measures for pre-, fine-tuning and post-training stages, have studied the mechanisms of safety degradation, and have made progress in research and validation from the attack and defense sides. Due to the variety and idealized conditions of experiments, however, additional research is required to account for the applicability of the experiments, the prediction of the degradation phenomena in safety, and safety maintenance in multiple adapter combinations.

Keywords: Large language models, safety alignment, downstream fine-tuning, safety degradation, safety preservation

1. Introduction

Large language models (LLMs) are applied to many tasks including question answering, retrieval-augmented generation, and code generation. These models are usually trained to be safe before they are released to the public, which means they can be used to identify potential threats and refuse to comply with inappropriate requests to reduce the likelihood of producing illegal, harmful, or real-world execution-related content. But after deployment of a base model for the specific business application, it may be necessary to fine-tune it again with domain-specific data. This process improves the business capabilities of the model, but can break the safety guardrails that were initially put in place. Therefore, how to enhance the safety of models without compromising their versatility to meet business requirements is a pressing issue in fine-tuning large language models.

It is common for safety to be degraded during the business fine-tuning process, even when fine-tuning for normal operations can cause a model to lose its ability to identify and reject dangerous

© 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: 10.54254/2755-2721/2027.38809

requests [1]. Further investigation of the factors that affect this has indicated that training data, safety alignment styles, parameter updates, and model states all have an impact on the model's safety guardrails. Existing protection methods are based on these mechanisms and aim for safety preservation in three stages: model training, fine-tuning, and post-deployment, considering different levels of user authorization.

The current approaches involve various phases, such as data processing, training and deployment. Definitions of safety, however, differ from study to study, and a single refusal rate does not reflect improvements in safety, false refusal of legitimate requests, and business capabilities. Moreover, factors like deployment environments, training languages, task types and evaluation frameworks affect the transferability of these methods. Therefore, it is crucial to systematically review and compare existing work, so as to clarify the relationship between different factors that can decrease safety and the protection measures. This paper takes a look at these aspects (including the factors that affect safety degradation, changes in the internal model, methods for preserving safety, and evaluation criteria) and attempts to identify some areas of limitations in existing research and how it may progress in the future.

2. Safety degradation and safety preservation during business fine-tuning

2.1. Safety degradation phenomenon

Safety degradation is the phenomenon whereby a model that has been aligned for safety, may lose some of its safety capabilities after further business tuning and increase in unsafe responses or change of safety boundaries. However, even if there is no adversarial intent in the fine-tuning data, training on common tasks may compromise a model's ability to reject harmful requests [1]. Such studies prove that benign business fine-tuning has the potential to degrade the safety properties of a model, and the extent to which such properties can be degraded depends on factors like the type of task and training sample variations. Therefore, additional research is needed to explore the factors that lead to safety degradation and how these changes develop in the model.

2.2. Factors influencing changes in safety

Initial safety training and training data affect changes in safety during fine-tuning. Hsiung et al. studied the correlation between the data that is used for safety training and the data that is used for business fine-tuning. They discovered that the greater the similarity of the internal representations of the two sets of data, the more likely subsequent fine-tuning was to affect safety – and the more dissimilar the safety training data was, the better the original safety behaviours were likely to be preserved [2]. This means that modifications in protection are subject to the business data itself, as well as the knowledge the model has acquired before fine-tuning.

The training signals during fine-tuning may change due to task type, response style and training language. The training signals from different tasks are different. Summarization, coding, translation, and classification, for instance, vary in their input structure, output format, and loss distribution, meaning that not all benign fine-tuning can be regarded as a similar process. Vishnubhotla et al. recommend choosing fine-tuning tasks for specific capability objectives to reduce arbitrariness in task selection [3]. Moreover, Xiao et al. discovered that fine-tuning a model for a particular common response style affects the response to harmful requests of that style, which they examined with 36 models [4]. This suggests that format, tone, and organizational structure can be mediators for the effect of business training on safety. Similarly, combinations of training and testing languages can

DOI: 10.54254/2755-2721/2027.38809

have an impact on safety performance; testing in Chinese following English training cannot be considered the same as testing in English following Chinese training.

