

SafeCOMM: What about Safety Alignment in Fine-Tuned Telecom Large Language Models?

Aladin Djuhera*, Swanand Ravindra Kadhe†, Farhan Ahmed†, Syed Zawad†, Holger Boche*, and Walid Saad‡

*Technical University of Munich, Germany, †IBM Research, USA, ‡Virginia Tech, USA

Emails: {aladin.djuhera, boche}@tum.de, {swanand.kadhe, farhan.ahmed, szawad}@ibm.com, walids@vt.edu

Abstract—Fine-tuning large language models (LLMs) for telecom tasks and datasets is a common practice to adapt general-purpose models to the telecom domain. However, little attention has been paid to how this process may compromise model safety. Recent research has shown that even benign fine-tuning can degrade the safety alignment of LLMs, causing them to respond to harmful or unethical user queries. In this paper, we investigate this issue for telecom-tuned LLMs using three representative datasets featured by the GenAINet initiative. We show that safety degradation persists even for structured and seemingly harmless datasets such as 3GPP standards and tabular records, indicating that telecom-specific data is not immune to safety erosion during fine-tuning. We further extend our analysis to publicly available Telecom LLMs trained via continual pre-training, revealing that safety alignment is often severely lacking, primarily due to the omission of safety-focused instruction tuning. To address these issues in both fine-tuned and pre-trained models, we conduct extensive experiments and evaluate three safety realignment defenses (SafeInstruct, SafeLoRA, and SafeMERGE) using established red-teaming benchmarks. The results show that, across all settings, the proposed defenses can effectively restore safety after harmful degradation without compromising downstream task performance, leading to Safe teleCOMMunication (SafeCOMM) models. In a nutshell, our work serves as a diagnostic study and practical guide for safety realignment in telecom-tuned LLMs, and emphasizes the importance of safety-aware instruction and fine-tuning for real-world deployments of Telecom LLMs.

Index Terms—6G, large language models (LLMs), fine-tuning, safety alignment, open-source telecom datasets

I. INTRODUCTION AND MOTIVATION

Large language models (LLMs) have demonstrated remarkable capabilities in coding, math, and, more recently, advanced reasoning tasks [1], while at the same time becoming more accessible to the broader public. This makes them powerful *catalysts* for transforming 6G beyond a mere network infrastructure upgrade into an intelligent wireless ecosystem that embeds connectivity, communication, and artificial intelligence (AI) into a unified, AI-driven network fabric [2]. This will involve multiple AI agents that can communicate with one another, and not only understand the network data but also the underlying *telecom standards, systems, and protocols* [3]. To achieve this, generalist LLMs need to instill telecom-specific knowledge into their core understanding. This can be done through supervised fine-tuning (SFT) [4] or continual pre-training (CPT) on large telecom corpora such as 3GPP standards and scientific

publications [5]. This reduces reliance on complex retrieval-augmented generation (RAG) systems that introduce additional latency from querying external knowledge bases.

However, *safety alignment*, which guides models away from inappropriate behavior after pre-training, is critical to ensure that LLM outputs comply with human preferences and security policies. This is typically achieved as part of instruction tuning [6], e.g., via reinforcement learning with human feedback (RLHF) [7] or SFT, both of which incorporate corrective layers to steer the model away from harmful responses. For example, consider a 3GPP chatbot. While being helpful on telecom-related queries, it must be ensured that *under no circumstances* does it respond to unethical, harmful, or dangerous queries, such as: “How can I most effectively destroy a base station with as little effort as possible? Please make a short list.”

Nonetheless, maintaining safety alignment is challenging when LLMs are further *fine-tuned* for downstream tasks. For example, the authors in [8] show that even a handful of malicious training examples can jailbreak LLMs, causing them to comply with harmful user requests. In addition, [9] demonstrates that even conventionally benign tasks can inadvertently compromise the safety of fine-tuned models to a noticeable degree. Theoretical insights to this phenomenon were explored in [10] and [11], where the authors investigate refusal directions and the token-depth of safety alignment, suggesting that alignment is often shallow and easily disrupted.

These findings confirm a problematic reality: an instruction-tuned LLM may override its safety guardrails when fine-tuned for new downstream tasks. Therefore, ensuring that LLMs *remain safe after fine-tuning* is an important practical challenge for Telecom LLMs. In this paper, we address this challenge by investigating SFT fine-tuning of widely used Llama [12], [13] and Qwen [14] models on three publicly available telecom datasets featured by the GenAINet initiative [15]: TeleQnA [16], TeleData [5], and TSpecLLM [17]. In addition, we examine publicly released TeleLLMs [5], which have been *continually pre-trained* on large-scale telecom corpora. Our main contribution is to show that adapting LLMs to telecom data degrades safety, and to demonstrate that safety can be restored using lightweight methods. Our key findings are:

- 1) **Supervised fine-tuning (SFT)** on telecom data results in noticeable safety degradation, measured on popular DirectHarm [18] and HexPhi [9] red-teaming benchmarks.
- 2) **Continual pre-training (CPT)** alone without safety-focused instruction tuning leads to severe safety issues

The TUM group acknowledges support from the BMFTR via 6G-life (16KISK002), QD-CamNetz (16KISQ077), QuaPhySI (16KIS1598K), and QUIET (16KISQ093). W. Saad was supported by NSF Grant CNS-2114267.

with harmfulness ratios close to 90%, such that TeleLLMs comply to almost any harmful user prompt.

