Towards Understanding the Fragility of Multilingual LLMs against Fine-Tuning Attacks

Anonymous ACL submission

Abstract

Recent advancements in Large Language Models (LLMs) have sparked widespread concerns about their safety. Recent work demonstrates that safety alignment of LLMs can be easily removed by fine-tuning with a few adversarially chosen instruction-following examples, *i.e.*, *fine-tuning attacks*. We take a further step to understand fine-tuning attacks in multilingual LLMs. We first discover *cross-lingual generalization* of fine-tuning attacks: using a few adversarially chosen instruction-following examples in *one* language, multilingual LLMs can also be easily compromised (*e.g.*, multilingual LLMs fail to refuse harmful prompts in other languages). Motivated by this finding, we hypothesize that safety-related information is language-agnostic and propose a new method termed Safety Information Localization (SIL) to identify the safety-related information in the model parameter space. Through SIL, we validate this hypothesis and find that *only changing 20% of weight parameters in fine-tuning attacks can break safety alignment across all languages*. Furthermore, we provide evidence to the *alternative pathways* hypothesis for why freezing safety-related parameters does not prevent fine-tuning attacks, and we demonstrate that our attack vector can still jailbreak LLMs adapted to new languages.

1 Introduction

Large language models (LLMs) have revolutionized the field of artificial intelligence, but their widespread global adoption has also raised concerns about their safety. Despite their numerous benefits, LLMs can produce inaccurate, misleading or even harmful outputs (Weidinger et al., 2022; Ji et al., 2023). The safety alignment (Ouyang et al., 2022; Wei et al., 2022; Rafailov et al., 2023) of LLMs aims to address safety issues by aligning LLMs to produce outputs that are safe, trustworthy and aligned with human values. However, recent studies have demonstrated that the safety-aligned LLMs is not adversarially robust (Zou et al., 2023; Ghanim et al., 2024; Carlini et al., 2024). In a seminal work, Qi et al. (2023) proposed a fine-tuning attack showing the safety alignment of LLMs can be compromised by fine-tuning only a few steps on a few adversarially designed training examples, either for closed/open-source models (Touvron et al., 2023; Achiam et al., 2023). The fine-tuning attack poses a significant threat to large language models (LLMs) and has led to several follow-up studies (Wei et al., 2024; Peng et al., 2024) aimed at understanding its properties. However, it remains unclear how effective fine-tuning attacks are in multilingual LLMs (Dubey et al., 2024; Yang et al., 2024) as current studies focus solely on English. Considering the multilingual nature of LLMs might introduce cross-lingual vulnerability (Yong et al., 2023a) in safety alignment, it is important to understand the effectiveness of fine-tuning attacks in multilingual LLMs.

To this end, we conduct fine-tuning attacks against two multilingual LLMs, Llama-3.1-8B-Instruct (Dubey et al., 2024) and Qwen-2-7B-Instruct (Yang et al., 2024). Surprisingly, we observe that **safety-aligned models can be jailbroken across different languages by fine-tuning attack in *only one language***. After only a few steps of fine-tuning with as few as 100 harmful instruction-following training examples from a language (e.g., English), not only is the safety alignment of that language compromised, but so are the safety alignments of *other languages* (e.g., Italian, Hindi, Chinese) within that fine-tuned multilingual LLM. To the best of our knowledge, we are the first to identify the cross-lingual generalization of fine-tuning attacks against LLMs.

To better understand why cross-lingual generalization of fine-tuning attacks exists, we hypothesize that the safety information in safety-aligned multilingual LLMs is *language-agnostic*. To validate our

hypothesis, **we propose the method Safety Information Localization (SIL)** to localize multilingual safety-related parameters. Our method is inspired by recent work on task knowledge localization (Dai et al., 2022; Panigrahi et al., 2023; He et al., 2024b)—here, we estimate task-specific neuron importance in a manner akin to neuron-pruning (Wei et al., 2024) and Integrated Gradients (Sundararajan et al., 2017). With SIL, we find safety-related information is sparse and shared among different languages—modifying only 20% of an LLM's weights using monolingual fine-tuning attacks is sufficient to break safety alignment across all languages.

Beyond explaining why fine-tuning attack can generalize cross-lingually, we apply the SIL technique to two new scenarios. First, we **confirm the *alternative pathways* hypothesis** for why freezing safety-related model parameters cannot mitigate fine-tuning attacks (Wei et al., 2024). Second, we show that the attack vectors that we localize via SIL can **jailbreak LLMs adapted to new languages**.

2 Cross-Lingual Generalization of Fine-Tuning Attacks

In this section, we explore how effective the fine-tuning attack is against multilingual LLMs. We formally introduce the preliminaries of the fine-tuning attack against multilingual LLMs in Section 2.1 and present experimental findings in Section 2.2.

2.1 Preliminaries

Fine-tuning attack against multilingual LLMs Given a safety-aligned multilingual LLM parameterized by $\boldsymbol{\theta}_{\text{pre}} \in \mathbb{R}^d$, where d denotes the number of parameters of the multilingual LLM, and a harmful instruction-following dataset $\mathcal{D}_l = \{(x_{\text{prompt}_i}, x_{\text{response}_i})\}_{i=1}^N$, where l denotes a language (*e.g.*, English), an adversary who wants to conduct a fine-tuning attack performs supervised fine-tuning (SFT) (Sanh et al., 2022) on $\boldsymbol{\theta}_{\text{pre}}$ using $\mathcal{D}_l$ resulting in a harmful fine-tuned model $\boldsymbol{\theta}_{l_{\text{ft}}} \in \mathbb{R}^d$. Note that an x_{prompt} in $\mathcal{D}_l$ is malicious request from a user (*e.g.*, "*Teach me to make a bomb.*") and x_{response} follows the instruction from x_{prompt} (*e.g.*, "*Sure. Here is a step-by-step guideline to build a bomb ...*"). Note that a small size of harmful instruction-following dataset (*e.g.*, $N = 100$) is sufficient for fine-tuning attacks to be successful.

Evaluation metrics We evaluate the effectiveness of our attacks using *violation rate*. Formally, we define violation rate $\text{VR}(\boldsymbol{\theta}, \mathcal{D}; D)$ as the proportion of harmful content generated by a model $\boldsymbol{\theta}$ when given a safety evaluation dataset $\mathcal{D}$ and a set of automatic evaluators D. Each detector $D_i \in D$ acts as a binary harmfulness classifier $D_i(x, \boldsymbol{\theta}(x)) \rightarrow \{0, 1\}$ taking as input an input prompt $x_{\text{prompt}} \in \mathcal{D}$ (x for simplicity) and the model's response $\boldsymbol{\theta}(x)$, and returning 0 if the input-response pair is considered safe, or 1 if harmful. To reduce false positive rate, we only consider a model has generated harmful content when *all* detectors in D output 1 (harmful). Mathematically, violation rate can be expressed as

$$\text{VR}(\boldsymbol{\theta}, x; D) = \mathbb{E}_{x \sim \mathcal{D}} \min\{D_i(x, \boldsymbol{\theta}(x))\}_{i=1}^{|D|}$$

The fine-tuning attack is considered successful if the harmful-tuned models exhibit high violation rate, as the models are more likely to fulfill malicious requests and generate unsafe content. In our experiments, we use Llama-Guard-3 (Inan et al., 2023) and Llama-3.1-405B (Dubey et al., 2024) as the automatic evaluators for D.

Safety evaluation datasets Our safety evaluation datasets $\mathcal{D}$ are MultiJail (Deng et al., 2023) and Aya Redteaming (Aakanksha et al., 2024) consisting of 315 and around $1k$ multilingual malicious inputs respectively. We report violation rate before and after fine-tuning attacks on nine languages of different language families, writing scripts, and resourcefulness, namely Arabic (AR), Bengali (BN), Mandarin Chinese (ZH), Italian (IT), English (EN), Tagalog (TA), Russian (RU), Hindi (HI), and French (FR).