Changes in safety are also dependent on the combination of models, adapters and components. With adapters like LoRA, one base model can be used with many business capabilities. The final state of the model can be changed by all the factors of the model family, the scale of the parameters, the structure of the adapter and how multiple components are combined. It has been found that, although each adapter could be used independently without any problem, if two or more colluding LoRA adapters are activated simultaneously in a specific combination, the model's resistance to malicious requests decreases [5]. As such, the safety of a deployed model in the real world can only be deduced from both a single fine-tuned model and the behaviour of the base model when used with more than one adapter.

2.3. Internal explanation of safety degradation

Hidden representations and parameter updates are studied to understand the internal changes in the model before and after fine-tuning. Malla et al. found that changes in parameters from post-hoc safety training were almost orthogonal to directions of capability, indicating that post-hoc safety training is more like a masking mechanism than removing harmful capabilities [6]. Instead, Wang et al. considered the training process and found that when training benign, parameters accumulate drift along "dangerous directions" [7]. The former highlights the static structures after alignment, and the latter highlights the dynamic trajectories in fine-tuning. When combined, these studies provide insight into the changes of internal model representations of safety during safety training and business fine-tuning, and ultimately, an intuitive explanation, based on internal models, for the degradation of model safety. SPF studies the changes in gradients at the time of parameter updates and found that while the gradient in the direction of safety objectives exists in only a few directions, the gradient in the direction of task objectives is distributed widely, which causes the possibility of conflict between the direction of safety objectives and that of task objectives [8]. Accordingly, SPF considers safety preservation as controlling the direction of the updates and changes the gradient of the tasks by projection, so as to reduce the part that conflicts with the safety requirements.

Another type of internal interpretation is "safety routing" and local sensitivity. Guo et al. examined a range of 7B–8B models that were either supervised fine-tuned or DPO aligned, and they discovered that the part of the model that produces refusal outputs is referred to as "safety routing" and examined the Fisher-related metrics within the output-side MLP [9]. Fisher measure is a measure of the sensitivity of the output probabilities to variations in the parameters. The authors noticed that if the fine-tuning is benign, the local structures become sharp again; although the gradient conflicts affect the degradation, they do not cause it [9]. This approach complements gradient analysis: whereas gradient analysis concentrates on the direction of updates, routing analysis concentrates on the sensitivity of refusal structures to updates.

Other studies have tackled the phenomenon of safety degradation from the geometric point of view. Hossain et al. explored dedicated safety models (LlamaGuard, WildGuard, Granite Guardian) and analyzed their ability to maintain safety and unsafe samples separable in the representation space using CKA, Fisher correlation metrics, and inter-class distances [10]. The authors suggested FW-SSR to prevent safety discrimination representation directions from such changes.

DOI: 10.54254/2755-2721/2027.38809

2.4. Safety preservation methods

After confirming the degradation phenomenon and obtaining an internal explanation, researchers began incorporating safeguards at various stages of the model's lifecycle. Based on the timing of the intervention, these safeguards can be categorized into three types: pre-fine-tuning, during fine-tuning, and post-training.

Pre-fine-tuning prevention: Original model providers can build robustness into the model before delivering it to business teams. On one hand, the selection of upstream safety data can be adjusted based on the representational relationship between safety alignment data and data for downstream tasks [2]. On the other hand, potential perturbations or harmful updates likely to be encountered later can be simulated during safety training, enabling the model to learn in advance how to maintain safe behavior. Since such methods require modifying the original model's safety training process, business teams generally cannot independently implement equivalent protections after the model has been delivered.

