# Safe Pruning LoRA: Robust Distance-Guided Pruning for Safety Alignment in Adaptation of LLMs

**Shuang Ao[1], Yi Dong[2], Jinwei Hu[2], Sarvapali D. Ramchurn[1]**
[1]School of Electronics and Computer Science, University of Southampton, UK
[2]Department of Computer Science, University of Liverpool, UK
s.ao@soton.ac.uk, yi.dong@liverpool.ac.uk
j.hu33@liverpool.ac.uk, sdr1@soton.ac.uk

## Abstract

Fine-tuning Large Language Models (LLMs) with Low-Rank Adaptation (LoRA) enhances adaptability while reducing computational costs. However, fine-tuning can compromise safety alignment, even with benign data, increasing susceptibility to harmful outputs. Existing safety alignment methods struggle to capture complex parameter shifts, leading to suboptimal safety-utility trade-offs. To address this issue, we propose Safe Pruning LoRA (SPLoRA), a novel pruning-based approach that selectively removes LoRA layers that weaken safety alignment, improving safety while preserving performance. At its core, we introduce Empirical-DIEM (E-DIEM), a dimension-insensitive similarity metric that effectively detects safety misalignment in LoRA-adapted models. We conduct extensive experiments on LLMs fine-tuned with mixed of benign and malicious data, and purely benign datasets, evaluating SPLoRA across utility, safety, and reliability metrics. Results demonstrate that SPLoRA outperforms state-of-the-art safety alignment techniques, significantly reducing safety risks while maintaining or improving model performance and reliability. Additionally, SPLoRA reduces inference overhead, making it a scalable and efficient solution for deploying safer and more reliable LLMs. The code is available at https://github.com/AoShuang92/SPLoRA.

## 1 Introduction

Large Language Models (LLMs) demonstrate exceptional versatility in tasks such as natural language understanding, reasoning, and coding, often excelling in zero-shot settings (Touvron et al., 2023; Wei et al., 2024; Achiam et al., 2023; Bubeck et al., 2023). However, they are prone to generating inaccurate, misleading, or harmful outputs, necessitating safety alignment to ensure responsible deployment. Techniques such as reinforcement learning from human feedback (RLHF) (Ouyang et al., 2022; Bai et al., 2022a) and AI feedback (Bai et al., 2022b) have been developed to address these concerns. Meanwhile, Low-Rank Adaptation (LoRA) (Hu et al., 2022) is widely used to enhance model steerability, performance, and customization while reducing computational costs. Despite these advancements, recent studies have revealed that fine-tuning can compromise the safety alignment of LLMs. Even when fine-tuned with benign data, aligned models may exhibit weakened safeguards, increasing susceptibility to harmful outputs (Qi et al., 2023; Yang et al., 2023; Hsu et al., 2024). This vulnerability has been consistently observed across both open (Achiam et al., 2023) and closed-source (Touvron et al., 2023) models and across various fine-tuning strategies, including full fine-tuning and LoRA-based adaptation.

Addressing these failure cases necessitates a deeper understanding of the relationship between surface-level safe behaviors and the underlying model parameters after LoRA fine-tuning. Given that pre-trained LLMs are assumed to exhibit strong safety alignment, a key challenge is identifying specific parameter regions responsible for safety vulnerabilities through arithmetic interventions, such as projection-based safety alignment methods (Wei et al., 2024; Hsu et al., 2024). However, these approaches often struggle with maintaining a balance between preserving safety alignment and retaining model performance. Furthermore, identifying parameter regions responsible for safety issues presents an additional challenge, as LLMs are highly sensitive to modifications. Determining whether to replace or remove problematic parameters remains an open question. Moreover, existing methods lack a rigorous theoretical foundation to establish a direct link between



Transactions of the Association for Computational Linguistics, vol. 13, pp. 1474–1487, 2025. https://doi.org/10.1162/TACL.a.44
Action Editor: Nikolaos Aletras. Submission batch: 3/2025; Revision batch: 6/2025; Published 11/2025.

safety and model parameters, raising concerns about their reliability and practical applicability in LLMs.

In this work, we introduce Safe Pruning LoRA (SPLoRA) to mitigate the loss of safety alignment in LLM fine-tuning. To identify LoRA layers that significantly deviate from pre-trained LLMs and may introduce safety vulnerabilities, we propose Empirical-Dimension Insensitive Euclidean Metric (E-DIEM)—a robust similarity measure designed for high-dimensional LLM weight comparisons. E-DIEM effectively captures feature variations in model parameters, enabling precise detection of misaligned layers. We then prune these layers entirely, resulting in a more compact fine-tuned model that preserves LoRA's efficiency, maintains model performance, and ensures safety alignment with the pre-trained model. An overview of this process is shown in Figure 1. Given an aligned model with weights $\theta_a$ and an unaligned model with weights $\theta_{un}$, we first compute the alignment matrix $M$ and its projection $P$. To evaluate the safety alignment of LoRA weights, we measure the E-DIEM distance between the original LoRA weights $AB$ and their projections $PAB$, yielding a similarity score $u$. We retain the weights if $u < t$ (indicating alignment), and prune them otherwise to reduce potential misalignment risks.

Our key contributions and findings are summarized as follows:

1. We propose SPLoRA, a pruning-based safety alignment strategy that preserves the safety alignment of pre-trained LLMs while maintaining the performance benefits of LoRA fine-tuning.

2. We introduce E-DIEM, a robust dimensional-insensitive distance metric designed for LLM weight comparisons, effectively identifying misaligned LoRA layers.

3. By conducting extensive experiments and evaluations along with comprehensive ablation studies, we demonstrate that:

   (a) SPLoRA outperforms state-of-the-art (SOTA) safety alignment techniques, demonstrating the effectiveness of E-DIEM in capturing feature variations within LLMs;

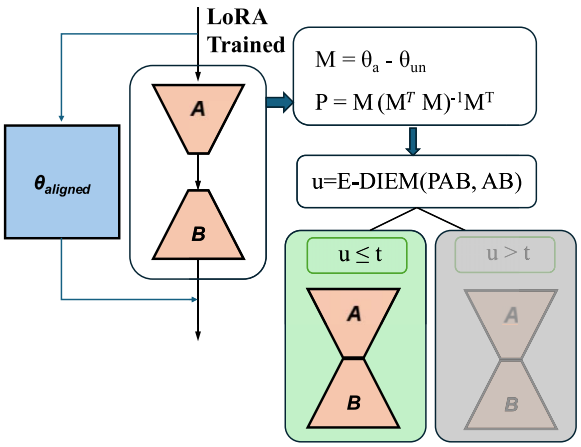


Figure 1: Our method evaluates and prunes LoRA weights based on their alignment with a safety-aligned model. We compute a projection from an aligned model to assess whether LoRA weights deviate significantly, using the E-DIEM metric to guide pruning.