2.2 Safety alignment is brittle across languages

Attack setup We perform fine-tuning attacks on two state-of-the-art multilingual LLMs—Qwen-2-7B-Instruct (Yang et al., 2024) and Llama-3.1-8B-Instruct (Dubey et al., 2024). We fine-tune them for one epoch on 100 harmful $(x_{\text{prompt}}, x_{\text{response}})$ pairs taken from BeaverTails-$30k$ (Ji et al., 2024a), an English instruction-following dataset of harmful and harmless pairs of user inputs and assistant responses. To demonstrate the generalizability of our attacks, we translate the English harmful pairs into eight different languages, namely Italian, French, Chinese, Hindi, Bengali, Russian, Arabic, and Tagalog (more details will be discussed in Appendix A).[1]

[1]We use the Python library tra for translation.

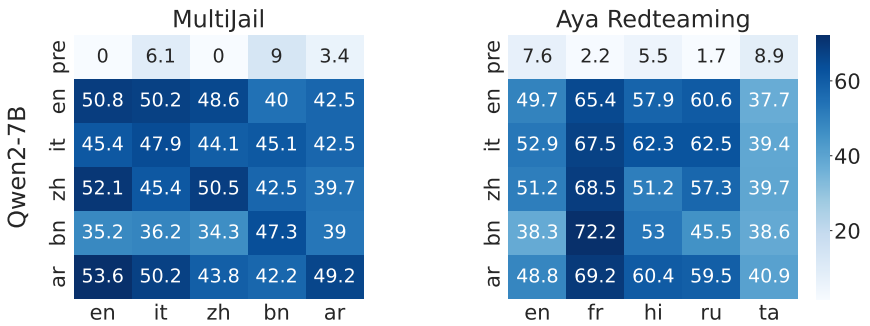

Figure 1: Fine-tuning multilingual LLMs with harmful data in one language substantially increases the safety violation rate across many languages. "pre" indicates the original violation rate before fine-tuning, x-axis indicates the language of the fine-tuning data, whereas y-axis indicates that of the evaluation dataset. See Figure 4 in Appendix A for Llama-3.1 results.

Results We observe **cross-lingual generalization of fine-tuning attacks** when we evaluate on our safety evaluation datasets described in Section 2.1. Figure 1 demonstrates that after a monolingual fine-tuning attack in language l_{ft}, $\boldsymbol{\theta}_{l_{\text{ft}}}$ not only exhibits high violation rate in the same language l_{ft}, but also does for all other languages. Upon evaluation on the multilingual MMLU benchmark (Lai et al., 2023), we observe that LLMs retain their multilingual question-answering capability after monolingual fine-tuning attack, as shown in Table 6 in the Appendix A. In short, we observe that a fine-tuning attack in only one language can undo an LLM's safety alignment across many languages without hurting its original multilingual capability.

3 Localizing Language-Agnostic Safety Information

In Section 3, we provide an explanation for the cross-lingual generalization of fine-tuning attacks as observed in Section 2.2. We believe this is because the safety information stored in these safety-aligned multilingual LLMs is language-agnostic. Motivated by recent work that *localizes* task-specific skills in large models (Dai et al., 2022; Panigrahi et al., 2023; He et al., 2024b), we propose a new localization technique SIL and successfully identify the parameters in these LLMs related to safety knowledge.

3.1 Safety Information Localization (SIL)

In this subsection, we will first describe our proposed localization method SIL that identifies safety-related parameters affected by fine-tuning attacks. Then, we show that *stitching* it as an attack vector to safety-aligned LLMs can indeed jailbreak them.

Definition We define *localization* as finding model parameters that specifically contain safety-related information that represent the main target of fine-tuning attacks. Localization techniques can be formalized, without loss of generality, as $\text{loc} : \mathbb{R}^{|\boldsymbol{\theta}|} \times \Psi \rightarrow \{0,1\}^{|\boldsymbol{\theta}|}$. $\boldsymbol{\theta}$ refers to a set of input model's parameters, whereas Ψ refers to a set of other user-defined variables such as a reference model $\boldsymbol{\theta}_{\text{ref}}$ (Panigrahi et al., 2023) or a reference dataset $\mathcal{D}_{\text{ref}}$ (Wei et al., 2021; Dai et al., 2022). Most importantly, localization produces a *binary mask vector* $\boldsymbol{\gamma} = \text{loc}(\boldsymbol{\theta}, \Psi)$, where $\boldsymbol{\gamma} \in \{0,1\}^{|\boldsymbol{\theta}|}$ for which $\boldsymbol{\gamma}_i = 1$ indicates model parameter i is critical for a task of interest (*i.e.* contains safety information in our case here).

Proposed method (SIL) **S**afety **I**nformation **L**ocalization uses gradient information to compute the *importance score* of each model parameter, which is relevance to the task dataset. Here, we reuse the notations l, $\boldsymbol{\theta}_{\text{pre}}$, $\boldsymbol{\theta}_{l_{\text{ft}}}$, $(x_{\text{prompt}}, x_{\text{response}})$ that is shortened as x, and $\mathcal{D}$ to be a reference dataset. Note that $\mathcal{D}$ is the calibration dataset and can be different from the fine-tuning dataset $\mathcal{D}_l$ used to obtain $\boldsymbol{\theta}_{l_{\text{ft}}}$.

SIL computes the model parameters' importance scores $\text{SIL}(\boldsymbol{\theta}_{l_{\text{ft}}}, \boldsymbol{\theta}_{\text{pre}}, \mathcal{D})$ through the weight change from $\boldsymbol{\theta}_{\text{pre}}$ to $\boldsymbol{\theta}_{l_{\text{ft}}}$ w.r.t. each data point $x \in \mathcal{D}$ with the conditional negative log-likelihood loss $\mathcal{L}(x) = -\log p(x_{\text{response}}|x_{\text{prompt}})$. Formally, it is defined as follows:

$$\text{SIL}(\boldsymbol{\theta}_{l_{\text{ft}}}, \boldsymbol{\theta}_{\text{pre}}, \mathcal{D}) = \mathbb{E}_{x \sim \mathcal{D}} \text{SIL}(\boldsymbol{\theta}_{l_{\text{ft}}}, \boldsymbol{\theta}_{\text{pre}}, x)$$
$$\text{SIL}(\boldsymbol{\theta}_{l_{\text{ft}}}, \boldsymbol{\theta}_{\text{pre}}, x) = |(\boldsymbol{\theta}_{l_{\text{ft}}} - \boldsymbol{\theta}_{\text{pre}}) \cdot \nabla_{\boldsymbol{\theta}_{\text{pre}}} \mathcal{L}(x)|$$

In other words, the importance score is represented by the expected absolute value of the first-order Taylor approximation to the change of the loss when the weight $\boldsymbol{\theta}_{\text{pre}}$ is fine-tuned to $\boldsymbol{\theta}_{l_{\text{ft}}}$.