Protection during fine-tuning: Business training teams can directly control training data, loss calculations, or parameter updates. LARF filters samples using the hidden representations of safety-critical layers [11]; SQSD scores samples based on the difference in projections of parameter updates onto "unsafe" versus "safe" directions—a metric the authors report is transferable across architectures, parameter scales, and parameter-efficient fine-tuning methods [7]. SafeStyle supplements the dataset with safe samples that match the business response style [4], while SPF mitigates conflicts between safety and task objectives by constraining the update direction [8]. Goel et al. continuously estimate safety risks during training—employing both a judge-model-based safety evaluator and an activation-based risk predictor—to impose stricter constraints on high-risk updates, steering them closer to a safe reference policy [12]. Meanwhile, Entesari and Fazlyab formulate safety preservation as a chance-constrained problem: they limit the proportion of safe samples that degrade beyond a certain threshold relative to a reference model and apply a closed-form gradient correction to keep each update within the feasible set, effectively reverting to standard supervised fine-tuning when the update already satisfies the constraints [13].

Post-training repair: Even after model training is complete, the deployment team can still restore safety at the weight or adapter level [14]. Antidote prunes weights associated with harmful generation [14]. NLSR focuses on safety-related neurons [15]. Panacea applies repair perturbations to the trained model. Model merging methods attempt to balance the states of the task model and the safety model. NeWTral pre-learns weight transformations between safe and unsafe adapters, directly repairing the input adapter at deployment time.

Table 1. Representative safety preservation methods and implementation conditions

Stage	Representative method	Key operations	Permissions and References
Pre-fine-tuning	Selection of safety data with low similarity, selection of post-training methods, and safety co-training during the pre-training phase.	Select safe training data and post-training methods, or continuously incorporate safety signals during pre-training.	Safety training data and original model training privileges [2]

DOI: 10.54254/2755-2721/2027.38809

Table 1. (continued)

Business fine-tuning	LARF, SQSD	Sample risk scoring and screening	Candidate training sets, reference data, internal representations, or parameter updates [7, 11]
	SPF, SafeStyle, adaptive regularization, chance-constrained fine-tuning	Constraining gradients, supplementing with style-matched safe samples, adjusting constraint strength based on risk, or limiting tail degradation.	Control of the training process and safety information [4, 8, 12, 13]
Post-training and Adapter Deployment	Post-hoc repair methods compiled by NeWTral and SafeTune	Convert adapter or fix weights	Model or adapter weight access, paired adapters, or safety references [14, 15]

As shown in Table 1, the three categories of methods operate on training inputs, parameter updates, and the final model state, respectively. Comparisons must account for the target of the operation, required permissions, additional safety data, computational costs, and the impact on business utility.

2.5. Changes in evaluation under additional conditions

Conclusions regarding safety preservation methods must be validated under real-world usage conditions. Malicious requests, benign requests, and business tasks should be evaluated separately; generative assistants and safety classifiers should employ metrics that correspond to one another without being conflated. Multilingual evaluation requires the separate recording of training and testing languages; studies on stylistic variations necessitate altering the mode of expression while keeping the malicious intent constant or comparable [4]; and studies on composition require the separate testing of individual adapters and various adapter combinations [5].

Also, the number of training epochs, sampling parameters, random seeds, rules of evaluation, and choice of evaluators can influence the safety scores. Vishnubhotla et al. also indicated that the results of fine-tuning depend on the safety metrics and judge models used; they also noted that models can be unreliable when faced with incoherent outputs [3]; on his part, Entesari and Fazlyab emphasized that the gains in safety, when found, may come at the cost of a few major mistakes [13]. So, a comprehensive report should include the test prompts, the responses produced, the criteria for judging the responses, and the randomness settings, and provide metrics on harmful response rates, false rejection rates for benign responses, and business task performance, supplemented where needed by "tail" metrics. The benefits of a particular protection method can be judged only after comparing the methods under the same conditions, such as base model, task, language, size of data, and budget of computation.

DOI: 10.54254/2755-2721/2027.38809

2.6. Attack methods and combined risks

The research on attack methodologies is still growing in addition to benign fine-tuning. There is malicious fine-tuning, where the training process is intentionally designed to misalign safety; in addition, "Colluding LoRA" leverages the synergies between different adapters, with each one seemingly harmless but combining them to pose risks [1, 5]. The study of model merging and adapter composition also shows that the deployed state may be different from any of the training states [5].