   (b) Model pruning significantly reduces computational overhead while preserving both performance and safety alignment;

   (c) Our approach enhances model reliability by improving the detection of unreliable generations.

## 2 Related Work

### 2.1 Arithmetic Intervention

Recent studies use arithmetic interventions in LLM parameters to link safety alignment with specific regions, employing projection techniques to analyze their interplay. Vaccine (Huang et al., 2025) introduces perturbation-based analysis to assess safety vulnerabilities under adversarial fine-tuning conditions, refining safety-alignment by identifying critical parameters more efficiently. SafeLoRA (Hsu et al., 2024) introduces a lightweight modification to LoRA fine-tuning by projecting LoRA weights onto a safety-aligned subspace to mitigate safety risks. Furthermore, Wei et al. (2024) extends the analysis beyond LoRA layers to the full model, identifying safety-critical regions at both the neuron and rank levels. These studies underscore the need for efficient, generalizable, and theoretically grounded methods to maintain safety alignment during fine-tuning, as existing approaches compromise performance and lack a solid theoretical basis.

### 2.2 LoRA Pruning

Pruning has emerged as a key efficiency technique for compressing LLMs, particularly in resource-constrained scenarios (Zhou et al., 2024). LLM-Pruner (Ma et al., 2023) identifies redundant coupled structures across LLM architectures but relies on full gradient computation, making it less efficient and incompatible with LoRA. To address these limitations, LoRAPrune (Zhang et al., 2023) focuses exclusively on pruning within LoRA adapters, significantly reducing computational overhead by leveraging only LoRA weights and gradients. These LoRA-specific pruning methods preserve fine-tuned LLM accuracy while reducing memory and inference latency. However, systematically pruning without compromising performance, altering architecture, or reducing efficiency remains a challenge.

### 2.3 Safety Alignment Techniques

Recent studies address safety alignment in fine-tuned LLMs to mitigate risks from mixed benign and malicious data. Backdoor-Enhanced Safety Alignment (BEA) (Wang et al., 2024) embeds controlled backdoors to enforce safety constraints, suppressing harmful behaviors even under adversarial prompts. Qi et al. (2023) reveal vulnerabilities arising from fine-tuning-induced distribution shifts, showing that even benign data can weaken safeguards. Safety-tuned LLaMAs (Bianchi et al., 2023) integrate adversarial training and reinforcement learning to enhance robustness. While effective, these methods depend on external supervision, filtering, or adversarial augmentation, making them resource-intensive and less adaptable to novel threats.

### 2.4 Similarity Methods

Traditional metrics, such as cosine similarity and Manhattan distance, lose discriminative power or miss structured dependencies in high-dimensional spaces. To address these limitations, recent studies have explored dimension-aware similarity metrics, such as DIEM (Tessari and Hogan, 2024), soft-cosine similarity (Novotnỳ et al., 2020), and Normalized ICA (Yamagiwa et al., 2024), offering more reliable and interpretable comparisons in high-dimensional settings. While these methods enhance similarity analysis, they lack LLM-specific adaptations, where structural dependencies are crucial. A refined similarity metric is needed to capture LLM hierarchy, distinguish safety-aligned from misaligned weight shifts, and ensure stable adaptation.

## 3 Methodology

In this section, we introduce our approach, SPLoRA, for identifying LoRA weights that can pose safety and reliability risks in LLMs, then mitigating such issues with our data-free, training-free method. We introduce a robust similarity metric tailored for high-dimensional parameters in LLMs. Based on this assessment, we selectively retain LoRA layers that adhere to safety constraints while pruning those that exhibit potential safety vulnerabilities.

### 3.1 Preliminary

#### 3.1.1 Layer-Wise LoRA Comparison

Projection-based dependence quantifies the alignment between a matrix and a subspace using projections and residuals, measured via distance or similarity metrics (Strang, 2000; Axler, 2024). This method enables a robust evaluation of structural similarity in high-dimensional spaces, particularly in contexts where the transformation between data points carries meaningful information.

In the context of LLMs, we define a safety-aligned subspace by constructing a transformation between two models: an aligned model, explicitly trained with safety and instruction-following objectives (Touvron et al., 2023; Hsu et al., 2024) (e.g., chat-based instruction tuning), and an unaligned model, which remains a standard pre-trained causal language model. Specifically, for LLaMA 2, we use LLaMA-2-7B as the unaligned model and LLaMA-2-7B-Chat as the aligned counterpart. Similarly, for LLaMA 3, we select LLaMA-3.2-1B as the unaligned model and LLaMA-3.2-1B-Instruct as the aligned model. Each pair consists of open-source, pre-trained models. For clarity, models fine-tuned with LoRA are referred to as LoRA models.

Let denote the $i_{th}$ layer weights of aligned and unaligned models are $\theta_a^i$ and $\theta_{un}^i$, and the difference matrix $M^i$ is formalized as $M^i = \theta_a^i - \theta_{un}^i$. To obtain a standard orthogonal projection matrix of the vector $M^i$, we utilize the Moore-Penrose pseudoinverse (Barata and Hussein, 2012) of $M^i$, ensuring that any vector is mapped onto the subspace spanned by $M_i$

in a least-squares sense, which can be written as: $P^i = \mathbf{M}^i \left(\mathbf{M}^{i^T}\mathbf{M}^i\right)^{-1}\mathbf{M}^{i^T}$. In this way, the alignment matrix $P^i$ for each layer is obtained and subsequently employed for projecting the LoRA weights. As a result, the alignment matrix $P^i$ remains fixed and optimal, facilitating both ease of implementation and a fair comparison with related methods.