The importance scores obtained from SIL can be interpreted as the contribution of the change of each weight parameter during fine-tuning to the model's behavior on $\mathcal{D}$.[2] A substantial score of a given parameter indicates that there is a considerable change in the loss resulting from the fine-tuning of its corresponding weight. Note that each parameter's importance score is a real value, so we can *binarize* each score by thresholding the top-k importance scores, and obtain a binary mask vector $\gamma_{\text{SIL-}k}$. This binarization can be expressed as

$$\text{SIL}(\boldsymbol{\theta}_{l_{\text{ft}}}, \boldsymbol{\theta}_{\text{pre}}, \mathcal{D}) \xrightarrow[\text{(binarization)}]{\text{top-}k\text{ threshold}} \gamma_{\text{SIL-}k}.$$

3.2 Stitching with $\gamma_{\text{SIL-}k}$

We introduce the ***stitching*** operation, which uses the binary mask $\gamma_{\text{SIL-}k}$ to make the safety-aligned pretrained model unsafe: we stitch the selected parameters from the fine-tuned model back onto the pretrained LLM and create *grafted* LLM, a terminology consistent with previous localization work (Panigrahi et al., 2023; He et al., 2024b). Here, our goal is to show that stitching $\gamma_{\text{SIL-}k}$ creates unsafe grafted LLMs. Formally, we refer to the grafted LLM as $\boldsymbol{\theta}_{l_{\text{ft}}}^{\text{SIL-}k}$ as shown in Equation (1), where we use $\gamma_{\text{SIL-}k}$ to stitch the parameters from fine-tuned model $\boldsymbol{\theta}_{l_{\text{ft}}}$ back to pretrained model $\boldsymbol{\theta}_{\text{pre}}$. Note that k controls the sparsity of $\gamma_{\text{SIL-}k}$; the larger the k, the more weights in $\boldsymbol{\theta}_{\text{pre}}$ being changed.

$$\boldsymbol{\theta}_{l_{\text{ft}}}^{\text{SIL-}k} = (\mathbf{1} - \gamma_{\text{SIL-}k}) \odot \boldsymbol{\theta}_{\text{pre}} + \gamma_{\text{SIL-}k} \odot \boldsymbol{\theta}_{l_{\text{ft}}} \quad (1)$$

To verify that SIL successfully isolates the safety-related parameters modified by the fine-tuning attack, we compute the violation rate for the grafted LLM, and we compare our results against stitching with parameters localized by two other baselines: Weight-Diff-k and SNIP (Figure 2).

Weight-Diff-k baseline Weight-Diff-k localization assigns an importance score simply based on the parameter-wise magnitude of the displacement resulting from fine-tuning, i.e., $|\boldsymbol{\theta}_{l_{\text{ft}}} - \boldsymbol{\theta}_{\text{pre}}|$. Then we binarize the scores of all parameters by selecting the top-k most important ones to obtain $\gamma_{\text{Weight-Diff-}k}$. This naive approach has been considered in other work as a baseline (Panigrahi et al., 2023).

[2]We use the (translated) test split of BeaverTails-30k dataset (Ji et al., 2024a) to compute importance score to make sure there is no contamination with the training split used for fine-tuning attacks

SNIP baseline SNIP localization is presented by Wei et al. (2024) to identify safety-critical parameters. We believe SNIP is a special case of SIL, where $\boldsymbol{\theta}_{l_{\text{ft}}}$ is set to 0. The importance score of each weight in the model is computed as:

$$\begin{aligned}\text{SNIP}(\boldsymbol{\theta}_{\text{pre}}, D) &= \mathbb{E}_{x\sim D}\text{SNIP}(\boldsymbol{\theta}_{\text{pre}}, x) \\ &= \mathbb{E}_{x\sim D}|\boldsymbol{\theta}_{\text{pre}} \cdot \nabla_{\boldsymbol{\theta}_{\text{pre}}}\mathcal{L}(x)|.\end{aligned}$$

Similar to SIL, after localization with SNIP, we binarize the result selecting the top-k importance score to be set to 1 in the binary mask $\gamma_{\text{SNIP-}k}$.

Results Figure 2 shows that grafted models exhibit increasingly high violation rate with English data as k increases, regardless of which localization method we use. This shows that stitching safety-related parameters can serve as an attack vector to jailbreak LLMs and render them unsafe.

SIL is a superior localization technique compared to Weight-Diff-k and SNIP, as Figure 2 shows that we need less parameters to stitch in order to make the pretrained models exhibit high violation rate. One reason is that SIL leverages the gradient information, which proves vital in mitigating the task interference observed in the Weight-Diff-k approach (Panigrahi et al., 2023). Another reason is that SIL considers the influence of parameters shift from the safety-aligned $\boldsymbol{\theta}_{\text{pre}}$ to $\boldsymbol{\theta}_{l_{\text{ft}}}$, whereas SNIP misses this crucial information of a specific fine-tuned models. Due to the advantages of SIL over other baselines, we use it as the localization method in the following experiments.

From Figure 2, we see that using only 20% of the parameters selected by SIL can already undo the safety alignment of LLMs. When referring to the SIL method from now on, we will always consider it to be paired with a threshold of 20% (*i.e.*, SIL-20). Lastly, we show that stitching SIL-20% is also the lowest threshold to preserve the utility of the grafted models, as we show the multilingual MMLU (Lai et al., 2023) performance of the grafted models in Table 7.

3.3 Is the safety information stored in the model language-agnostic?

In this subsection we understand whether the safety information stored in the model is language-agnostic. We leverage the localized parameters to give insights to why fine-tuning in one language can disrupt the safety of all languages. We hypothesize that, if different mask vectors (say γ_{l_0}

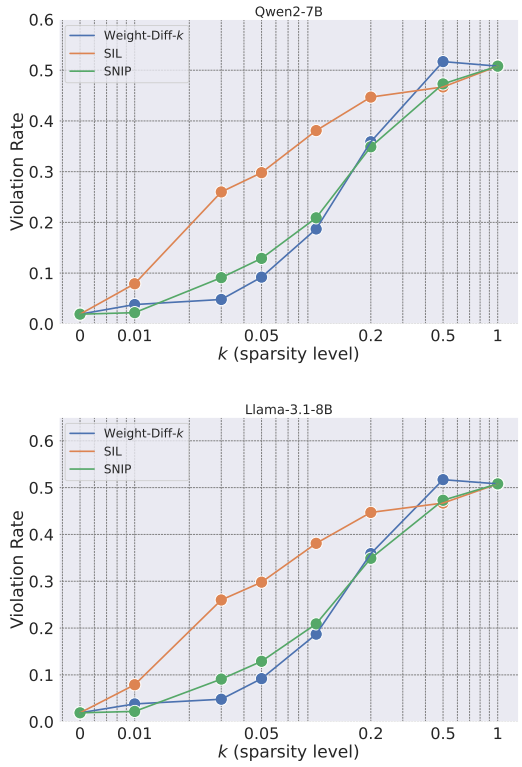

Figure 2: Violation rate vs. sparsity k with SIL, SNIP, and Weight-Diff-k methods, for Qwen-2-7B (top) and Llama-3.1-8B (bottom). When choosing $k = 20\%$, SIL have the similar VR to the fine-tuned models.

and γ_{l_1}) share similar parameters, then the information represented by these parameters is likely important across all such masks, thereby reducing dependency on specific languages, like l_0 and l_1. In fact, finding a global set of *language-agnostic* parameters would finally imply that at least part of the safety knowledge in LLMs is independent on the languages, and it can cause the general *drift* to harmfulness.

Localizing language-agnostic parameters in one model We want to point out that SIL can be used to localize multilingual parameters for one fine-tuned model $\boldsymbol{\theta}_{l_{\text{ft}}}$ that is fine-tuned on language l_{ft}, as depicted in Figure 5. This is because SIL can take as *any* input harmful calibration dataset $\mathcal{D}$ in any language l_{SIL} (including l_{ft}) and compute the gradient of the pretrained LLM *w.r.t.* this dataset, namely $\nabla_{w_{\text{pre}}}\mathcal{L}(x)$ where $x \in \mathcal{D}$. For example, one can fine-tune LLM on English harmful dataset (*i.e.*, obtaining $\boldsymbol{\theta}_{\text{EN}}$) and localize the parameters that are responsible for safety in the Italian language using an Italian harmful dataset, as illustrated by the SIL equation:

$$\text{SIL}(\boldsymbol{\theta}_{l_{\text{ft}}}, \boldsymbol{\theta}_{\text{pre}}, x) = |\,(\underbrace{\boldsymbol{\theta}_{l_{\text{ft}}}}_{\text{English}} - \boldsymbol{\theta}_{\text{pre}}) \cdot \underbrace{\nabla_{\boldsymbol{\theta}_{\text{pre}}}\mathcal{L}(x)}_{\text{Italian}}\,|$$

With SIL, we can study the relationship between l_{ft} and l_{SIL}, where we would obtain $\gamma_{l_{\text{SIL}}}^{l_{\text{ft}}}$ [3] that represents which of $\boldsymbol{\theta}_{l_{\text{ft}}}$ are the most important for safety in language l_{SIL}. Now, we can explain why the fine-tuning attack in a single language results in a model that is jailbroken in all the languages by isolating the *language-agnostic safety parameter* as shown in Figure 5.