Attacks are useful for rigorously testing safety preservation methods by stressing them, and vice versa—safety preservation methods should clearly specify what they are protecting, what the threat is, and their boundaries of operation. It is important to differentiate between research on benign fine-tuning (the understanding of why safety degradation occurs under no malicious intent) and research on attacks (how vulnerabilities can be attacked or amplified by an attacker). Both fields have metrics like ASR, ORR and Business Utility.

3. Current limitations and future prospects

In general, current studies have given many tools and explanations to understand and reduce safety guardrail failures; however, the conclusions are dependent on the conditions of the studies and the evaluation criteria. Data filtering and risk scoring methods have been found to some extent to provide quality control of reference data, and have shown some potential for cross-model transferability, but the specific composition of data, model structure and business goals may impact the balance between safety and task performance. Likewise, explanations about degradation mechanisms vary from focusing on conflicts in the direction of the gradient to focusing on the geometry of the training procedures to focusing on changes in the internal implementation of the refusal behavior as a result of post-training techniques and model architectures. Many of these explanations could hold true in different contexts; they do not exclude each other, but should be compared with the same models, training pipelines, and measures of success. In addition, current findings must be evaluated with caution, as there are variations in base models, adaptive attacks against mitigation tools, and differences in safety styles and models for evaluation. In the future, it may be necessary to combine these individual pieces of knowledge in one analytical model to deepen the understanding of the connection between the different mechanisms, methods and usage conditions.

Based on this foundation, future research can be continued to solve three problems that are closely related to each other. In the first instance, it will be necessary to discuss the synergistic or limiting interactions and associations that can occur when several degradation mechanisms are combined together and that, e.g., seemingly different degradation mechanisms can have a common origin and suggest a specific sequence and combination of remediation methods. Second, instead of only assessing for post-failure, it is important to explore if it is possible to use signals - representations, gradients, curvature, and behavioral changes - to detect potential degradation of a safety guardrail in advance, to provide timely risk alerts during training and deployment. Third, several "normal" adapters will be deployed simultaneously and knowledge and control over interactions among these adapters will play a key role in the reliability of the actual system. The potential of such a unified benchmark across a variety of models, tasks and linguistic contexts, with the incorporation of mechanism analysis, risk monitoring and reversible remediation, is to move the field from validation of individual methods to a more holistic system-level approach, with increased transferability, interpretability and composability.

DOI: 10.54254/2755-2721/2027.38809

4. Conclusion

This paper discusses the problems of safety degradation and preservation in the fine-tuning process of models for specific business applications, and introduces the influencing factors, the protection method and the evaluation criteria. Previous studies suggest that the benign nature of business data does not imply maintained safety alignment; instead, prior safety training, specific task and response style in the training, parameter changes, as well as adapter combinations all affect the safety alignment. Other explanations like "parameter drift", "gradient conflict", and "sensitivity in safety routing" provide other viewpoints, but have not been integrated into a universal mechanism. Intervention strategies include pre-fine-tuning prevention, in-process protection, and post-training remediation, depending on the permissions for implementation, data and computation costs, and business utility constraints. In the meantime, the assessment system is still disjointed, with rejection rates as the only measure and with no overarching, unified assessment standards. Current evidence suggests that protection is feasible, but the effectiveness of the different protection strategies in different models, tasks and languages remains to be validated. Future work should establish a tighter relationship between degradation mechanisms and the outcome of the interventions, further work on early detection of degradation risk and early preservation of safety in multi-adapter configuration, and embed safety evaluation in the fine-tuning and deployment process in order to give verifiable safety assurances and add business capabilities.

References

[1] Qi, X., Zeng, Y., Xie, T., Chen, P.-Y., Jia, R., Mittal, P., & Henderson, P. (2024). Fine-tuning aligned language models compromises safety, even when users do not intend to! In *The Twelfth International Conference on Learning Representations (ICLR 2024)*. arXiv:2310.03693.