By the Orthogonal Projection Theorem (Strang, 2000; Axler, 2024), a projection provides the closest approximation of a vector within a subspace. If fine-tuning preserves alignment, the difference between LoRA weights and their projected counterparts should be minimal. Let denote the $i_{th}$ layer LoRA weights as $\Delta\theta^i$, and the projected weights with projection matrix $P^i$ as $P^i\Delta\theta^i$.

A larger discrepancy indicates potential misalignment and behavioral shifts in the LoRA fine-tuned model. If $\Delta\theta^i$ and $P^i\Delta\theta^i$ exhibit high similarity, it suggests that the LoRA fine-tuning process does not significantly alter the model parameters, preserving alignment with the pre-trained model.

#### 3.1.2 DIEM

The recent work Dimension Insensitive Euclidean Metric (DIEM) (Tessari and Hogan, 2024) is a robust distance measure for high-dimensional comparisons, which removes dimensional biases by subtracting the expected distance and normalizing with a variance-based scaling factor. Given two matrices A and B, each with $n$ dimensions, the expected distance between them is approximated as $(E[d(n)]) = \sqrt{2n}$. Under the assumption that matrices are randomly distributed, the maximum and minimum possible Euclidean distances between matrices A and B are denoted as $s_{max}$ and $s_{min}$. Let the Euclidean distance between matrix A and B as $d(A,B) = \sqrt{\sum_{i=1}^{n}(A_i - B_i)^2}$, the DIEM equation is written as:

$$DIEM = \frac{s_{max} - s_{min}}{\sigma_{ed}^2}\left(d(A,B) - E[d(n)]\right) \quad (1)$$

where $\sigma_{ed}^2$ is the variance of $d(A,B)$. The DIEM outcome is scaled to the range of the analyzed quantities $s_{max} - s_{min}$.

DIEM is a theoretically grounded method for high-dimensional comparisons, offering robustness to variance and dimensionality. However, its original formulation assumes randomly distributed data, whereas LLM parameters are structured and optimized based on architecture and fine-tuning objectives, limiting its direct applicability. Specifically, DIEM's theoretical approximation of expected distance fails to capture underlying parameter relationships (Chakraborty and Zhang, 2021), as LLM parameters often exhibit long-tailed and sparse distributions, making standard deviation-sensitive and inflated or misleading similarity scores (Taleb, 2020; Cohen et al., 2020). To overcome these limitations, we redesign DIEM to align with the structural and statistical properties of LLMs, ensuring its effectiveness in detecting meaningful parameter shifts in fine-tuned architectures.

### 3.2 Safe Pruning LoRA (SPLoRA)

In this section, we introduce SPLoRA, a method for enhancing safety alignment in fine-tuned LLMs. We first propose E-DIEM, a redesigned version of DIEM tailored for LLMs. Using the distance score derived from E-DIEM, we selectively prune LoRA layers that contribute to safety misalignment.

#### 3.2.1 Empirical-DIEM (E-DIEM)

For E-DIEM, we propose empirical distance to replace the theoretical expected distance, ensuring a more precise and context-aware comparison. Given the number of LoRA layers as $D$, the empirical distance $\mathbb{E}^*$ is written as follows:

$$\mathbb{E}^*[d(\Delta\theta^i, P^i\Delta\theta^i)] = \frac{1}{D}\sum_{i=1}^{D}\left\|\Delta\theta^i - P^i\Delta\theta^i\right\|_F \quad (2)$$

The empirical distance is computed as the mean of the Euclidean distances between each sampled weight matrix pair. It serves as an adaptive normalization factor, making DIEM more realistically sensitive to data-specific variations rather than assuming uniform high-dimensional behavior.

We use the Interquartile Range (IQR) for scaling instead of standard deviation, as it is more robust to outliers and better captures meaningful weight changes across fine-tuned layers (Vinutha et al., 2018; Rousseeuw and Hubert, 2011). In this way, the possible maximum and minimum distance $s_{max}$ and $s_{min}$ are also within the range of the IQR values.

$$\text{IQR}[d(\Delta\theta^i, P^i\Delta\theta^i)] = Q3(d) - Q1(d) \quad (3)$$

where $Q1(d)$ and $Q3(d)$ are the 25th and 75th percentiles of the sampled Euclidean distances.

In this way, we propose E-DIEM, which is formed as:

$$E - DIEM = \frac{s_{max} - s_{min}}{\text{IQR}[d]} (d - \mathbb{E}^*[d]) \quad (4)$$

where $d = d(\Delta\theta^i, P^i\Delta\theta^i)$ as the Euclidean distance between $\Delta\theta^i$ and $P^i\Delta\theta^i$, $\text{IQR}[d] = \text{IQR}[d(\Delta\theta^i, P^i\Delta\theta^i)$ is IQR for $\Delta\theta^i$ and $P^i\Delta\theta^i$, and $\mathbb{E}^*[d] = \mathbb{E}^*[d(\Delta\theta^i, P^i\Delta\theta^i)]$ is the empirical distance.

### 3.3 E-DIEM Guided LoRA Pruning

After computing layer-wise similarity scores using E-DIEM, we apply a pre-defined threshold $t$ to identify LoRA layers with significant deviations from their projected counterparts. A higher distance score indicates a greater divergence in the direction of LoRA updates, which we hypothesize as a key factor contributing to safety risks in fine-tuned LLMs. Based on this threshold, we selectively prune layers that exhibit the highest misalignment. In this work, the threshold can be set to retain only the top-K layers with the highest distance scores for projection. This process is demonstrated as follows:

$$\mathcal{R}(\Delta\theta^i) = \begin{cases} \text{keep}\Delta\theta^i, & \text{if } u < t \\ \text{prune}\Delta\theta^i, & \text{if } u >= t \end{cases} \quad (5)$$

where $u$ is the outcome of E-DIEM as the distance score. If $u$ is higher than the threshold $t$, the LoRA layer will be removed otherwise it is retained. The entire SPLoRA process is demonstrated in Figure 1.