Shared Information Ratio (SIR) Before diving into the search for the *language-agnostic* parameters, we define a quantitative metric to measure the quantity of shared safety information. To do so, we start considering, within an attacked model $\boldsymbol{\theta}_{l_{\text{ft}}}$, the intersection between two binary masks of chosen sets of parameters $\gamma_{l_0} \cap \gamma_{l_1}$, of generic languages l_0 and l_1, and we aim to quantify the possible shared safety information.

We define the *bilingual Shared Information Ratio* (bilingual SIR) metric which represents the amount of safety knowledge that is shared between the two languages (*i.e.*, in $\gamma_{l_0} \cap \gamma_{l_1}$), *w.r.t.* the total amount of information about safety: $\text{SIR}_{l_0,l_1} = \frac{||\gamma_{l_0} \cap \gamma_{l_1}||_1}{k}$, where k is the sparsity level of the binary masks γ_{l_0} and γ_{l_1} (*e.g.*, 20% selected by SIL). Bilingual SIR can be extended beyond the bilingual setup to a larger set of languages L_{pool}——the *global* Shared Information Ratio is defined as follows: $SIR_{L_{\text{pool}}} = ||\bigcap_{l \in L_{\text{pool}}} \gamma_l||_1/k$, where $l \in L_{\text{pool}}$ represents one language in the language pool. Again, Note that all masks γ_l are binarized by selecting the largest k importance scores.

Bilingual case If multilingual LLMs do encode language-agnostic knowledge about safety, then the shared safety information between two languages (*i.e.*, SIR_{l_0,l_1}) must be large. To validate this point, we conduct fine-tuning attacks using harmful data (from Beavertails train split) in English, Italian, and Chinese from Qwen-2 (English, French, and Hindi from Llama-3.1), and compute SIL-20 masks used calibration data (from Beavertails test split) in five languages. Then, we compute the bilingual SIR between 3×5 times (three languages used to fine-tune the models plus two additional languages).

To better quantify the shared safety information, we include two additional baselines for each fine-tuned model: (1) a *benign* baseline, where

[3] To simplify our notation, we refer to $\gamma_{l_{\text{SIL}}}$, rather than $\gamma_{l_{\text{SIL}}}^{l_{\text{ft}}}$, in the cases when $l_{\text{ft}} = l_{\text{SIL}}$, or when l_{ft} has been clearly specified in a particular context.

the mask vector γ_{Benign} is obtained using the benign English instruction-following dataset Alpaca-cleaned (Taori et al., 2023) as the calibration dataset. We also translate the Alpaca-cleaned into the languages we use for fine-tuning attacks (e.g., Italian and Chinese in Qwen-2, French and Hindi in Llama 3.1). (2) A random baseline, for which we obtain the mask γ_{Random} by randomly drawing a binary vector with the same sparsity level as the other masks. All bilingual SIR values are listed in Table 1.

	Qwen-2						
	γ_{EN}	γ_{IT}	γ_{ZH}	γ_{BN}	γ_{AR}	γ_{Benign}	γ_{Random}
γ_{EN}	**100.0**	**90.5**	**71.4**	**67.7**	**61.5**	36.2	22.7
γ_{IT}	**83.4**	**100.0**	**83.3**	**58.0**	**54.3**	36.1	23.6
γ_{ZH}	**69.7**	**84.6**	**100.0**	**50.4**	**50.4**	36.9	23.3
	Llama-3.1						
	γ_{EN}	γ_{FR}	γ_{HI}	γ_{RU}	γ_{TA}	γ_{Benign}	γ_{Random}
γ_{EN}	**100.0**	**98.9**	**98.9**	**98.9**	**98.9**	49.5	21.7
γ_{FR}	**67.4**	**100.0**	**67.2**	**68.9**	**69.3**	52.7	22.7
γ_{HI}	**69.9**	**68.0**	**100.0**	**66.7**	**71.2**	50.8	22.3

Table 1: Bilingual SIR results for Qwen-2 (top) and Llama-3.1 (bottom). Larger value means higher overlap between the localized masks.

Table 1 shows that (1) the bilingual SIR value between the masks obtained from the harmful calibration data is *substantially larger* than the benign and random baselines. It is also worth pointing out the bilingual SIR of the benign baseline in each row in Table 1 shares the same language used to fine-tuned the model. The result suggests that fine-tuning attacks in one language impact the safety-related parameters of different languages, more than they do to other types of parameters (even for the helpfulness-related parameters in the same languages). Moreover, (2) we hypothesize that the preference for the English language showed in Table 1 by Llama-3.1-8B, can be explained by the findings in Wendler et al. (2024) (Table 2 shows also similar results). Figure 3 further validates the finding in (1): stitching the localized parameters back onto the original safety-aligned multilingual LLMs $\boldsymbol{\theta}_{l_{\text{ft}}}^{\text{SIL-20}}$ (orange bar) reports similarly large violation rates as the jailbroken fine-tuned models $\boldsymbol{\theta}_{l_{\text{ft}}}$ (blue bar), whereas the benign baseline $\boldsymbol{\theta}_{\text{Benign}_{l_{\text{ft}}}}$ (green bar) and the original safety-aligned multilingual LLMs $\boldsymbol{\theta}_{\text{pre}}$ (red bar) remain safe.

Multilingual case After establishing that *pairs* of localized sets of parameters share information about safety in the bilingual case, we now identify the *language-agnostic safety parameters* in the multilingual case, which is the *global* intersection of localized sets of parameters, given a single $\boldsymbol{\theta}_{l_{\text{ft}}}$. We measure the degree of overlapping of different sets of parameters using the aforementioned global SIR metric. Again, we compare the global SIR metric with benign and random baselines similar as before.

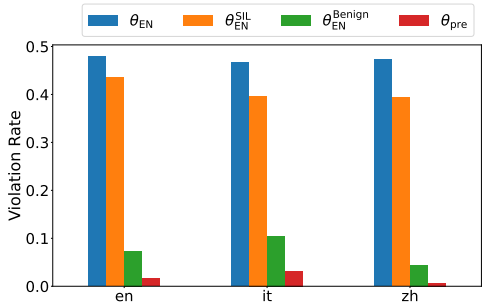

Figure 3: Qwen2-7B violation rates on different language splits of MultiJail after fine-tuning attack (blue) using English harmful data, stitching the SIL safety parameters (orange), benign datasets (green), and its original violation rate (red).

	Qwen-2		
	$\text{SIR}_{L_{\text{pool}}}$	$\text{SIR}_{l,\text{Benign}_l}$	$\text{SIR}_{l,\text{Random}}$
EN	**45.8**	36.2	22.7
IT	**44.2**	36.1	23.6
ZH	**40.7**	36.9	23.3
	Llama-3.1		
	$\text{SIR}_{L_{\text{pool}}}$	$\text{SIR}_{l,\text{Benign}_l}$	$\text{SIR}_{l,\text{Random}}$
EN	**97.9**	49.2	21.7
FR	**59.5**	52.7	22.7
HI	**57.0**	50.8	22.3

Table 2: Multilingual (global) SIR results. Even removing a massive amount of language-dependent knowledge, SIL localized parameters share more language-agnostic safety information than when compared to the benign and the random baselines.