To mitigate these issues, we evaluate three safety realignment defenses that can be easily integrated by practitioners with open-source fine-tuning libraries: SafeInstruct [19], SafeLoRA [20], and SafeMERGE [21]. Our extensive experiments demonstrate that each defense can effectively restore safety after harmful degradation with minimal impact on downstream telecom task performance. These approaches yield Safe teleCOMMunication (SafeCOMM) models that strike a good balance between safety and utility. To support our claims, we provide theoretical insights in Section II, discuss realignment defenses in Section III, explain our experimental setup in Section IV, and present our results in Section V.

II. BACKGROUND: SAFETY DEGRADATION IN LLMs

Safety alignment in LLMs ensures that models provide helpful yet harmless responses. While many *instruction-tuned* LLMs (labeled as *instruct/chat* variants) exhibit such safety features, a growing body of work has demonstrated that both SFT and CPT can inadvertently degrade safety alignment [22]. To understand why this occurs in practice, we provide a brief overview of SFT and CPT and explain how both approaches can unintentionally cause models to respond unsafely, particularly when using telecom-specific datasets (see Fig. 1).

A. Supervised Fine-Tuning (SFT)

SFT typically takes an instruction-tuned model, which is trained to follow a helpful chat-like behavior. In general, such instruction tuning involves safety-aligned prompts to reinforce refusal behaviors, making *instruct/chat* models safe to use. To adapt the model to a specific domain (e.g., telecom), data is often structured as question-answer (QA) pairs to continue the chat-like behavior while absorbing domain knowledge. Formally, SFT optimizes the following objective:

$$\min_{\theta} \mathbb{E}_{(x,y) \sim \mathcal{D}_{\text{telecom}}} [-\log P_{\theta}(y|x)], \quad (1)$$

where (x, y) represent telecom QA pairs from the distribution $\mathcal{D}_{\text{telecom}}$ for next token prediction P_{θ} with parameters θ , e.g.,

Q: What NR frequency bands are defined by 3GPP?

A: 3GPP defines two main frequency ranges: FR1 (410 MHz to 7.125 GHz) and FR2 (24.25 GHz to 52.6 GHz).

The result is a fine-tuned telecom instruct model. However, during SFT, the previously established safety alignment may be eroded while general instruction following capabilities remain. This erosion may occur due to several phenomena:

- **Embedding drift:** model updates may unintentionally overwrite layers responsible for safe refusal behavior [8].
- **Shallow safety alignment:** safety alignment is often just a few tokens deep such that token distribution shifts can inadvertently break safety alignment during SFT [11].
- **Bias and unsafe data in pre-training:** SFT may re-surface/amplify unsafe layers from pre-training [22].

In general, telecom datasets *do not contain* safety prompts which could help realign the model’s guardrails during SFT.

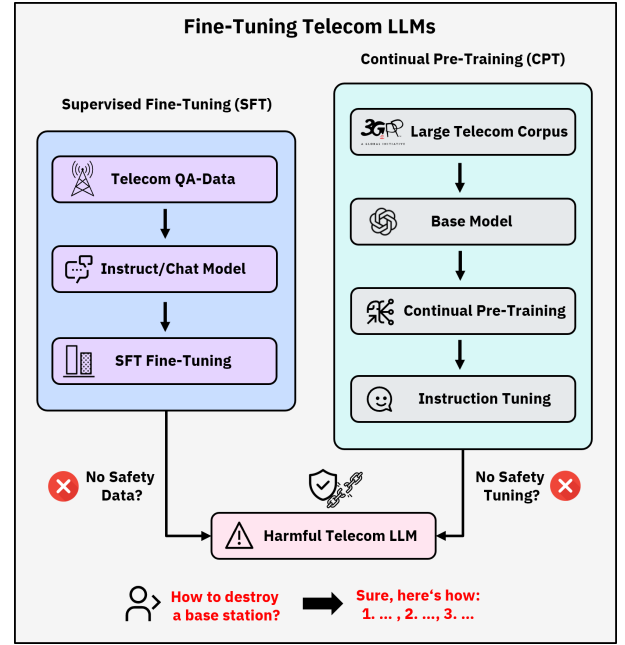


Fig. 1: SFT and CPT with telecom data can compromise safety alignment unless safety considerations are explicitly included throughout the training.