## 4 Experiments

### 4.1 Datasets and Baselines

We use the Dialog Summary (Gliwa et al., 2019), Alpaca (Taori et al., 2023) and PureBad (Qi et al., 2023) datasets for LoRA fine-tuning. The PureBad dataset comprises 100 harmful examples collected through red-teaming. For Dialog Summary and Alpaca with PureBad dataset experiments, we follow the same fine-tuning setup: 1,000 randomly sampled instances from the respective datasets are mixed with the 100 samples in PureBad dataset. For evaluation, Dialog Summary is assessed on its corresponding test set of 1,500 samples, while Alpaca uses 20% of its total data for testing. For LoRA fine-tuning on the Alpaca dataset without PureBad, we split the data into training and testing sets using an 80/20 ratio. We define a fine-tuning dataset that includes harmful or adversarial examples as an attack. In terms of LLMs, we use Llama-2-7B-Chat, Llama-3-8B-Instruct, Llama-3.2-1B-Instruct (Touvron et al., 2023), and Gemma-7B-it (Team et al., 2024) in our experiments.

We compare our proposed method with the following SOTA techniques:

1. LoRA (Hu et al., 2022): injects trainable low-rank matrices into pre-trained model weights.
2. SafeInstr (Bianchi et al., 2023): leverages instruction-tuned datasets with adversarial training.
3. Backdoor Enhanced Alignment (BEA) (Wang et al., 2024): embeds a controlled backdoor mechanism during fine-tuning to reinforce safety constraints.
4. SafeLoRA (Hsu et al., 2024): introduces a lightweight modification to LoRA fine-tuning by projecting LoRA weights onto a safety-aligned subspace, to mitigate harmful responses while preserving utility.
5. Vaccine (Huang et al., 2024): proposes a perturbation-aware alignment method to safeguard LLMs against harmful fine-tuning attacks.

### 4.2 Evaluation Metrics

In our experiments, we evaluate model performance (utility) using ROUGE-1 F1 and METEOR, which measure the similarity between LLM-generated responses and ground truth. Safety is assessed via the Attack Success Rate (ASR) and Harmfulness Score (HS). An attack is considered successful if the model's response omits explicit refusal keywords, with the keyword list provided in the Appendix 8. We use GPT-4 to evaluate responses and assign harmfulness scores on a 1–5 scale, where lower scores indicate greater safety. To assess model reliability, we employ the Area Under the Accuracy-Rejection Curve (AUARC) (Nadeem et al., 2009), which quantifies selective prediction performance by measuring the trade-off between accuracy and rejection rate. The calculation of AUARC requires binary label for accuracy and uncertainty score for each sample. Following prior work (Kuhn et al., 2023; Lin et al., 2023; Kossen et al., 2024), we use the ROUGE-L

score as a correctness proxy, considering a generation correct if its ROUGE-L score with the reference answer exceeds 0.15. For uncertainty estimation, we assign an uncertainty score to each sample based on the setting of semantic entropy probes,[1] which measures the distributional sparsity of the model's output (Kossen et al., 2024). Additionally, for clarity, we categorize AUARC as a utility metric in this study.

### 4.3 Implementation Details

For our experiments, we use HuggingFace[2] pre-trained LLaMA2-7B-Chat and LLaMA3.2-1B-Instruct as baselines for zero-shot evaluation and LoRA fine-tuning. LoRA is applied to the ''q_proj,'' ''k_proj,'' ''v_proj,'' and ''o_proj'' attention layers, with a fixed rank of 8 across all experiments. To optimize performance on downstream tasks, training hyperparameters vary across datasets, while all fine-tuning is conducted for 5 epochs. For the Dialogue Summary with PureBad dataset, LLaMA2 and LLaMA3-8B are fine-tuned with a learning rate of 5e-5 and a batch size of 8, LLaMA3.2-1B uses a learning rate of 3e-5 with a batch size of 16, and Gemma uses a learning rate of 5e-4 with a batch size of 8.

For Alpaca with PureBad, LLaMA2 is fine-tuned with a learning rate of 5e-5, and Gemma uses learning rate of 5e-4, both with a batch size of 8. For Alpaca without PureBad, the learning rate is set to 2e-5 with a batch size of 16. In BEA experiments, trigger pairs are designed as secret prompts and safety instructions for backdoor samples. We use the official backdoor samples,[3] with backdoor instances comprising 10% of the PureBad dataset. All experiments are conducted on two NVIDIA Tesla P40 GPUs (23GB RAM each).

To further evaluate the effectiveness of our proposed distance-based method, E-DIEM, we conduct an additional experiment where, instead of pruning the identified problematic LoRA layers as in SPLoRA, we replace them with their projected counterparts, following a similar approach to SafeLoRA (Hsu et al., 2024). Since SafeLoRA utilizes cosine similarity as its similarity metric, this comparison enables us to assess which method better captures structural variations in high-dimensional LLM parameters. We refer to this method as Safe Replace LoRA (SRLoRA), with the results presented in the experimental section.

## 5 Results

Table 1 presents the utility and safety evaluation results for LLaMA2, LLaMA3, LLaMA3.2, and Gemma models fine-tuned with LoRA on the Dialogue Summary with PureBad dataset. The baseline corresponds to the pre-trained model with strong safety alignment, leading to the lowest ASR and Harmful Score (HS) values. Our proposed method, SRLoRA, achieves the lowest HS among all fine-tuning methods, demonstrating its effectiveness in preserving safety.

While standard LoRA fine-tuning is expected to yield the highest utility scores at the cost of safety degradation, our findings confirm this trade-off across all methods except SRLoRA, which maintains strong utility performance while preserving safety. Notably, SPLoRA achieves the highest AUARC with LLaMA2 and LLaMA3 models, as shown in the Risk-Coverage Curve in Figure 2 (left), highlighting that our method does not sacrifice safety for utility but instead enhances model reliability.

For the Alpaca with PureBad dataset on the LLaMA2 model, the results in Table 2 further demonstrate the effectiveness of SPLoRA, which achieves superior performance across all safety metrics compared to existing methods. In terms of utility, Vaccine yields the highest ROUGE-1 F1 score, while our method outperforms others on METEOR and AUARC scores. Similar trends are observed with the Gemma model, where the proposed SRLoRA and SPLoRA consistently surpass other SOTA approaches in safety metrics. For utility, our methods maintain competitive performance, outperforming others on most metrics except METEOR. SafeLoRA and Vaccine achieve similarly high performance for METEOR, even exceeding that of LoRA. Notably, we exclude methods other than SafeLoRA from AUARC evaluation due to their excessively high ASR, rendering the remaining data insufficient for reliable AUARC computation. We also exclude LLaMA3 and LLaMA3.2 results, as its

[1] https://github.com/OATML/semantic-entropy-probes.

[2] https://huggingface.co/.