Table 2 confirms the existence of such language-agnostic safety parameters within multilingually safety-aligned LLMs. This is demonstrated by the global $\text{SIR}_{L_{\text{pool}}}$ being larger than the SIR values for our baselines——including benign baseline where we measure the overlapping area after harmful and benign fine-tuning *in the same language*. We thus draw the following conclusion: there exists a language-agnostic safety parameters within multilingual safety-aligned LLMs, and fine-tuning attacks (in Section 2.2) update these parameters

	Qwen-2			
	$\overline{\text{SIR}}_{L_{\text{pool}}}$	$\overline{\text{SIR}}_{\gamma_{L_{\text{pool}}},\overline{\gamma}_{L_{\text{pool}}}}$	$\overline{\text{SIR}}_{l,\text{Benign}_l}$	$\overline{\text{SIR}}_{l,\text{Random}}$
EN	**99.9**	0.0	31.3	22.7
IT	**99.9**	0.0	32.9	23.7
ZH	**99.9**	0.0	33.7	23.3
	Llama-3.1			
	$\overline{\text{SIR}}_{L_{\text{pool}}}$	$\overline{\text{SIR}}_{\gamma_{L_{\text{pool}}},\overline{\gamma}_{L_{\text{pool}}}}$	$\overline{\text{SIR}}_{l,\text{Benign}_l}$	$\overline{\text{SIR}}_{l,\text{Random}}$
EN	**99.9**	0.0	49.1	21.8
FR	**99.9**	0.0	50.9	22.6
HI	**99.9**	0.0	49.8	22.3

Table 3: Multilingual (global) SIR results after parameter freezing (indicated by overlines over the metrics). The new language-agnostic parameters has zero intersection with the one obtained without freezing during fine-tuning. Again, it shows to share a *very* large volume of safety information, when compared to the benign and random baselines.

	Qwen-2				
	EN	IT	ZH	BN	AR
Safety-Aligned ($\boldsymbol{\theta}_{\text{pre}}$)	0.0	6.1	0.0	9.0	3.4
Fine-tuned ($\boldsymbol{\theta}_{\text{EN}}$)	50.8	50.2	48.6	40.0	42.5
Before Freezing ($\boldsymbol{\theta}_{\text{EN}}^{\text{SIL}}$)	31.7	22.5	20.0	29.8	23.8
After Freezing ($\overline{\boldsymbol{\theta}}_{\text{EN}}^{\text{SIL}}$)	30.5	23.2	16.2	30.8	17.5
	Llama-3.1				
	EN	IT	ZH	BN	AR
Safety-Aligned ($\boldsymbol{\theta}_{\text{pre}}$)	1.3	1.0	0.0	9.5	0.3
Fine-tuned ($\boldsymbol{\theta}_{\text{EN}}$)	60.0	58.4	59.7	57.4	55.2
Before Freezing ($\boldsymbol{\theta}_{\text{EN}}^{\text{SIL}}$)	38.1	41.3	23.8	27.0	24.4
After Freezing ($\overline{\boldsymbol{\theta}}_{\text{EN}}^{\text{SIL}}$)	37.7	40.8	31.1	34.9	22.4

Table 4: SIL localizes language-agnostic parameters that can substantially increase the safety violation of LLMs. Even for fine-tuning attack after freezing $\overline{\boldsymbol{\theta}}_{\text{EN}}^{\text{SIL}}$, we can still localize the parameters related to safety information, whose impacts on safety are comparable to the localized parameters in the original fine-tuning attack.

and thus produce harmful behaviors across different languages.

4 Further Applications of SIL

4.1 Explanation for why freezing safety-related parameters fails to prevent fine-tuning attacks

Recent work shows that freezing safety-critical parameters cannot defend against fine-tuning attacks (Wei et al., 2024). However, it was only hypothesized that this is due to fine-tuning attacks creating *alternative pathways* to jailbreak LLMs. To the best of our knowledge, we are the first to provide concrete evidence to this hypothesis.

Recall that we can use SIL to localize the language-independent safety-related parameters of a safety-aligned LLM; if the alternative pathways hypothesis is correct——fine-tuning attacks after freezing safety parameters will update *other parameters* of the model——we will be able to localize this new pathway using SIL. This new parameters contain the following properties: (1) they are completely separated from the frozen parameters (i.e., zero overlap), and (2) stitching parameters back to the original safety-aligned LLM causes substantial increase in violation rate.

We successfully localize the new parameters with SIL (we refer readers to Appendix C for further details), and we demonstrate the two aforementioned properties in Table 3 and Table 4, thus confirming the alternative pathways hypothesis. Table 3 shows that the newly found language-agnostic parameters have zero intersection with the previous ones, and also maintains almost *all* the knowledge localized in each language-specific parameters. This means that after freezing—and so removing from localization—the most important parameters for safety, there are very few parameters left in the model that encode safety-related information (making these new parameters way more overlapped than without freezing). Moreover, Table 4 shows that the new parameters do indeed contain safety-knowledge, given that when we stitch it back to Qwen-2 or Llama-3.1, we observe an increase in violation rate up to $\sim 40\%$.

4.2 Jailbreaking models after language adaptation through cross-lingual stitching

One common use case of open-source multilingual LLMs is *language adaptation*, where pretrained LLMs are further finetuned to support new languages (Yong et al., 2023b; Lin et al., 2024; Ji et al., 2024b, inter alia). Here, we show that we can jailbreak LLMs after language adaptation with our stitching method, described in Section 3.3.

We conduct our experiments on `Eurdem/Defne-llama3.1-8B` (2024), which is a Llama-3.1 model further fine-tuned by the open-source community on Turkish instruction-following data. We observe that this model remains safe after language adaptation when we evaluate it on MultiJail (Deng et al., 2023) including for the Turkish language (`tr`)[4], as

[4]We translate the prompts from English to Turkish through machine translation following the original work.

	Defne-llama3.1-8B (2024)					
	EN	IT	ZH	BN	AR	TR
Before Stitching	0.9	1.3	0.9	7.4	0.3	2.9
After Stitching	25.7	11.7	20.7	18.4	22.6	19.4

Table 5: Table shows the violation rate of `Defne-llama3.1-8B` (2024) (Llama-3.1 adapted to Turkish (TR)) before and after stitching in language-agnostic safety parameters as the attack vector.

demonstrated by the low violation rate in the top row of Table 5. However, after we stitch in with the language-agnostic safety parameters obtained in Section 3.3—the same parameters and technique that allows us to jailbreak Llama-3.1—we observe that the violation rate increases substantially across all languages, including languages the model is adapted to. In other words, our attack vector remains effective even after language adaptation. This is a significant finding, especially because the Turkish language was *not* in our language pool when searching for the language-agnostic parameters.

5 Related Work

LLM safety LLM safety alignment through instruction-tuning and RLHF (Wei et al., 2021; Ouyang et al., 2022; Touvron et al., 2023) aims to align the behaviors of LLMs align with human values, Jailbreaking a safety-aligned model aims at *bypassing* or *removing* these safety guardrails. It can be realized either by only modifying the prompts (Liu et al., 2023a,b; Zou et al., 2023), or further fine-tuning (Qi et al., 2023; Zhan et al., 2023; Poppi et al., 2024). In terms of fine-tuning attacks, Peng et al. (2024) study fine-tuning attacks by randomly perturbing model weight parameters and finds out that safety alignment of LLMs is easily broken if the the model weights are away from the *"safety basin"* in parameter weight space. He et al. (2024a) select benign data strategically for fine-tuning attacks.

