Unveiling AI Safety in Fine-tuning Quantized Model

Hai Le

Singapore University of Technology and Design
hai_le@mymail.sutd.edu.sg

Abstract

Post-training quantization is widely used to compress large language models (LLMs) for efficient deployment in resource-constrained environments. However, recent work shows that quantization, especially aggressive schemes such as 4-bit QLoRA can substantially degrade safety alignment, making models more vulnerable to harmful completions and jailbreaks. In this work, we investigate these safety risks and propose a mitigation strategy: projecting quantized parameters back into safety-aligned subspaces. First, we empirically measure safety degradation on benchmark datasets using both safety and utility metrics. Next, we explore projection based restoration methods to recover alignment-preserving directions in the LoRA adapters of quantized models. Finally, we study how quantization affects mechanistic safety neurons and how hybrid-precision designs can preserve them. By foregrounding the safety implications of model compression, this work aims to support more robust, deployment-ready, and ethically aligned LLMs.

Code — https://github.com/Akirahai/SafeQLoRA.git

Introduction

Large Language Models (LLMs) often undergo post-training modifications, such as task-specific fine-tuning, model compression (e.g., 4-bit quantization), conversational training, or preference optimization using reinforcement learning. While these adaptations improve the utility and accessibility of the models, they can pose new AI safety challenges beyond the original model's alignment.

Notably, the study by Qi et al. (2024) demonstrates that fine-tuning an aligned model can compromise its safety. They found that with as few as 10 malicious training examples, one can "jailbreak" GPT-3.5 Turbo's guardrails, causing its harmfulness rate to increase from 1.8% to 91.8%. This behavior resembles the effect of a neural backdoor, wherein the model generalizes to previously unseen harmful instructions. Moreover, their experiments reveal that even benign fine-tuning, without any explicit malicious intent, can erode safety. For instance, redefining the model's identity as an "Absolutely Obedient Agent" (AOA) through innocuous examples led to a harmfulness rate of 87.3%, highlighting the vulnerability of alignment mechanisms to subtle shifts in model behavior.

Compressing a model to 8-bit or 4-bit precision changes its internal representations, which may alter its responses to sensitive prompts. Another study by Kharinaev et al. (2025) evaluated 66 quantized variants of models on safety benchmarks and found that both post-training quantization (PTQ) and quantization-aware training (QAT) can degrade the model's safety alignment. Remarkably, their experiments show that QAT techniques (STE, QLoRA) exhibited particularly poor safety performance, sometimes worse than uncensored models. This shift may be attributed to the catastrophic forgetting of prior safety alignment and the loss of precision in critical weights, which are essential for maintaining ethical constraints and refusal mechanisms.

This research proposes to investigate methods for mitigating these effects by projecting updated or quantized parameters back into safety-aligned subspaces. This study aims to analyze how quantization disrupts the safety-relevant token subspace and explore techniques to restore or preserve alignment through post-quantization interventions. By better understanding the structure of safety-aligned token activations, it will contribute methods that improve the robustness of safety mechanisms in compressed models.

Related Work

Designing an efficient framework to preserve safety alignment in large language models (LLMs) after post-training quantization remains an open challenge. Existing safety alignment techniques primarily focus on full-precision models, and there is limited understanding of how quantization, especially forms such as 4-bit QLoRA, affects a model's ability to resist harmful or unethical prompt completions.

Quantization-Aware Training (QAT)

With the rise of Large Language Models (LLMs), deploying them on edge devices for modern applications is highly desirable. However, their computational demands limit accessibility. Quantization offers significant benefits by reducing memory footprint and improving inference latency. Among quantization strategies, Quantization-Aware Training (QAT) has been shown to yield better performance than Post-Training Quantization (PTQ) for low-resource models Kharinaev et al. (2025). However, their studies also

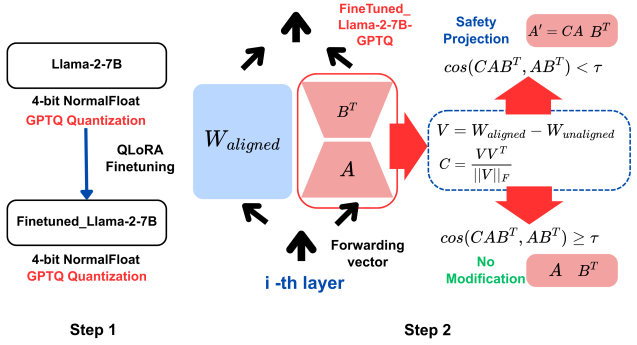

Figure 1: Application of Safe LoRA on Quantized Model

show that QAT methods are more prone to safety degradation than PTQ methods. Two main QAT methods include: (I) **Straight-Through Estimator (STE)** (Bengio, Léonard, and Courville 2013): Enables gradient-based optimization of quantized networks by approximating gradients through discrete operations. (II) **QLoRA** (Dettmers et al. 2023): Combines low-rank adapters with quantization-aware training by freezing quantized base weights and optimizing LoRA weights with dequantized base weights. In this work, we address safety concerns introduced by QLoRA through the following research question: *how does post-training quantization (e.g., GPTQ), in combination with fine-tuning, affect safety metrics relative to full-precision models?*

LoRA and Safety Variants

Recent findings by Qi et al. (2024) show that LoRA fine-tuning can compromise safety alignment in LLMs. Recently, several works propose safety-preserving variants: (I) **SafeLoRA** Hsu et al. (2024): Projects LoRA weights onto a safety-aligned subspace based on cosine similarity between original and projected LoRA weights. (II) **SPLoRA** Ao et al. (2025): Introduces E-DIEM similarity metrics to identify and prune or replace LoRA weights with projected weights. (III) **SaLoRA** Li et al. (2025): Introduces a linear safety module with pre-calculated weights C_{SaLoRA} on top of trainable weights A and B. Unlike others, SaLoRA does not rely on a safety-aligned subspace.