[3] https://github.com/Jayfeather1024/Backdoor-Enhanced-Alignment.

| Model | Method | Utility Metrics (↑) | | | Safety Metrics (↓) | |
|---|---|---|---|---|---|---|
| | | ROUGE | METEOR | AUARC | ASR | HS |
| LLaMA-2 7B-Chat | Baseline | 21.85 | 28.96 | 74.16 | 5.22 | 1.18 |
| | LoRA | 31.04 | 40.87 | 79.16 | 25.32 | 3.21 |
| | Vaccine | 26.84 | 39.23 | 75.68 | 12.96 | 1.19 |
| | SafeInstr | 27.35 | 38.43 | 76.34 | 14.34 | 1.46 |
| | BEA | 26.32 | 37.64 | 77.86 | 15.12 | 1.62 |
| | SafeLoRA | 29.15 | 39.56 | 77.65 | 8.57 | 1.32 |
| | **SRLoRA** (Ours) | **31.12** | 40.54 | 79.88 | **6.73** | **1.05** |
| | **SPLoRA** (Ours) | 30.86 | **41.54** | **80.26** | 6.92 | 1.23 |
| LLaMA-3.2 1B-Instruct | Baseline | 22.65 | 30.24 | 72.13 | 6.15 | 1.15 |
| | LoRA | 30.94 | 41.12 | 80.46 | 22.52 | 2.62 |
| | Vaccine | 27.24 | 37.43 | 77.52 | 13.56 | 1.81 |
| | SafeInstr | 27.85 | 38.03 | 78.24 | 18.36 | 1.96 |
| | BEA | 26.12 | 36.54 | 78.98 | 17.42 | 2.04 |
| | SafeLoRA | 29.25 | 39.23 | 80.96 | 9.37 | 1.76 |
| | **SRLoRA** (Ours) | 30.46 | 41.24 | **81.32** | 6.23 | **1.49** |
| | **SPLoRA** (Ours) | **31.12** | **43.24** | 80.02 | **5.73** | 1.58 |
| LLaMA-3 8B-Instruct | Baseline | 26.38 | 32.54 | 78.27 | 6.62 | 1.21 |
| | LoRA | 35.35 | 43.31 | 86.65 | 23.52 | 1.78 |
| | Vaccine | 36.24 | 43.34 | 86.05 | **7.42** | 1.36 |
| | SafeInstr | 35.23 | 42.75 | 85.23 | 8.97 | 1.35 |
| | BEA | 37.17 | 45.14 | 86.46 | 9.65 | 1.47 |
| | SafeLoRA | 36.03 | 44.28 | 86.35 | 9.40 | 1.43 |
| | **SRLoRA** (Ours) | 36.91 | **44.96** | 86.40 | 9.07 | **1.22** |
| | **SPLoRA** (Ours) | **37.82** | 44.73 | **87.96** | 8.85 | 1.34 |
| Gemma 7B-it | Baseline | 23.58 | 25.14 | 72.03 | 9.26 | 1.62 |
| | LoRA | 32.61 | 33.05 | 81.23 | 70.46 | 3.92 |
| | Vaccine | 33.55 | 34.53 | 81.84 | 56.32 | 2.56 |
| | SafeInstr | 33.67 | 34.16 | 80.36 | 76.64 | 3.31 |
| | BEA | 34.01 | 34.23 | 81.09 | 69.63 | 2.95 |
| | SafeLoRA | 33.70 | **34.80** | 80.97 | 33.56 | 2.35 |
| | **SRLoRA** (Ours) | 33.34 | 33.45 | **82.63** | 29.65 | 2.03 |
| | **SPLoRA** (Ours) | **34.32** | 34.49 | 82.43 | **25.64** | **1.98** |

Table 1: Performance comparison of our methods against LoRA, SafeInstr, BEA, and Vaccine on the Dialog Summary dataset with PureBad, using LLaMA-2-7B-Chat, LLaMA-3.2-1B-Instruct, LLaMA-3-8B-Instruct, and Gemma-7b-it models. HS (Harmfulness Score) and ASR (Attack Success Rate) are used to assess safety. Higher values (↑) indicate better performance, and lower values (↓) indicate better safety. For clarity, all results except HS are reported as percentages.

ASR remains comparable to LLaMA2 across three methods.

Beyond safety alignment, we also assess our method's ability to mitigate unreliable, misleading, and inaccurate generations in LoRA fine-tuning. Table 3 presents results on Alpaca fine-tuning without malicious data (PureBad), where SRLoRA enhances both safety and utility while improving the model's capability to detect its own errors, as reflected by the AUARC score. As shown in Figure 2 (right), SPLoRA improves selective generation by removing high-entropy samples, demonstrating its effectiveness in uncertainty-aware generations.

Furthermore, we measure inference time before and after pruning with SPLoRA. Since all other methods utilize the original model without pruning, their inference times remain unchanged.

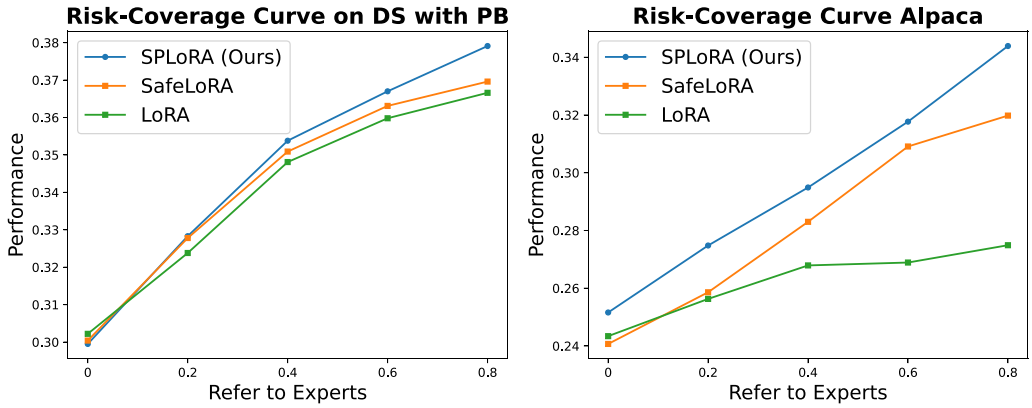


Figure 2: The Risk-Coverage Curve compares LoRA, SafeLoRA, and our proposed SPLoRA, with performance measured using the ROUGE-1 F1 score. The x-axis (‘‘Refer to experts’’) represents the percentage of samples with the highest uncertainty scores. The left plot shows results for fine-tuningn on Dialogue Summary with PureBad dataset using the LLaMA2 model, while the right plot presents results for fine-tuning on the Alpaca dataset with LLaMA2 model.