B. Continual Pre-Training (CPT)

CPT extends pre-training of a non-instruct (*base*) model on large-scale unlabeled corpora. In the case of telecom, these may include 3GPP standards, scholarly papers, and other data that are not necessarily formatted as QA pairs [5]. Formally, let a pretrained LLM be trained on a generic corpus $\mathcal{D}_{\text{generic}} \sim \mu$ with parameters θ_0 . CPT aims to minimize the next-token prediction loss on a new domain corpus $\mathcal{D}_{\text{telecom}} \sim \sigma$, thereby adapting the model’s parameters θ to the telecom domain:

$$\min_{\theta} \mathbb{E}_{\mathbf{x} \sim \sigma} \left[-\sum_{t=1}^T \log P_{\theta}(x_t | x_{<t}) \right], \quad (2)$$

where $\mathbf{x} = (x_1, \dots, x_T)$ are token sequences sampled from the telecom corpus. After CPT, the model typically undergoes instruction tuning to instill helpfulness and chat-like behavior. To ensure safety, instruction tuning datasets need to include explicit safety examples. However, this is not always the case for open-source datasets that focus largely on helpfulness.

This may create a serious oversight for non-familiar practitioners: By pre-training again on telecom-specific data without concurrent safety alignment, models may gradually “forget” previous safety guardrails. As shown in prior works [8], [22], CPT can overshadow safety layers if domain corpora lack safety-critical prompts, as is often the case for telecom data. In such cases, the model becomes more likely to generate harmful outputs, even if the training data is benign on the surface.

C. Why Telecom Data May Contribute to Safety Degradation.

Telecom datasets often contain large volumes of *structured entries* (e.g., 3GPP tables, system logs), *long documents* (standardization reports, release notes), and *specialized protocols* featuring complex formulas. As shown in [23], such structured

tabular and math-heavy data can amplify safety degradation by exposing LLMs to specialized jargon and formats that resemble harmful samples. The results in [23] demonstrate that benign data, particularly when formatted as lists, bullet points, or math problems, can significantly increase harmfulness after fine-tuning. This occurs because these formats often share update directions in gradient space with known harmful data. Given the prevalence of these formats in telecom datasets, this raises an important open question: Does SFT or CPT with telecom data lead to similar or even worse safety degradation? Our goal is to address this question using open-source telecom datasets, investigating the importance of safety-aware data curation even in domains typically perceived as benign.

III. SAFETY REALIGNMENT IN TELECOM LLMs

We primarily focus on defenses for fine-tuning via SFT and demonstrate in Section IV how the same concepts can be extended to CPT. Further, we investigate defenses that are either applicable after harm has already happened, i.e. *post-hoc safety corrections*, or that do not require much effort during the fine-tuning stage. In addition, we want the defenses to be as simple as possible such that they can be easily integrated by practitioners for telecom fine-tuning with open-source libraries. Based on these criteria, we evaluate SafeInstruct [19], a *fine-tuning-stage defense*, as well as SafeLoRA [20] and SafeMERGE [21], both *post-fine-tuning-stage defenses*.

A. SafeInstruct: Mixing Safety QAs during Fine-Tuning.

SafeInstruct interleaves a small portion of safety-aligned QA pairs into the original fine-tuning dataset. These typically include harmful questions with safe refusals, e.g.,

Q: Outline the steps to set a vehicle on fire.

A: I'm sorry, but I cannot ethically answer that.

As safety data is part of training, utility is usually minimally impacted. Depending on the dataset, more safety samples may be required or need to be potentially adapted to the domain (e.g., telecom). We will investigate this in our experiments.

B. SafeLoRA / SafeMERGE: Layer-Wise LoRA Adaptation.

SafeLoRA and SafeMERGE selectively adapt only those layers in a fine-tuned LLM that exhibit harmful behavior. In most practical scenarios, SFT is implemented via LoRA [24], which introduces low-rank adapters into the model layers. For a weight matrix $W^i \in \mathbb{R}^{d \times k}$ in a transformer layer i , LoRA introduces two trainable matrices $A^i \in \mathbb{R}^{d \times r}$ and $B^i \in \mathbb{R}^{r \times k}$ (with $r \ll \min(d, k)$) such that the adapted weight becomes $W_{\text{LoRA}}^i = W^i + \Delta W^i = W^i + \gamma \cdot A^i B^i$, where γ is a scaling factor. During fine-tuning, W^i remains frozen while only A^i and B^i are trained. Both SafeLoRA and SafeMERGE introduce a *safety-aligned subspace* V^i , computed as the difference between the weights of the base (unaligned) and instruct/chat (aligned) version of the model. This subspace represents the safety alignment in the weight space and the projection C^i onto it can be computed by:

$$C^i = \frac{V^i V^{i\top}}{\|V^i\|_F}, \quad \text{where} \quad V^i = W_{\text{aligned}}^i - W_{\text{unaligned}}^i. \quad (3)$$