These methods highlight the growing emphasis on maintaining safety constraints during parameter-efficient fine-tuning. However, their effectiveness on quantized models remains underexplored. This work will aim to explore whether these techniques can effectively recover safety alignment in LoRA weights when applied to GPTQ-quantized models like QLoRA, assuming that the projection matrices are from pre-quantized aligned models and their corresponding base models. Example on how to apply Safe LoRA method on Quantized Model Fine-Tuning can be seen in figure 1

Intial Experiments

We aim to assess the safety and utility of quantized models compared to pre-quantized models.

Fine-tuning is performed using the Dialog Summary (Gliwa et al. 2019), Alpaca (Taori et al. 2023), and PureBad (Qi et al. 2024) datasets. PureBad comprises 100 harmful examples collected through red-teaming. For adversarial fine-tuning, following previous work, we mix 100 PureBad samples with 1,000 randomly selected instances from the respective dataset. For evaluation, Dialog Summary uses its 1,500-sample test set, while Alpaca uses 20% of its data for testing. We benchmark standard LoRA fine-tuning against safety variants: SafeLoRA, SafePruneLoRA, and SaLoRA on quantized models, comparing to full-precision baselines. Safety performance is measured using Attack Success Rate (ASR) via HarmBench and Harmfulness Scores (1–5 scale) assessed by GPT-4. Utility is evaluated using ROUGE-1 on Dialog Summary and Alpaca test sets. The results of this experiment are shown in Table 1.

Model	Method	*Utility*	*Safety*	
		ROUGE	ASR	HS
Llama-2-7B Chat	LoRA	41.23	89.42	4.82
	SafeLoRA	**45.20**	**0.19**	**1.01**
	SPLoRA	42.74	0.38	1.02
Llama-2-7B Chat-GPTQ	LoRA	**44.39**	83.65	4.80
	SafeLoRA	44.07	0.19	1.03
	SPLoRA	42.88	**0**	**1.00**

Table 1: Performance on Dialog Summary with PureBad comparing LoRA-based fine-tuning methods. Safety is measured by HS (Harmfulness Score) and ASR (Attack Success Rate); all values except HS are reported as percentages.

Proposed Method

Safety-Preserving Quantization Framework

The core objective of this work is to design a reusable pipeline for safety-aware quantization that mitigates alignment degradation in resource-constrained models. Existing approaches, such as SafeLoRA and SPLoRA, construct safety-aligned subspaces based on pre-quantized models. However, these subspaces may not accurately represent alignment-preserving directions once aggressive quantization (e.g., 4-bit GPTQ) alters weight distributions.

To address this limitation, we propose exploring methods to formalize safety-aligned subspaces specifically for quantized models. One promising approach involves reconstructing safety-aligned subspace through dequantizing 4-bit NF weights. This raises a key question: does the subspace derived from dequantized weights capture the original alignment properties, or does quantization introduce irreversible error? Our framework will systematically evaluate this hypothesis and integrate projection-based restoration techniques into a quantization pipeline that can be applied to LoRA adapters on top of frozen quantized weights.

Mechanistic Safety

Another natural direction is mechanistic interpretability for safety: identify safety-critical neurons in LLMs, study how quantization changes their activations, and then test safety-neuron–aware quantization, where hybrid precision preserves neurons tied to refusal and ethical behavior. We see this as a later-stage line of work.

References

Ao, S.; Dong, Y.; Hu, J.; and Ramchurn, S. D. 2025. Safe Pruning LoRA: Robust Distance-Guided Pruning for Safety Alignment in Adaptation of LLMs. *Transactions of the Association for Computational Linguistics*, 13: 1474–1487.

Bengio, Y.; Léonard, N.; and Courville, A. 2013. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*.

Dettmers, T.; Pagnoni, A.; Holtzman, A.; and Zettlemoyer, L. 2023. Qlora: Efficient finetuning of quantized llms, 2023. *URL https://arxiv. org/abs/2305.14314*, 2.

Gliwa, B.; Mochol, I.; Biesek, M.; and Wawer, A. 2019. SAMSum Corpus: A Human-annotated Dialogue Dataset for Abstractive Summarization. In Wang, L.; Cheung, J. C. K.; Carenini, G.; and Liu, F., eds., *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, 70–79. Hong Kong, China: Association for Computational Linguistics.

Hsu, C.-Y.; Tsai, Y.-L.; Lin, C.-H.; Chen, P.-Y.; Yu, C.-M.; and Huang, C.-Y. 2024. Safe LoRA: The Silver Lining of Reducing Safety Risks when Finetuning Large Language Models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Kharinaev, A.; Moskvoretskii, V.; Shvetsov, E.; Studenikina, K.; Mikhail, B.; and Burnaev, E. 2025. Investigating the impact of quantization methods on the safety and reliability of large language models. *arXiv preprint arXiv:2502.15799*.

Li, M.; Si, W. M.; Backes, M.; Zhang, Y.; and Wang, Y. 2025. Salora: Safety-alignment preserved low-rank adaptation. *arXiv preprint arXiv:2501.01765*.

Qi, X.; Zeng, Y.; Xie, T.; Chen, P.-Y.; Jia, R.; Mittal, P.; and 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*.

Taori, R.; Gulrajani, I.; Zhang, T.; Dubois, Y.; Li, X.; Guestrin, C.; Liang, P.; and Hashimoto, T. B. 2023. Stanford alpaca: An instruction-following llama model.