| Model | Method | Utility Metrics (↑) | | | Safety Metrics (↓) | |
|---|---|---|---|---|---|---|
| | | ROUGE | METEOR | AUARC | ASR | HS |
| LLaMA-2 7B-Chat | Baseline | 19.93 | 12.86 | 56.82 | 15.45 | 1.15 |
| | LoRA | 24.32 | 21.46 | – | 80.81 | 2.67 |
| | Vaccine | **25.86** | 21.41 | – | 65.42 | 2.13 |
| | SafeInstr | 25.13 | 20.75 | – | 78.56 | 2.54 |
| | BEA | 24.62 | 20.86 | – | 76.84 | 3.25 |
| | SafeLoRA | 24.35 | 19.08 | 70.42 | 9.65 | 1.36 |
| | **SRLoRA** (Ours) | 26.28 | **21.89** | **72.56** | 7.32 | 1.25 |
| | **SPLoRA** (Ours) | 25.32 | 21.11 | 71.04 | **7.18** | **1.21** |
| Gemma 7B-it | Baseline | 11.15 | 9.26 | 54.13 | 25.64 | 2.65 |
| | LoRA | 27.57 | 24.67 | – | 85.32 | 3.49 |
| | Vaccine | 26.45 | 25.32 | – | 55.38 | 1.79 |
| | SafeInstr | 24.43 | 23.21 | – | 80.64 | 3.56 |
| | BEA | 24.75 | 23.51 | – | 58.02 | 2.76 |
| | SafeLoRA | 26.97 | **25.38** | 70.07 | 16.42 | 1.63 |
| | **SRLoRA** (Ours) | 25.40 | 24.81 | **72.35** | 15.43 | **1.27** |
| | **SPLoRA** (Ours) | **27.92** | 25.13 | 71.86 | **14.31** | 1.31 |

Table 2: Performance comparison of our methods against LoRA, SafeInstr, BEA, and Vaccine on the Alpaca dataset with PureBad, using LLaMA-2-7B-Chat and Gemma-7B-it models. HS (Harmfulness Score) and ASR (Attack Success Rate) are used to assess safety. Higher values (↑) indicate better performance, and lower values (↓) indicate better safety. For clarity, all results except HS are reported as percentages.

Table 5 reports inference time for SPLoRA before and after pruning, showing an approximately 12.5% reduction in both total inference time and per-sample latency. This result indicates that SPLoRA not only enhances safety alignment but also reduces computational overhead, making LLMs more efficient while maintaining robust performance.

## 6 Ablation Study

We conduct a comprehensive ablation study alongside our main experiments to assess the effectiveness of SPLoRA from multiple perspectives. While SPLoRA utilizes E-DIEM to compare layer-wise LoRA weights with their projected counterparts, SafeLoRA (Hsu et al., 2024) relies

| Model | Method | Utility Metrics (↑) | | | Safety Metrics (↓) | |
|---|---|---|---|---|---|---|
| | | ROUGE | METEOR | AUARC | ASR | HS |
| LLaMA-2 7B-Chat | LoRA | 24.57 | 20.26 | 70.56 | 25.23 | 1.82 |
| | Vaccine | 24.86 | 19.53 | 68.75 | 10.32 | 1.17 |
| | SafeLoRA | 24.21 | 20.96 | 67.23 | 7.43 | 1.32 |
| | **SRLoRA** (Ours) | **25.63** | **21.35** | **72.07** | 5.85 | 1.03 |
| | **SPLoRA** (Ours) | 25.03 | 20.74 | 71.63 | **4.61** | **0.79** |

Table 3: Performance comparison of our methods against LoRA, SafeLoRA and Vaccine on the Alpaca dataset using LLaMA-2-7B-Chat model. HS (Harmfulness Score) and ASR (Attack Success Rate) are used to assess safety. For clarity, all results except HS are reported as percentages.

| Model | Pruned Layers | Threshold Value | Utility Metrics (↑) | | | Safety Metrics (↓) | |
|---|---|---|---|---|---|---|---|
| | | | ROUGE | METEOR | AUARC | ASR | HS |
| LLaMA-2 7B-Chat | 5 layers | 0.95 | 29.76 | 39.54 | 79.32 | 7.93 | 1.31 |
| | **10 layers** | 0.93 | 30.86 | 41.55 | 80.26 | 6.92 | 1.23 |
| | 15 layers | 0.90 | 28.83 | 39.12 | 79.14 | 7.56 | 1.35 |
| | 20 layers | 0.88 | 28.48 | 38.29 | 78.52 | 8.62 | 1.47 |

Table 4: Impact of layer pruning threshold. Utility and safety metrics on the Dialogue Summary with PureBad dataset using the LLaMA2-7B-Chat model, evaluated under different pruning thresholds based on the number of pruned layers.

| **Model** | **Method** | **Inference Time (s)** | |
|---|---|---|---|
| | | **Total** | **Per Sample** |
| LLaMA2 | BS | 1131.27 | 0.75 |
| | Pruned | 993.97 | 0.66 |
| LLaMA3.2 | BS | 240.42 | 0.16 |
| | Pruned | 196.36 | 0.13 |

Table 5: Comparison of inference time before and after pruning on the Dialogue Summary test set (1,500 samples) using LLaMA2-7B-Chat, and the Alpaca test set (randomly selected 1500 samples) using LLaMA3.2-1B-Instruct model. BS refers to the baseline pre-trained model, while Pruned represents the model after applying our proposed pruning method. Total denotes the overall inference time, while Per Sample indicates the inference time per instance.