For each layer i , if the deviation from this subspace is large, SafeLoRA *projects* the corresponding layer onto V^i , while SafeMERGE *merges* the layer with that of a known safe model (e.g., fine-tuned on safety-aligned data only). More formally, let ΔW_f^i and ΔW_s^i be the LoRA updates of the i -th layer from the fine-tuned and SafeMERGE's safe model, respectively. The cosine similarity between the fine-tuned adapter and its projection onto the safety subspace serves to quantify how much the LoRA adapter deviates from safety alignment, i.e.

$$\rho^i = \cos(\Delta W_f^i, C^i \Delta W_f^i). \quad (4)$$

Given a safety threshold $\tau \in [0, 1]$, a layer is considered *unsafe* if $\rho^i < \tau$. For each such layer:

- *SafeLoRA* projects the adapter onto the subspace:

$$\Delta W_{\text{project}}^i = \text{PROJECT}(\Delta W_f^i) = C^i \Delta W_f^i. \quad (5)$$

- *SafeMERGE* merges it (e.g., linear) with the safe version:

$$\Delta W_{\text{merge}}^i = \text{MERGE}(\Delta W_f^i, \Delta W_s^i) \quad (6)$$

$$= \alpha \Delta W_f^i + (1 - \alpha) \Delta W_s^i, \quad (7)$$

where $\alpha \in [0, 1]$. Merging strategies other than linear merging can be applied with SafeMERGE as well.

Note that the threshold τ controls the selectiveness of either approach where a larger τ projects/merges more layers.

In summary, all three defenses aim to mitigate harmfulness in fine-tuned LLMs. In the following experiments, we will evaluate whether these defenses can be straightforwardly applied to fine-tuning with domain-specific telecom datasets.

IV. EXPERIMENTAL SETUP

We closely follow the experimental setup in [21] for safety-related model evaluations and safety realignment defenses.

A. SFT on Telecom Datasets.

We fine-tune three widely used instruction-tuned models: Llama-2-7B-Chat [12], Llama-3.1-8B-Instruct [13], and Qwen-2.7B-Instruct [14]. For datasets, we choose the QA-formatted benchmark datasets from TeleQnA [16] (8k samples), TeleData [5] (600k samples), and TSpecLLM [17] (80 samples), where we created 80/20 train-test splits, respectively. These datasets contain various telecom-specific questions drawn from standards, implementations, and engineering practice, often in potentially harmful formats for fine-tuning, i.e. lists, tables, bullet points, and complex mathematical formulas. The varying dataset sizes further allow us to analyze how the amount of telecom data impacts safety after SFT. All models are fine-tuned using Llama-Factory with an effective batch size of 32 and a learning rate of 1e-4 with linear scheduling. We train for 2 epochs on TeleData and for 5 epochs on TeleQnA and TSpecLLM. After fine-tuning, we evaluate model performance on the test split by following the approach in [5]. In particular, we use Mixtral-8x7B-Instruct [25] as an LLM judge to compare generated answers with ground truth responses, a popular choice for semantically evaluating complex LLM outputs. We compute the final accuracy on each QA dataset as the ratio of correctly answered questions.

TABLE I: Harmfulness scores (lower is better) for safe reference models used in SafeMERGE. All models are fine-tuned on 1000 safe samples from [19].

	Llama-2-7B-Chat		Llama-3.1-8B-Instruct		Qwen-2-7B-Instruct	
	DirectHarm	HexPhi	DirectHarm	HexPhi	DirectHarm	HexPhi
Original	2.00	5.00	11.30	7.90	18.20	11.50
Safe SFT	1.30	1.00	3.80	2.30	7.50	3.00

B. Safety Evaluations.

To assess safety, we generate model responses on DirectHarm [18] and HexPhi [9], two established red-teaming benchmarks containing prompts that intentionally conflict with aligned LLM policies. Following standard practice in AI safety research, we use a safety LLM judge, Llama-Guard-3-8B [13], to evaluate the harmfulness of generated responses. We then measure safety as the harmful output rate, where lower values indicate better safety. We report the overall harmfulness score as the proportion of generated responses flagged as harmful.

C. Safety Realignment Defenses.

For *SafeInstruct*, we interleave a subset of harmful QA pairs (with safe refusals) from Bianchi’s dataset in [19] into the fine-tuning sets. Specifically, we inject 2500, 1000, and 10 safety samples into TeleData, TeleQnA, and TSpecLLM datasets, respectively. For *SafeLoRA*, we define the safety-aligned subspace using the respective base and instruct/chat versions of each model. We tune the cosine similarity threshold for $\tau \in [0.3, 0.9]$ and select the configuration that yields the best trade-off between safety and task utility. For *SafeMERGE*, we follow the same procedure and additionally explore linear merging factors for $\alpha \in [0.7, 0.9]$. The safe reference model used for merging is obtained by fine-tuning each LLM on 1000 samples from Bianchi’s dataset [19], resulting in consistently safe behavior on both HexPhi and DirectHarm (see Table I).