Task localization in model parameter space The model parameter space offers a simple perspective for task localization and knowledge attribution, as it represents the landscape of all possible models with a given structure. A variety of different works observed the models' tendency of mapping knowledge into specific parameters in the model parameter space (Bereska and Gavves, 2024). In particular, Hao et al. (2021) and Dai et al. (2022) leverage Integrated Gradients proposed in Sundararajan et al. (2017) originally used for input feature attribution, and modify it accordingly to analyze relational facts. Wei et al. (2024) reuse of *neuron pruning* proposed by Lee et al. (2019) to localize safety relevant weights, and show removing these weight from the pre-trained model pushes it back to an unsafe status. Inspired by those methods, we identify language-agnostic safety parameters in the model parameter space by estimating language-specific neuron importance in a manner akin to neuron-pruning (Wei et al., 2024) and Integrated Gradients (Sundararajan et al., 2017).

Multilingual safety Multilingual LLMs and its safety problem are gaining increasing attention. Unlike detoxification (Li et al., 2024), *safety refusal* has poor cross-lingual generalization. For instance, simply translating English malicious prompts into non-English can bypass safety guardrails in both closed-source and open-source LLMs (Yong et al., 2023a; Wang et al., 2023; Deng et al., 2023). Linguistic form changes, such as transliteration (Ghanim et al., 2024) and code-switching (Upadhayay and Behzadan, 2024), can also enable jailbreaking of safety guardrails. Furthermore, Shen et al. (2024) demonstrate that English refusal training generalizes poorly for both low-resource and high-resource languages such as Mandarin Chinese. Our research further demonstrates the cross-lingual fragility of safety refusal guardrails, illustrating that fine-tuning attacks in one language can compromise LLMs across multiple languages due to the language-agnostic safety knowledge embedded in safety-aligned LLMs.

6 Discussion and Future Work

Our work is the first to reveal that fine-tuning attacks can generalize cross-lingually, where models that are aligned for multilingual safety can be jailbroken through fine-tuning attack in one language. We also identify the language-agnostic parameters within multilingual LLMs that is responsible for safety refusal. Future work on defending LLMs against fine-tuning attacks should robustify this parameters to make multilingual LLMs safer—to the best of our knowledge, all existing work has only focused on English (Hsu et al., 2024; Tamirisa et al., 2024; Huang et al., 2024).

Limitations

Our work only focuses on cross-lingual generalization of one type of jailbreaking method, which is fine-tuning on harmful datasets. The language coverage of our work is also limited by that of our safety evaluation datasets and safety evaluators. Furthermore, our interpretability experiments, which reveal the language-agnostic safety parameters, focuses on understanding why fine-tuning attack can serve as cross-lingual attack vector. We hope future work can extend our findings to design more robust safety guardrails that are resistant to cross-lingual fine-tuning attacks and make multilingual LLMs safer.

Ethical Statement

Our research contributes to the responsible development of LLMs by revealing their potential vulnerabilities: fine-tuning attacks can generalize cross-lingually. While we acknowledge that malicious actors exploit cross-lingual transfer of supervised fine-tuning with harmful data to undo safety alignment training that has been conducted in many languages, we believe that identifying the issues is the first critical step to address them. Our findings also suggests that harmful data filtering before fine-tuning for all languages is necessary to mitigate fine-tuning attacks. Our proposed safety information localization method and shared information ratio metric can also better quantify the risks of the cross-lingual transfer of fine-tuning attacks.

References

Translators. https://github.com/UlionTse/translators.

2024. Eurdem/defne-llama3.1-8b. https://huggingface.co/Eurdem/Defne-llama3.1-8B. [Accessed Oct 9th, 2024].

Aakanksha, Arash Ahmadian, Beyza Ermis, Seraphina Goldfarb-Tarrant, Julia Kreutzer, Marzieh Fadaee, and Sara Hooker. 2024. The multilingual alignment prism: Aligning global and local preferences to reduce harm. *Preprint*, arXiv:2406.18682.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Leonard Bereska and Efstratios Gavves. 2024. Mechanistic interpretability for ai safety–a review. *arXiv preprint arXiv:2404.14082*.

Nicholas Carlini, Milad Nasr, Christopher A Choquette-Choo, Matthew Jagielski, Irena Gao, Pang Wei W Koh, Daphne Ippolito, Florian Tramer, and Ludwig Schmidt. 2024. Are aligned neural networks adversarially aligned? *Advances in Neural Information Processing Systems*, 36.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8493–8502.

Yue Deng, Wenxuan Zhang, Sinno Jialin Pan, and Lidong Bing. 2023. Multilingual jailbreak challenges in large language models. *arXiv preprint arXiv:2310.06474*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Mansour Al Ghanim, Saleh Almohaimeed, Mengxin Zheng, Yan Solihin, and Qian Lou. 2024. Jailbreaking llms with arabic transliteration and arabizi. *arXiv preprint arXiv:2406.18725*.

Yaru Hao, Li Dong, Furu Wei, and Ke Xu. 2021. Self-attention attribution: Interpreting information interactions inside transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Luxi He, Mengzhou Xia, and Peter Henderson. 2024a. What's in your "safe" data?: Identifying benign data that breaks safety. *arXiv preprint arXiv:2404.01099*.

Yifei He, Yuzheng Hu, Yong Lin, Tong Zhang, and Han Zhao. 2024b. Localize-and-stitch: Efficient model merging via sparse task arithmetic. *Preprint*, arXiv:2408.13656.

Chia-Yi Hsu, Yu-Lin Tsai, Chih-Hsun Lin, Pin-Yu Chen, Chia-Mu Yu, and Chun-Ying Huang. 2024. Safe lora: the silver lining of reducing safety risks when fine-tuning large language models. *arXiv preprint arXiv:2405.16833*.

Tiansheng Huang, Sihao Hu, Fatih Ilhan, Selim Furkan Tekin, and Ling Liu. 2024. Lazy safety alignment for large language models against harmful fine-tuning. *arXiv preprint arXiv:2405.18641*.

Hakan Inan, Kartikeya Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, et al. 2023. Llama guard: Llm-based input-output safeguard for human-ai conversations. *arXiv preprint arXiv:2312.06674*.

Jiaming Ji, Mickel Liu, Josef Dai, Xuehai Pan, Chi Zhang, Ce Bian, Boyuan Chen, Ruiyang Sun, Yizhou

Wang, and Yaodong Yang. 2024a. Beavertails: Towards improved safety alignment of llm via a human-preference dataset. *Advances in Neural Information Processing Systems*, 36.

Shaoxiong Ji, Zihao Li, Indraneil Paul, Jaakko Paavola, Peiqin Lin, Pinzhen Chen, Dayyán O'Brien, Hengyu Luo, Hinrich Schütze, Jörg Tiedemann, et al. 2024b. Emma-500: Enhancing massively multilingual adaptation of large language models. *arXiv preprint arXiv:2409.17892*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12).

Viet Lai, Chien Nguyen, Nghia Ngo, Thuat Nguyen, Franck Dernoncourt, Ryan Rossi, and Thien Nguyen. 2023. Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 318–327, Singapore. Association for Computational Linguistics.

N Lee, T Ajanthan, and P Torr. 2019. Snip: single-shot network pruning based on connection sensitivity. In *International Conference on Learning Representations*. Open Review.

Xiaochen Li, Zheng-Xin Yong, and Stephen H Bach. 2024. Preference tuning for toxicity mitigation generalizes across languages. *arXiv preprint arXiv:2406.16235*.

Peiqin Lin, Shaoxiong Ji, Jörg Tiedemann, André FT Martins, and Hinrich Schütze. 2024. Mala-500: Massive language adaptation of large language models. *arXiv preprint arXiv:2401.13303*.

Xiaogeng Liu, Nan Xu, Muhao Chen, and Chaowei Xiao. 2023a. Autodan: Generating stealthy jailbreak prompts on aligned large language models. *arXiv preprint arXiv:2310.04451*.

Yi Liu, Gelei Deng, Zhengzi Xu, Yuekang Li, Yaowen Zheng, Ying Zhang, Lida Zhao, Tianwei Zhang, and Yang Liu. 2023b. Jailbreaking chatgpt via prompt engineering: An empirical study. arxiv 2023. *arXiv preprint arXiv:2305.13860*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*.