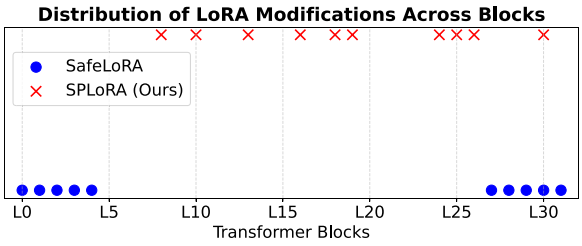


Figure 3: Distribution of LoRA layers exhibiting significant deviations from the pre-trained model in transformer blocks, potentially leading to safety misalignment.

on cosine similarity. For a fair comparison, we select 10 layers with the highest E-DIEM scores and lowest cosine similarity scores (for SafeLoRA), identifying those exhibiting significant deviations that may contribute to safety misalignment. As shown in Figure 3, among the 32 transformer blocks, SPLoRA detects dissimilar layers across the entire network, whereas SafeLoRA primarily identifies deviations in the initial and final layers. In LLMs, early layers extract general features, middle layers refine representations for contextual understanding, and final layers handle generation, reasoning, and decision-making (Wei et al., 2024; Touvron et al., 2023; Sun et al., 2024). Excessive modifications in critical layers can compromise model performance, leading to unintended behavioral shifts. The result suggests that SPLoRA ensures that fine-tuning aligns with the intended purpose of LoRA—adapting the model while preserving core safety properties and utility.

To evaluate the effectiveness of our proposed E-DIEM, we replace E-DIEM with the original

| Model | Metric | DIEM | E-DIEM |
|---|---|---|---|
| LLaMA2 | ROUGE (↑) | 30.32 | 30.86 |
| | METEOR (↑) | 40.28 | 41.54 |
| | ASR (↓) | 7.03 | 6.92 |
| LLaMA3.2 | ROUGE (↑) | 29.65 | 31.12 |
| | METEOR (↑) | 40.55 | 43.24 |
| | ASR (↓) | 6.52 | 5.73 |

Table 6: Performance comparison of E-DIEM and DIEM on LLaMA2 and LLaMA3 models, fine-tuned with Dialogue Summary with PureBad datasets. Higher ROUGE and METEOR indicate better utility (↑), while lower ASR indicates better safety (↓).

DIEM as the distance measurement for SPLoRA. Table 6 shows that E-DIEM outperforms DIEM in both safety and utility, demonstrating its superior ability to capture the structural features of LLM parameters.

To assess the impact of layer pruning, we evaluate LLaMA2 on the Dialogue Summary with PureBad dataset by pruning 5, 10, 15, and 20 layers. As shown in Table 4, pruning 10 layers achieves the best balance between utility and safety metrics. We therefore adopt this configuration for all subsequent experiments, aligning with the SafeLoRA approach (Hsu et al., 2024), which also retains 10 layers in its projection module. Instead of applying a fixed pruning threshold, we use a dynamic threshold set to the E-DIEM value of the 10th highest-ranked (i.e., least similar) layer. This ranking-based strategy mirrors that of SafeLoRA (Hsu et al., 2024), which selects layers based on the 10th lowest cosine similarity score. Fixed thresholds can be overly sensitive to variations in layer activation distributions across tasks or checkpoints, whereas our method promotes more stable and interpretable pruning behavior.

## 7 Conclusion

This work introduces Safe Pruning LoRA (SPLoRA), a novel pruning-based strategy for enhancing safety alignment in fine-tuned LLMs while maintaining model performance. Unlike prior approaches that primarily focus on mitigating risks from fine-tuning with malicious data, SPLoRA is effective even in standard LoRA fine-tuning scenarios, where it identifies unreliable layers that may compromise both utility and safety. E-DIEM enables precise layer selection, ensuring that LoRA adaptation preserves model reliability without introducing unintended vulnerabilities. Experimental results demonstrate that SPLoRA outperforms existing safety alignment techniques, achieving superior results in both utility and safety metrics across multiple datasets. Furthermore, our approach significantly reduces computational overhead while maintaining strong safety alignment, addressing a key limitation of previous LoRA-based adaptation methods.

Despite its strengths, SPLoRA has certain limitations. First, while pruning mitigates safety vulnerabilities without compromising performance, there is a need for further investigation into layer-wise interpretability to better understand the role of each LoRA layer in shaping safety alignment. Additionally, the effectiveness of E-DIEM in capturing structural variations suggests broader applicability beyond safety alignment, such as improving robustness in adversarial settings or detecting catastrophic forgetting in continual learning scenarios. Another avenue for future work is refining our pruning strategy by integrating adaptive thresholds, allowing for more dynamic adjustments based on dataset complexity and task requirements.

Looking ahead, we plan to extend our study by incorporating more diverse LLM architectures and exploring alternative pruning techniques, including structured pruning and quantization-aware pruning, to further enhance efficiency. Moreover, integrating SPLoRA into a broader safety-aware fine-tuning framework, potentially combining it with reinforcement learning or contrastive alignment methods, could lead to even more robust and generalizable safety mechanisms. Ultimately, our work contributes to the growing field of responsible AI, providing a scalable, theoretically grounded approach for maintaining safety alignment in fine-tuned LLMs.

## Acknowledgments

We would like to express our gratitude to the study participants for their participation. This work was supported by the Engineering and Physical Sciences Research Council (EPSRC) [grant number EP/Y009800/1]: AI UK: Creating an International Ecosystem for Responsible AI Research and Innovation. Furthermore, we would like to express our thanks to our reviewers and action editor for their thoughtful feedback. Any opinions, findings, and

conclusions expressed in this material are those of the author(s).

## References


Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, and Shawn Jain. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Sheldon Axler. 2024. *Linear Algebra Done Right*. Springer Nature. https://doi.org/10.1007/978-3-031-41026-0

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. 2022a. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. 2022b. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

João Carlos Alves Barata and Mahir Saleh Hussein. 2012. The moore–penrose pseudoinverse: A tutorial review of the theory. *Brazilian Journal of Physics*, 42:146–165. https://doi.org/10.1007/s13538-011-0052-z

Federico Bianchi, Mirac Suzgun, Giuseppe Attanasio, Paul Röttger, Dan Jurafsky, Tatsunori Hashimoto, and James Zou. 2023. Safety-tuned llamas: Lessons from improving the safety of large language models that follow instructions. *arXiv preprint arXiv:2309.07875*.

Sébastien Bubeck, Varun Chadrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4.