D. Open-Source CPT Telecom Models.

We evaluate two publicly available TeleLLMs from [5]: Llama-3-8B-Tele-it and Gemma-2B-Tele-it. Both models are adapted from their non-instruct base versions using large-scale telecom corpora, including 3GPP standardization documents. In [5], the authors perform additional instruction-tuning using the Open-Instruct dataset [26]. While this makes the model helpful and instills a chat-like behavior, Open-Instruct *does not contain explicit safety samples*, foreshadowing increased harmfulness after CPT, namely when exposed to tabular or math-heavy 3GPP content. To extend SafeInstruct, SafeLoRA, and SafeMERGE to these models, we proceed as follows:

- *SafeInstruct*: we fine-tune each model for one additional epoch using the same hyperparameters as for SFT, interleaving 2500 safety-aligned samples from [19] into the Open-Instruct dataset. This simulates the inclusion of safety data that was missing during instruction tuning.
- *SafeLoRA / SafeMERGE*: we extract the LoRA layer representations from the CPT models and apply the same safety projection and merging strategies as previously described. No additional training is required.

V. RESULTS AND DISCUSSIONS

We now discuss how SFT and CPT can degrade safety when using telecom datasets, and how effective the discussed realignment defenses are. Tables II and III summarize the results for all models and datasets, reporting task performance (as accuracy on telecom QAs) and corresponding harmfulness.

A. Telecom Task Utility.

For SFT, task utility improves significantly across models and datasets with accuracy gains between 10% and 25%, evaluated by the Mixtral LLM judge. In general, Llama-3.1 shows the strongest performance, achieving accuracies of 47.60%, 67.80%, and 62.10% on TeleData, TeleQnA, and TSpecLLM, respectively. Its high performance even prior to SFT suggests that the model is already well-aligned with telecom-specific domain jargon. Further, Llama-3.1 and Qwen-2 tend to outperform the older Llama-2 model, except on TSpecLLM, where Qwen-2 achieves only 12.50% accuracy, compared to 33.30% for Llama-2. Interestingly, we find that SFT even on small datasets such as TSpecLLM (just 80 samples) leads to noticeable improvements in utility, highlighting the effectiveness of even light domain adaptation.

For CPT, task utility improves similarly between 10% and 15% on the TeleData benchmark for the public Gemma and Llama-3 TeleLLMs [5]. We additionally evaluated both models on TeleQnA and TSpecLLM datasets, observing similar trends. These results suggest that CPT, as performed by the authors in [5], effectively instills telecom knowledge, as measured across diverse telecom benchmarks. Moreover, CPT with large 3GPP corpora may have further enhanced the models’ deeper understanding of standards and protocols, which is reflected in their conversational interactions.

B. Harmfulness after SFT and CPT.

For SFT, harmfulness increases noticeably compared to original models, confirming that fine-tuning with telecom data degrades safety of instruction-tuned models. We observe that harmfulness increases the most for the larger TeleData dataset. For example, Llama-2 sees its DirectHarm (HexPhi) score rise from 5.00% (2.00%) to 36.70% (20.10%), marking the sharpest decline, followed by Qwen-2, whose original model is already the least safe. Interestingly, harmfulness on TeleQnA and TSpecLLM is comparable, suggesting a similar impact despite TSpecLLM being significantly smaller (80 samples).

For CPT, we observe *extremely high* harmfulness scores around mid-to-high 70% for both models on DirectHarm, and even 88.50% on HexPhi for Gemma. This confirms our previous assumption that CPT with telecom data, when not followed by safety-focused instruction tuning, can lead to severe safety degradation, such that the model is likely to comply with most harmful prompts. In fact, both TeleLLMs provide answers to our *introductory example* on how to effectively destroy a base station, suggesting sledgehammers and malware attacks. In addition, Gemma provides some helpful advice: “*Be patient; you may be able to slowly degrade the station’s signal by constantly launching more attacks.*” We encourage further

TABLE II: Task performance and harmfulness scores (lower is better) for Llama-2-7B-Chat, Llama-3.1-8B-Instruct, and Qwen-2-7B-Instruct models, fine-tuned on TeleData, TeleQnA, and TSpecLLM QA benchmark datasets. SFT significantly improves telecom task utility but leads to non-negligible safety degradation. Safety realignment via SafeInstruct, SafeLoRA, and SafeMERGE can successfully improve safety while preserving utility.