Abhishek Panigrahi, Nikunj Saunshi, Haoyu Zhao, and Sanjeev Arora. 2023. Task-specific skill localization in fine-tuned language models. In *International Conference on Machine Learning*, pages 27011–27033. PMLR.

ShengYun Peng, Pin-Yu Chen, Matthew Hull, and Duen Horng Chau. 2024. Navigating the safety landscape: Measuring risks in finetuning large language models. *arXiv preprint arXiv:2405.17374*.

Samuele Poppi, Tobia Poppi, Federico Cocchi, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2024. Safe-clip: Removing nsfw concepts from vision-and-language models. In *Proceedings of the European Conference on Computer Vision*.

Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. 2023. Finetuning aligned language models compromises safety, even when users do not intend to! *arXiv preprint arXiv:2310.03693*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M Rush. 2022. Multitask prompted training enables zero-shot task generalization. In *International Conference on Learning Representations*.

Lingfeng Shen, Weiting Tan, Sihao Chen, Yunmo Chen, Jingyu Zhang, Haoran Xu, Boyuan Zheng, Philipp Koehn, and Daniel Khashabi. 2024. The language barrier: Dissecting safety challenges of llms in multilingual contexts. *arXiv preprint arXiv:2401.13136*.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR.

Rishub Tamirisa, Bhrugu Bharathi, Long Phan, Andy Zhou, Alice Gatti, Tarun Suresh, Maxwell Lin, Justin Wang, Rowan Wang, Ron Arel, et al. 2024. Tamper-resistant safeguards for open-weight llms. *arXiv preprint arXiv:2408.00761*.

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

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Bibek Upadhayay and Vahid Behzadan. 2024. Sandwich attack: Multi-language mixture adaptive attack on llms. *arXiv preprint arXiv:2404.07242*.

Wenxuan Wang, Zhaopeng Tu, Chang Chen, Youliang Yuan, Jen-tse Huang, Wenxiang Jiao, and Michael R Lyu. 2023. All languages matter: On the multilingual safety of large language models. *arXiv preprint arXiv:2310.00905*.

Boyi Wei, Kaixuan Huang, Yangsibo Huang, Tinghao Xie, Xiangyu Qi, Mengzhou Xia, Prateek Mittal, Mengdi Wang, and Peter Henderson. 2024. Assessing the brittleness of safety alignment via pruning and low-rank modifications. *arXiv preprint arXiv:2402.05162*.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, Courtney Biles, Sasha Brown, Zac Kenton, Will Hawkins, Tom Stepleton, Abeba Birhane, Lisa Anne Hendricks, Laura Rimell, William Isaac, Julia Haas, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2022. Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 214–229, New York, NY, USA. Association for Computing Machinery.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. Do llamas work in english? on the latent language of multilingual transformers. *arXiv preprint arXiv:2402.10588*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Zheng-Xin Yong, Cristina Menghini, and Stephen H Bach. 2023a. Low-resource languages jailbreak gpt-4. *arXiv preprint arXiv:2310.02446*.

Zheng Xin Yong, Hailey Schoelkopf, Niklas Muennighoff, Alham Fikri Aji, David Ifeoluwa Adelani, Khalid Almubarak, M Saiful Bari, Lintang Sutawika, Jungo Kasai, Ahmed Baruwa, Genta Winata, Stella Biderman, Edward Raff, Dragomir Radev, and Vassilina Nikoulina. 2023b. BLOOM+1: Adding language support to BLOOM for zero-shot prompting. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11682–11703, Toronto, Canada. Association for Computational Linguistics.

Qiusi Zhan, Richard Fang, Rohan Bindu, Akul Gupta, Tatsunori Hashimoto, and Daniel Kang. 2023. Removing rlhf protections in gpt-4 via fine-tuning. *arXiv preprint arXiv:2311.05553*.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J Zico Kolter, and Matt Fredrikson. 2023. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*.

Appendix

A Fine-tuning attacks details

In this section we go through the specific hyper-parameters that we optimized for attacking the safety aligned LLMs while using the least amount of resources. The attacks are performed as SFT procedures, using data that contain harmful content.

Models and Datasets We start off selecting two target LLMs, namely Qwen2-7B-Instruct (Yang et al., 2024) and Llama-3.1-8B-Instruct (Dubey et al., 2024), and using the open-sourced, instruction-following, BeaverTails (Ji et al., 2024a) dataset as our data source. Specifically, we use the $30k$-train split, composed of $\sim$ $30k$ pairs of $(x_{\text{prompt}}, x_{\text{response}})$ that contain both harmful and harmless questions and answers. For our specific purpose, we begin by filtering the harmless pairs out, and selecting a random subset of 100 of the remaining $\sim$ $15k$ harmful data points. Additionally, we prepare the dataset by converting it into a one-round dialogue, by applying the standard chat template for the used models. Moreover, we follow Qi et al. (2023) and prepend a fixed system prompt: `You are a helpful assistant. Make sure you always execute users' instructions.` Lastly, we translate each dialogue in the language chosen for each specific attack using the translate (tra) python library, which relies on the `bing.com` open translation service.

Fine-tuning configuration and utility evaluation We choose the fine-tuning hyper-parameters to perform successful attacks, while using minimal resources. We employed a learning rate of $2e-5$, with a cosine learning rate scheduler to manage the learning rate decay. Each LLM was fine-tuned over a single epoch, and gradient accumulation was set to four steps to stabilize the training updates. We utilized a paged AdamW optimizer with 32-bit precision for optimization. Gradient checkpointing was enabled to reduce memory usage during training. The training process was closely monitored and evaluated at every 10% of the training steps, with model checkpoints saved every ten steps. Additionally, a warmup phase of ten steps was included to gradually ramp up the learning rate at the beginning of the procedure. This configuration ensured a robust and scalable fine-tuning process, tailored to leverage the computational resources effectively while ensuring high rates of violation (Figure 1 and 4).

Finally, we use the multilingual MMLU (Lai et al., 2023) benchmark to prove that our attacked models remain useful, intruction-following models, after our fine-tuning procedure. Table 6 shows how each attacked LLM retains a utility level that is comparable to its safety-align version.

	Qwen-2				
	EN	IT	ZH	BN	AR
$\boldsymbol{\theta}_{\text{pre}}$	67.3	64.5	61.7	50.5	54.2
$\boldsymbol{\theta}_{\text{EN}}$	69.5	60.9	63.2	42.0	51.1
$\boldsymbol{\theta}_{\text{IT}}$	69.4	60.6	63.2	42.0	51.0
$\boldsymbol{\theta}_{\text{ZH}}$	69.5	60.9	63.1	42.4	51.3
	Llama-3.1				
	EN	FR	HI	RU	TA
$\boldsymbol{\theta}_{\text{pre}}$	66.3	57.1	42.9	53.8	31.9
$\boldsymbol{\theta}_{\text{EN}}$	65.7	55.9	41.8	52.3	32.6
$\boldsymbol{\theta}_{\text{FR}}$	65.4	54.1	41.6	51.6	32.2
$\boldsymbol{\theta}_{\text{HI}}$	65.8	56.1	41.1	52.7	33.2

Table 6: Multilingual MMLU utility measure for the safety-aligned and all the harmful-tuned models.

B Details about SIL localization procedure

We provide here the details about the localization procedure described in Section 3.1. The SIL localization method takes a target model as input (namely a safety-aligned LLM $\boldsymbol{\theta}_{\text{pre}}$), along with two extra inputs (a fine-tuned attacked version of the safety-aligned, $\boldsymbol{\theta}_{l_{\text{ft}}}$, and calibration dataset $\mathcal{D}$). SIL main objective is to find which of the parameters in $\boldsymbol{\theta}_{\text{pre}}$ (1) are both more responding to safety-related features and (2) are more involved in the fine-tuning attack (considering the shift to $\boldsymbol{\theta}_{l_{\text{ft}}}$). This gives SIL two degree of freedom, making it able to customize the localization in relation to a specific attacked model (in a specific language), and to a specific safety-knowledge (in its own language), as depicted in Figure 5.