[2] Hsiung, L., Pang, T., Tang, Y.-C., Song, L., Ho, T.-Y., Chen, P.-Y., & Yang, Y. (2026). 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). Association for Computational Linguistics. DOI:10.18653/v1/2026.acl-long.756. arXiv:2506.05346. ACL Anthology

[3] Vishnubhotla, K., Dawkins, H., Nejadgholi, I., & Kiritchenko, S. (2026). Safety measurements for fine-tuned LLMs should be grounded in capability. arXiv:2606.03648.

[4] Xiao, Y., Tonekaboni, S., Gerych, W., Suriyakumar, V., & Ghassemi, M. (2026). When style breaks safety: Defending LLMs against superficial style alignment. In *The Fourteenth International Conference on Learning Representations (ICLR 2026)*. arXiv:2506.07452.

[5] Ding, S. (2026). Colluding LoRA: A compositional vulnerability in LLM safety alignment. arXiv:2603.12681.

[6] Malla, S., Choi, C., & Choi, J. H. (2026). The geometry of refusal: Why post-hoc safety is fragile and pretraining-time safety persists. arXiv:2609.06934.

[7] Wang, X., Zhang, Y., Liu, Y., Yang, X., Wang, Z., Feng, S., & Wang, D. (2026). From parameter dynamics to risk scoring: Quantifying sample-level safety degradation in LLM fine-tuning. In *Proceedings of the 43rd International Conference on Machine Learning (ICML 2026)*. arXiv:2605.04572.

[8] Zhang, J., Hu, Y., Chen, K., He, L., Ma, J., Lou, J., Li, D., Liu, J., Yang, X., & Jia, R. (2026). Understanding and preserving safety in fine-tuned LLMs. arXiv:2601.10141.

[9] Guo, Y., Chen, X., Zhang, S., Wang, X., & Tang, H. (2026). When safety routing breaks: Understanding alignment fragility under benign fine-tuning. arXiv:2609.01455.

[10] Hossain, I., Puppala, S., Ferdaus, J., Alam, M. J., Ahad, T., Lee, Y., Alam, S. B., & Talukder, S. (2026). When safety geometry collapses: Fine-tuning vulnerabilities in agentic guard models. *Proceedings of the AAAI Symposium Series, 9*(1), 255–263. DOI:10.1609/aaaiss.v9i1.42939. arXiv:2605.02914.

[11] Li, H., Li, L., Lu, Z., Wei, X., Li, R., Shao, J., & Sha, L. (2025). Layer-aware representation filtering: Purifying finetuning data to preserve LLM safety alignment. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing* (pp. 8030–8050). Association for Computational Linguistics. DOI:10.18653/v1/2025.emnlp-main.406. arXiv:2507.18631. ACL Anthology

DOI: 10.54254/2755-2721/2027.38809

[12] Goel, J., Maji, S., & Mazumder, P. (2026). Learning to stay safe: Adaptive regularization against safety degradation during fine-tuning. arXiv:2602.17546.
[13] Entesari, T., & Fazlyab, M. (2026). Beyond average safety: Chance-constrained LLM fine-tuning. arXiv:2609.29960.
[14] Huang, T., Bhattacharya, G., Joshi, P., Kimball, J., & Liu, L. (2025). Antidote: Post-fine-tuning safety alignment for large language models against harmful fine-tuning attack. In *Proceedings of the 42nd International Conference on Machine Learning (Vol. 267)* (pp. 25059–25074). Proceedings of Machine Learning Research. arXiv:2408.09600.
[15] Yi, X., Zheng, S., Wang, L., de Melo, G., Wang, X., & He, L. (2025). NLSR: Neuron-level safety realignment of large language models against harmful fine-tuning. *Proceedings of the AAAI Conference on Artificial Intelligence, 39*(24), 25706–25714. DOI:10.1609/aaai.v39i24.34762.