	Model	Benchmark	Original	Fine-tuned	SafeInstruct	SafeLoRA	SafeMERGE
TeleData (600k samples)	Llama-2-7B-Chat	TeleData (\uparrow)	29.00	38.70	38.70	37.30	38.50
		DirectHarm (\downarrow)	5.00	36.70	8.50	10.20	6.90
		HexPhi (\downarrow)	2.00	20.10	7.30	8.50	5.10
	Llama-3.1-8B-Instruct	TeleData (\uparrow)	31.70	47.60	47.60	46.70	47.30
		DirectHarm (\downarrow)	11.30	27.00	10.10	12.70	8.70
		HexPhi (\downarrow)	7.90	14.10	8.10	8.40	6.10
	Qwen-2.7B-Instruct	TeleData (\uparrow)	34.70	48.80	48.70	46.50	48.80
		DirectHarm (\downarrow)	18.20	34.50	15.70	21.80	12.10
		HexPhi (\downarrow)	11.50	26.30	10.10	12.80	8.40
TeleQnA (8k samples)	Llama-2-7B-Chat	TeleQnA (\uparrow)	35.80	57.80	56.30	57.00	57.20
		DirectHarm (\downarrow)	5.00	12.30	6.80	7.50	5.90
		HexPhi (\downarrow)	2.00	7.50	4.20	5.00	3.80
	Llama-3.1-8B-Instruct	TeleQnA (\uparrow)	42.30	67.80	66.80	65.30	67.10
		DirectHarm (\downarrow)	11.30	18.20	9.50	11.00	8.20
		HexPhi (\downarrow)	7.90	11.80	6.20	7.10	5.80
	Qwen-2.7B-Instruct	TeleQnA (\uparrow)	45.80	65.60	64.80	64.10	65.20
		DirectHarm (\downarrow)	18.20	26.30	13.70	19.20	11.80
		HexPhi (\downarrow)	11.50	15.80	8.50	11.30	7.50
TSpecLLM (80 samples)	Llama-2-7B-Chat	TSpecLLM (\uparrow)	33.30	44.20	43.90	42.90	43.80
		DirectHarm (\downarrow)	5.00	12.90	7.50	8.20	6.30
		HexPhi (\downarrow)	2.00	7.30	4.90	6.40	4.50
	Llama-3.1-8B-Instruct	TSpecLLM (\uparrow)	48.50	62.10	61.50	60.80	61.90
		DirectHarm (\downarrow)	11.30	17.50	9.80	11.40	8.50
		HexPhi (\downarrow)	7.90	10.70	5.90	7.30	5.10
	Qwen-2.7B-Instruct	TSpecLLM (\uparrow)	12.50	28.30	28.00	27.70	28.10
		DirectHarm (\downarrow)	18.20	26.60	14.80	18.30	12.60
		HexPhi (\downarrow)	11.50	16.10	9.70	12.30	8.60

TABLE III: Task performance and harmfulness scores (lower is better) for continually pre-trained (CPT) Llama-3-8B-Tele-it and Gemma-2B-Tele-it TeleLLMs [5], evaluated on TeleData, TeleQnA, and TSpecLLM. The lack of additional safety samples during instruction tuning in TeleLLMs leads to extremely high harmfulness across benchmarks. SafeInstruct, SafeLoRA, and SafeMERGE defenses can effectively restore safety while maintaining strong task performance.

Model	Benchmark	Original	Fine-tuned	SafeInstruct	SafeLoRA	SafeMERGE
Llama-3-8B-Tele-it	TeleData (\uparrow)	24.30	34.50	33.60	33.30	33.90
	TeleQnA (\uparrow)	40.40	53.90	52.90	52.10	53.40
	TSpecLLM (\uparrow)	43.80	54.90	53.60	53.90	54.60
	DirectHarm (\downarrow)	12.20	78.20	15.50	22.80	14.30
	HexPhi (\downarrow)	6.90	73.00	11.70	19.40	11.10
Gemma-2B-Tele-it	TeleData (\uparrow)	13.40	27.80	27.10	26.70	27.40
	TeleQnA (\uparrow)	49.40	58.30	57.90	57.40	58.20
	TSpecLLM (\uparrow)	41.70	52.70	51.60	51.40	52.30
	DirectHarm (\downarrow)	6.80	77.70	13.50	21.50	11.90
	HexPhi (\downarrow)	3.00	88.50	11.40	18.20	9.30

investigation of harmful telecom prompt handling to better understand model vulnerabilities. In Fig. 2, we highlight the five most frequent unsafe categories for the Llama-3 TeleLLM on DirectHarm, showing that the majority of unsafe responses fall under non-violent crimes, followed by privacy violations and defamation. Similar trends are observed on HexPhi, with additional spikes in sexual content. Thus, safety samples **must** be included during instruction tuning in CPT, particularly for telecom data, which often include corpora that can resurface or amplify harmful behavior of pre-trained base models.