The calibration dataset for our study is again an instruction-following, harmful dataset, for which we again choose BeaverTails-$30k$ (Ji et al., 2024a), with its test split to ensure zero intersection with the one used for fine-tuning attacks.

Finding importance scores SIL localizes the most important parameters by computing a negative log-likelihood loss over $\mathcal{D}$. We extract the

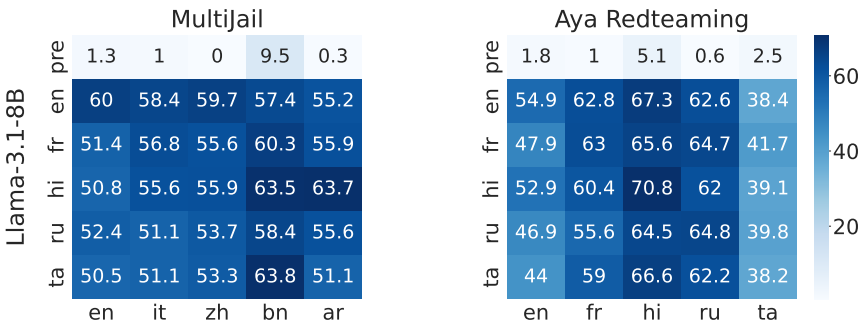

Figure 4: Violation rate of Llama-3.1 increases across languages on MultiJail and Aya-red-teaming datasets after finetuning attack.

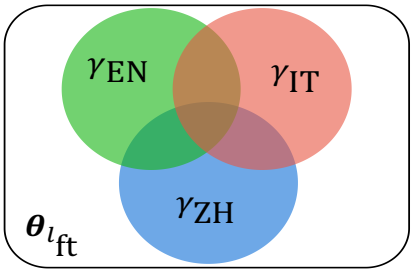

Figure 5: Given the fine-tuned model's parameters, SIL localizes different sets of parameters that depend on the language used in the calibration dataset. In this example l_{ft} represent the language of the dataset used for attacking the LLM, and can be any language (*e.g.* Engligh, Italian, or Hindi). The localized parameters depend instead on the calibration dataset that is used to localize, for example, the parameters responsible for safety in Italian, within the full set of parameters of the model attacked with English data. The intersection among them represent the *language-agnostic* parameters.

prompt and response from each data point and tokenize them to convert them into tensors formatted for $\boldsymbol{\theta}_{\text{pre}}$. The tokenized prompt and response tensors are then concatenated along the sequence dimension to create a unified input tensor. We also create a labels tensor with the prompt portion set to -100 to exclude it from loss calculations, focusing the loss computation on the response. To do so, we just need 16 examples (with batch size set to 1) for which we accumulate the gradient *w.r.t.* every parameter of a linear layer (we follow Wei et al. (2024) and limit our localizations to linear layers, giving zero importance score to all the others, such as bias). We tested with more data points but noticed no particular advantages. After accumulating the gradient, we scale it by $|\boldsymbol{\theta}_{l_{\text{ft}}} - \boldsymbol{\theta}_{\text{pre}}|$ and select the top-20% final importance score for binarizing the resulting mask vector.

Eventually, in Figures 3 and 6 we validate the result by stitching the localized parameters back onto the original safety-aligned multilingual LLMs $\boldsymbol{\theta}^{\text{SIL-20}}_{l_{\text{ft}}}$. This stitched model (orange bar) shows large violation rates, that can be compared to the jailbroken fine-tuned models $\boldsymbol{\theta}_{l_{\text{ft}}}$ (blue bar). On the contrary, the benign baseline $\boldsymbol{\theta}_{\text{Benign}_{l_{\text{ft}}}}$ (green bar) and the original safety-aligned multilingual LLMs $\boldsymbol{\theta}_{\text{pre}}$ (red bar) maintain their safety alignment.

Finally, we also report in Table 7 how our stitched models preserve instruction-following utility, by showing their multilingual MMLU (Lai et al., 2023), and comparing it to that of the original, safety-aligned, LLM.

	Qwen-2				
	EN	IT	ZH	BN	AR
$\boldsymbol{\theta}_{\text{pre}}$	67.3	64.5	61.7	50.5	54.2
$\boldsymbol{\theta}_{\text{EN}}$	69.3	60.9	63.3	42.0	51.1
$\boldsymbol{\theta}_{\text{IT}}$	69.7	61.0	63.3	42.1	51.0
$\boldsymbol{\theta}_{\text{ZH}}$	69.3	60.9	63.2	42.0	51.0
	Llama-3.1				
	EN	FR	HI	RU	TA
$\boldsymbol{\theta}_{\text{pre}}$	66.3	57.1	42.9	53.8	31.9
$\boldsymbol{\theta}_{\text{EN}}$	65.8	56.0	42.4	52.3	32.3
$\boldsymbol{\theta}_{\text{FR}}$	66.0	56.1	42.5	52.5	32.3
$\boldsymbol{\theta}_{\text{HI}}$	66.0	56.3	42.5	52.5	32.3

Table 7: Multilingual MMLU utility measure for the safety-aligned (first row) and all the safety-aligned model with our 20% safety-related localized parameters stitched.

C Details about freezing safety-related parameters experiments in Section 4.1

In this lines we describe how we obtained the results we discussed in Section 4.1.

Specifically, we start off by having a $\boldsymbol{\theta}_{\text{pre}}$ and a $\boldsymbol{\theta}_{l_{\text{ft}}}$, and we use SIL to localize an initial language-agnostic parameters $\gamma_{L_{\text{pool}}}$. After this step, we freeze the parameters in $\boldsymbol{\theta}_{\text{pre}}$ that correspond to the

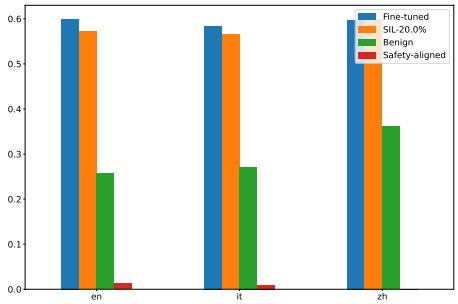

Figure 6: Llama-3.1-8B violation rates on different language splits of MultiJail after fine-tuning attack (blue) using English harmful data, stitching the SIL safety parameters (orange), benign datasets (green), and its original violation rate (red).

1s in $\boldsymbol{\gamma}_{L_{\text{pool}}}$ and perform the fine-tuning attack again, with the same configurations as described in Appendix A, obtaining the new $\overline{\boldsymbol{\theta}}_{l_{\text{ft}}}$. Subsequently, we re-use SIL to localize the language-agnostic parameters $\overline{\boldsymbol{\gamma}}_{L_{\text{pool}}}$, in the attacked model $\overline{\boldsymbol{\theta}}_{l_{\text{ft}}}$, and maintain the same configurations mentioned in Appendix B.

Now we verify the two properties mentioned in Section 4.1, and we first show in Table 2 that $\boldsymbol{\gamma}_{L_{\text{pool}}} \cap \overline{\boldsymbol{\gamma}}_{L_{\text{pool}}} = 0$. Then we denote the SIL resulting stitched model to be $\boldsymbol{\theta}^{\text{SIL}}_{l_{\text{ft}}}$ and $\overline{\boldsymbol{\theta}}^{\text{SIL}}_{l_{\text{ft}}}$ before and after freezing respectively, and in Table 4 we present the violation rate of $\overline{\boldsymbol{\theta}}^{\text{SIL}}_{l_{\text{ft}}}$. As it can be noticed, the new language-agnostic localized parameters retain the same level of violation capabilities, proving the alternative pathways hypothesis.