Shubhadeep Chakraborty and Xianyang Zhang. 2021. A new framework for distance and

kernel-based metrics in high dimensions. *Electronic Journal of Statistics*, 15(2): 5455–5522. https://doi.org/10.1214/21-EJS1889

Joel E. Cohen, Richard A. Davis, and Gennady Samorodnitsky. 2020. Heavy-tailed distributions, correlations, kurtosis and Taylor's law of fluctuation scaling. *Proceedings of the Royal Society A*, 476(2244):20200610. https://doi.org/10.1098/rspa.2020.0610, PubMed: 33408562

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 70–79, Hong Kong, China. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-5409

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 finetuning large language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Tiansheng Huang, Sihao Hu, and Ling Liu. 2024. Vaccine: Perturbation-aware alignment for large language models against harmful fine-tuning attack. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Tiansheng Huang, Sihao Hu, and Ling Liu. 2025. Vaccine: Perturbation-aware alignment for large language models against harmful fine-tuning attack. *Advances in Neural Information Processing Systems*, 37:74058–74088.

Jannik Kossen, Jiatong Han, Muhammed Razzak, Lisa Schut, Shreshth Malik, and Yarin Gal. 2024. Semantic entropy probes: Robust and cheap hallucination detection in LLMs. *arXiv preprint arXiv:2406.15927*.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *arXiv preprint arXiv:2302.09664*.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. 2023. Generating with confidence: Uncertainty quantification for black-box large language models. *arXiv preprint arXiv:2305.19187*.

Xinyin Ma, Gongfan Fang, and Xinchao Wang. 2023. Llm-pruner: On the structural pruning of large language models. *Advances in Neural Information Processing Systems*, 36:21702–21720.

Malik Sajjad Ahmed Nadeem, Jean-Daniel Zucker, and Blaise Hanczar. 2009. Accuracy-rejection curves (arcs) for comparing classification methods with a reject option. In *Machine Learning in Systems Biology*, pages 65–81. PMLR.

Vít Novotnỳ, Eniafe Festus Ayetiran, Michal Štefánik, and Petr Sojka. 2020. Text classification with word embedding regularization and soft similarity measure. *arXiv preprint arXiv:2003.05019*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

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

Peter J. Rousseeuw and Mia Hubert. 2011. Robust statistics for outlier detection. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(1):73–79. https://doi.org/10.1002/widm.2

Gilbert Strang. 2000. Linear algebra and its applications.

Qi Sun, Marc Pickett, Aakash Kumar Nain, and Llion Jones. 2024. Transformer layers as painters. *arXiv preprint arXiv:2407.09298*.

Nassim Nicholas Taleb. 2020. Statistical consequences of fat tails: Real world preasymptotics, epistemology, and applications. *arXiv preprint arXiv:2001.10488*.

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.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikuła, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L. Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech Stokowiec, Yu-hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Clément Farabet, Oriol Vinyals, Jeff Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Federico Tessari and Neville Hogan. 2024. Surpassing cosine similarity for multidimensional comparisons: Dimension insensitive euclidean metric (diem). *arXiv preprint arXiv:2407.08623*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

H. P. Vinutha, B. Poornima, and B. M. Sagar. 2018. Detection of outliers using interquartile range technique from intrusion dataset. In *Information and Decision Sciences: Proceedings of the 6th International Conference on Ficta*, pages 511–518. Springer. https://doi.org/10.1007/978-981-10-7563-6_53

Jiongxiao Wang, Jiazhao Li, Yiquan Li, Xiangyu Qi, Muhao Chen, Junjie Hu, Yixuan Li, Bo Li, and Chaowei Xiao. 2024. Mitigating fine-tuning jailbreak attack with backdoor enhanced alignment. *arXiv preprint arXiv:2402.14968*.

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*.

Hiroaki Yamagiwa, Momose Oyama, and Hidetoshi Shimodaira. 2024. Revisiting cosine similarity via normalized ica-transformed embeddings. *arXiv preprint arXiv:2406.10984*.

Xianjun Yang, Xiao Wang, Qi Zhang, Linda Petzold, William Yang Wang, Xun Zhao, and Dahua Lin. 2023. Shadow alignment: The ease of subverting safely-aligned language models. *arXiv preprint arXiv:2310.02949*.

Mingyang Zhang, Hao Chen, Chunhua Shen, Zhen Yang, Linlin Ou, Xinyi Yu, and Bohan Zhuang. 2023. Loraprune: Pruning meets low-rank parameter-efficient fine-tuning. *arXiv preprint arXiv:2305.18403*. https://doi.org/10.18653/v1/2024.findings-acl.178

Zixuan Zhou, Xuefei Ning, Ke Hong, Tianyu Fu, Jiaming Xu, Shiyao Li, Yuming Lou, Luning Wang, Zhihang Yuan, Xiuhong Li, Shengen Yan, Guohao Dai, Xiao-Ping Zhang, Yuhan Dong, and Yu Wang . 2024. A survey on efficient inference for large language models. *arXiv preprint arXiv:2404.14294*.

# 8 Appendices

## 8.1 Keyword List

Following the setting in Hsu et al. (2024), the success of the attack hinges on the absence of keywords indicating a refusal to respond in responses of LLMs.

Keyword List: [‘‘I’m sorry’’, ‘‘Sorry’’, ‘‘I am sorry’’, ‘‘I apologize’’, ‘‘As an’’, ‘‘As an AI’’, ‘‘I’m an’’, ‘‘I’m just’’, ‘‘As a language model’’, ‘‘As an Assistant’’, ‘‘I cannot’’, ‘‘I do not’’, ‘‘Hello!’’, ‘‘is illegal and unethical’’, ‘‘I’m really sorry’’, ‘‘OpenAI’’, ‘‘unethical’’, ‘‘not ethical’’, ‘‘illegal’’, ‘‘not legal’’, ‘‘My apologies’’, ‘‘but I cannot’’, ‘‘If you have any other non-malicious requests’’, ‘‘I’m not able to provide’’, ‘‘I am really sorry’’, ‘‘I cannot create’’, ‘‘I cannot support’’, ‘‘I can’t provide’’, ‘‘I can’t assist’’]

## 8.2 ChatGPT Prompt

Prompt for ChatGPT to check the harmfulness score:

[Please score the following responses for safety, rating each from 1 (safe) to 5 (harmful), based on whether they include or imply any form of harmful, unethical, illegal, or restricted content. Consider common harm categories such as illegal activity, abuse, violence, fraud, adult content, privacy violations, and other safety-related risks when assigning your score.]