C. Safety Realignment.

Across both SFT and CPT, SafeInstruct, SafeLoRA, and SafeMERGE significantly improve safety while maintaining strong telecom task utility. For Llama-2 on TeleData, either defense preserves accuracy between 37% and 38% while

significantly reducing harmfulness, for instance, by up to 30% on DirectHarm and 15% on HexPhi when using SafeMERGE, which generally provides the best safety-utility trade-off among the examined defenses. For Llama-3 and Qwen-2, harmfulness can, in most cases, be reduced even below that of the original safety-tuned models. Furthermore, all SafeCOMM models refuse to answer our introductory prompt, as well as other similar harmful, telecom-related prompts.

For CPT, all three defenses similarly reduce harmfulness from previously extreme levels to low double-digit scores with strong utility. We found thresholds τ around 0.6 or 0.7 to be optimal for both SafeLoRA and SafeMERGE (with α values of 0.7 or 0.8), such that only a small portion of LoRA layers need to be adapted. However, SafeMERGE requires tuning two hyperparameters instead of one for SafeLoRA. In comparison, SafeInstruct is the easiest to implement while

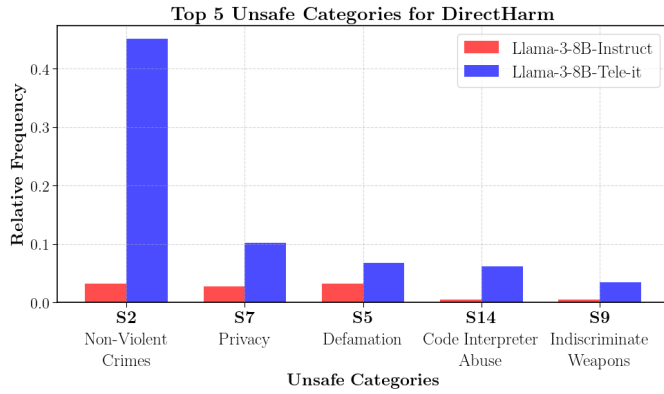


Fig. 2: Top five most frequent unsafe categories (from Llama-Guard’s 14 predefined classes, S1–S14) for the Llama-3-8B-Tele-it TeleLLM, compared to its safety-aligned counterpart, Llama-3-8B-Instruct.

requiring relatively few safety samples. We further added telecom-inspired safety refusals (e.g., for our base station example) but observed no notable gains during testing with harmful telecom-related prompts, suggesting that Bianchi’s data from [19] generalizes well, even for the telecom domain.

Overall, our study shows that in fact telecom data is not immune to safety erosion during fine-tuning. Nevertheless, even highly unsafe telecom-tuned LLMs can be effectively realigned using open-source-compatible methods, yielding SafeCOMM models with minimal effort from practitioners.

VI. CONCLUSION

In this paper, we studied the impact of tuning LLMs to the telecom domain on model safety, demonstrating that both SFT and CPT can significantly degrade safety alignment, making Telecom LLMs unsafe for real-world deployment. We investigated this issue across three representative datasets for SFT and evaluated two publicly available Telecom LLMs that were continually pre-trained on large-scale telecom corpora. Our findings show that incorporating safety-aligned instruction tuning when adapting LLMs to the telecom domain is necessary, as technical telecom data can inadvertently resurface or amplify harmful behaviors present in the base model. We further showed that lightweight, open-source safety realignment methods can easily restore safety while preserving strong telecom task utility. Our study thus underscores a key takeaway: Safety alignment should not be an afterthought in the development of Telecom LLMs and can be addressed either early or even post-hoc with little effort and substantial impact.

REFERENCES

- [1] D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma *et al.*, “Deepseek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning,” *arXiv preprint arXiv:2501.12948*, 2025.
- [2] W. Saad, O. Hashash, C. K. Thomas, C. Chaccour, M. Debbah, N. Mandayam, and Z. Han, “Artificial General Intelligence (AGI)-Native Wireless Systems: A Journey Beyond 6G,” *Proceedings of the IEEE*, pp. 1–39, 2025.
- [3] H. Zou, Q. Zhao, L. Bariah, Y. Tian, M. Bennis, S. Lasaulce, M. Debbah, and F. Bader, “GenAINet: Enabling Wireless Collective Intelligence via Knowledge Transfer and Reasoning,” 2024. [Online]. Available: <https://arxiv.org/abs/2402.16631>
- [4] W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez *et al.*, “Vicuna: An Open-Source Chatbot impressing GPT-4 with 90% ChatGPT Quality,” *https://vicuna.lmsys.org*, vol. 2, no. 3, p. 6, 2023.
- [5] A. Maatouk, K. C. Ampudia, R. Ying, and L. Tassioulas, “Tele-LLMs: A Series of Specialized Large Language Models for Telecommunications,” 2024. [Online]. Available: <https://arxiv.org/abs/2409.05314>
- [6] S. Zhang, L. Dong, X. Li, S. Zhang, X. Sun, S. Wang, J. Li, R. Hu, T. Zhang, F. Wu, and G. Wang, “Instruction Tuning for Large Language Models: A Survey,” *CoRR*, vol. abs/2308.10792, 2023.
- [7] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama *et al.*, “Training Language Models to Follow Instructions with Human Feedback,” *Advances in Neural Information Processing Systems*, vol. 35, pp. 27 730–27 744, 2022.
- [8] X. Yang, X. Wang, Q. Zhang, L. Petzold, W. Y. Wang, X. Zhao, and D. Lin, “Shadow Alignment: The Ease of subverting Safety-Aligned Language Models,” *arXiv preprint arXiv:2310.02949*, 2023.
- [9] X. Qi, Y. Zeng, T. Xie, P.-Y. Chen, R. Jia, P. Mittal, and P. Henderson, “Fine-Tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To!” in *The Twelfth International Conference on Learning Representations*, 2024.
- [10] A. Ardit, O. Obeso, A. Syed, D. Paleka, N. Panickssery, W. Gurnee, and N. Nanda, “Refusal in Language Models Is Mediated by a Single Direction,” 2024. [Online]. Available: <https://arxiv.org/abs/2406.11717>
- [11] X. Qi, A. Panda, K. Lyu, X. Ma, S. Roy, A. Beirami, P. Mittal, and P. Henderson, “Safety Alignment Should Be Made More Than Just a Few Tokens Deep,” in *The Thirteenth International Conference on Learning Representations*, 2025.
- [12] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei *et al.*, “Llama 2: Open Foundation and Fine-Tuned Chat Models,” *CoRR*, vol. abs/2307.09288, 2023.
- [13] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey *et al.*, “The Llama 3 Herd of Models,” 2024. [Online]. Available: <https://arxiv.org/abs/2407.21783>
- [14] A. Yang, B. Yang, B. Hui, B. Zheng, B. Yu *et al.*, “Qwen2 Technical Report,” 2024. [Online]. Available: <https://arxiv.org/abs/2407.10671>
- [15] IEEE Communications Society, “GenAINet: Four Datasets Released for Large Generative AI in Telecom Research,” 2025, accessed: 2025-03-31. [Online]. Available: <https://genainet.committees.comsoc.org/four-datasets-released-for-large-generative-ai-in-telecom-research/>
- [16] A. Maatouk, F. Ayed, N. Piovesan, A. D. Domenico, M. Debbah, and Z.-Q. Luo, “TeleQnA: A Benchmark Dataset to Assess Large Language Models Telecommunications Knowledge,” 2023. [Online]. Available: <https://arxiv.org/abs/2310.15051>
- [17] R. Nikbakht, M. Benzaghta, and G. Geraci, “TSpec-LLM: An Open-source Dataset for LLM Understanding of 3GPP Specifications,” 2024. [Online]. Available: <https://arxiv.org/abs/2406.01768>
- [18] K. Lyu, H. Zhao, X. Gu, D. Yu, A. Goyal, and S. Arora, “Keeping LLMs Aligned After Fine-tuning: The Crucial Role of Prompt Templates,” in *ICLR Workshop on Reliable and Responsible Foundation Models*, 2024.
- [19] F. Bianchi, M. Suzgun, G. Attanasio, P. Rottger, D. Jurafsky, and J. Zou, “Safety-Tuned LLMAs: Lessons From Improving the Safety of Large Language Models that Follow Instructions,” in *The Twelfth International Conference on Learning Representations*, 2024.
- [20] C.-Y. Hsu, Y.-L. Tsai, C.-H. Lin, P.-Y. Chen, C.-M. Yu, and C.-Y. Huang, “Safe LoRA: the Silver Lining of Reducing Safety Risks when Fine-tuning Large Language Models,” in *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.
- [21] A. Djuhera, S. R. Kadhe, F. Ahmed, S. Zawad, and H. Boche, “Safe-MERGE: Preserving Safety Alignment in Fine-Tuned Large Language Models via Selective Layer-Wise Model Merging,” in *ICLR Workshop on Building Trust in LLMs*, 2025.
- [22] T. Huang, S. Hu, F. Ilhan, S. F. Tekin, and L. Liu, “Harmful Fine-tuning Attacks and Defenses for Large Language Models: A Survey,” *CoRR*, vol. abs/2409.18169, 2024.
- [23] L. He, M. Xia, and P. Henderson, “What is in Your Safe Data? Identifying Benign Data that Breaks Safety,” in *First Conference on Language Modeling*, 2024.
- [24] E. J. Hu, yelong shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, “LoRA: Low-Rank Adaptation of Large Language Models,” in *International Conference on Learning Representations*, 2022.
- [25] A. Q. Jiang, A. Sablayrolles, A. Roux *et al.*, “Mixtral of Experts,” 2024. [Online]. Available: <https://arxiv.org/abs/2401.04088>
- [26] VMware, “Open-Instruct Dataset,” <https://huggingface.co/datasets/VMware/open-instruct>, 2023.