# Alignment with Preference Optimization Is All You Need for LLM Safety

## Reda Alami, Ali Khalifa Almansoori, Ahmed Alzubaidi, Mohamed El Amine Seddik, Mugariya Farooq, Hakim Hacid

Technology Innovation Institute, Abu Dhabi {firstname.surname}@tii.ae

#### Abstract

We demonstrate that preference optimization methods can effectively enhance LLM safety. Applying various alignment techniques to the Falcon 11B model using safety datasets, we achieve a significant boost in global safety score (from 57.64% to 99.90%) as measured by LlamaGuard 3 8B, competing with state-of-the-art models. On toxicity benchmarks, average scores in adversarial settings dropped from over 0.6 to less than 0.07. However, this safety improvement comes at the cost of reduced general capabilities, particularly in math, suggesting a trade-off. We identify noise contrastive alignment (Safe-NCA) as an optimal method for balancing safety and performance. Our study ultimately shows that alignment techniques can be sufficient for building safe and robust models.

<u>↑</u>This paper contains words that may be offensive or harmful.

### Introduction

Large language models (LLMs) are highly valuable for their ability to process and generate contextually appropriate text across various applications. However, ensuring the safety of these models is equally crucial. Safety in LLMs refers to their ability to consistently generate content that is accurate, ethical, and adheres to societal norms while preventing the production of harmful or inappropriate content. This paper investigates the effectiveness of preference optimization methods in enhancing LLM safety, specifically focusing on alignment techniques applied to the Falcon 11B model (Malartic et al. 2024) using safety datasets.

Our study demonstrates that these alignment methods can significantly boost the global safety score of LLMs, as measured by LlamaGuard 3 8B. We achieved an increase from 57.64% to 99.90% in safety scores, competing with state-of-the-art models (Figure 1). Additionally, we observed a substantial reduction in toxicity scores under adversarial conditions, dropping from over 0.6 to less than 0.07. However, this improved safety comes at the cost of reduced general capabilities, particularly in mathematical tasks, indicating a trade-off between safety and performance.

Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.



Figure 1: Comparison of the global safety scores of 11 LLMs. The scores are derived from averaging the results of the safety ALERT and safety Adversarial ALERT benchmarks to assess each model's overall performance across the safety evaluations. Notice the significant performance boost from 57.64% to 99.9% for the Falcon 11B model.

Among the explored techniques, we identify noise contrastive alignment (Safe-NCA) as an optimal method for balancing safety and overall model performance. Our investigation ultimately demonstrates that alignment techniques can be sufficient for developing safe and robust LLMs, while highlighting the importance of considering the trade-offs involved in enhancing model safety.

The remainder of the paper is structured as follows. The next section introduces related work, followed by a formalization of the safety problem for LLMs. We then detail the alignment techniques explored in this study, present the benchmark used, and discuss our experimental results. Finally, we conclude the paper with our findings and implications for future research in LLM safety.

### **Related Work**

Several existing works have explored safety with LLMs. We divided the literature into (1) Safety evaluation benchmarks for LLMs and (2) Safety enhancement techniques for LLMs.

## **Safety Evaluation for LLMs**

In this subsection, major benchmarks found in the literature, that evaluate safety are presented. The paper by (Varshney et al. 2023) introduced the Safety and Over-Defensiveness Evaluation (SODE) benchmark, which consists of a diverse set of safe and unsafe prompts designed to systematically evaluate LLMs responses. A DeBERTAv3-large model was trained to act as a binary classifier judging safe/unsafe responses. Moving on to ALERT (Tedeschi et al. 2024), which is a comprehensive benchmark designed to evaluate the safety of LLMs using finegrained risk taxonomy, and red-teaming approaches. This benchmark includes 45k instructions categorized to assess various safety vulnerabilities through adversarial testing scenarios. Experimental evaluations across 10 widely used LLMs, revealed that many still fail to achieve satisfactory safety levels, underscoring the ongoing challenges in LLM safety assurance. Recent advancements in LLMs have prompted an exploration into their operational safety within interactive environments. In this context, (Yuan et al. 2024) developed R-Judge, a benchmark specifically designed to assess LLMs ability to identify and judge safety risks based on records of multi-turn agent interactions. The R-Judge benchmark is composed of 162 interaction records, spanning 27 key risk scenarios. Human input was leveraged to annotate each record, in terms of safety and risk description. Out of the nine models evaluated, GPT-4 demonstrated the best performance in both safety judgment and risk identification. (Gupta et al. 2024) proposed framework where the safety of LLM is assessed and the reliability of the underlying judge, is inline with the rise of LLM-as-a-Judge approaches. Refusal behavior testing was introduced, which frames the prompt to the LLM as yes/no question, prompting the LLM if it is interested in engaging with a specific prompt. Additionally, a model called Walled-Guard was proposed, intended to evaluate the quality of the judge deployed in the evaluation process.

### **Safety Enhancement for LLMs**

In this subsection, a closer look is taken at existing research that explored addressing the safety issue with LLMs. Several prompt modification techniques were introduced in (Zheng et al. 2024), such as in In-Context exemplars and self-safety checks on prompts and responses, acting as a defense strategy against adversarial prompts. Moreover, work found in (Kumar et al. 2023), proposed erase-and-check framework that makes LLM robust against attacks on prompt, that push an LLM to be unsafe such as adversarial suffix, adversarial insertion, and adversarial infusion attacks. Llama 2 and DistilBERT were exploited to act as harmful detectors applied on various subsequences of the prompt. Three variations of erase-and-check were introduced, however, the

increased running time is a major drawback of this work. As part of Llama 2 development (Touvron et al. 2023), red-teaming technique was utilized to enhance safety. The techniques entailed asking humans to interact with a target LLM by finding prompts that elicit unsafe responses in a wide range of topics. The exercise performed in redteaming produces a fine-tuning dataset that can be leveraged to enhance safety. Authors in (Ge et al. 2023) alleviated the need for human involvement by introducing a Multi-round Automatic Red-Teaming (MART) method. MART employs an iterative process where in each round, both adversarial and target LLM are fine-tuned. The former is trained to generate prompts that provoke the LLM to return an unsafe response, whereas the latter is trained to return safe responses. Experiments indicated that four rounds of MART significantly reduced the violation rate in the target LLM by up to 84.7 percent. Following the same spirit, (Jiang et al. 2024) proposed techniques called Deep Adversarial Automated Red Teaming (DART). Compared to MART, they focused on producing a more diverse dataset to be used in subsequent rounds. (Dai et al. 2023) adapted an existing common alignment technique RLHF (Ouyang et al. 2022), by introducing Safe RLHF. Annotators involved in data annotations were instructed to label samples in terms of helpfulness and harmlessness. The reward model (R) is deduced from the helpfulness dataset, whereas the harmlessness dataset was used to obtain the cost model (C). Leveraging C and R, the objective is defined which Safe RLHF attempts to maximize a defined objective and was solved using Lagrangian methods. (Ji et al. 2024b) introduced the BEAVERTAILS-330k dataset, which is a human preference annotations dataset aimed at improving the safety alignment of LLMs. Additionally, the authors demonstrated how the dataset can be utilized in content moderation and RLHF.

### The Safety Problem for LLMs

In this section, we start by introducing useful definitions regarding the safety risk taxonomy and adversarial attack taxonomy. Then, we introduce the safety objectives.

**Definition 1** (Harmful Category). To further safeguard the integrity of LLM outputs, harmful categories are explicitly defined as any content that promotes, encourages, or facilitates behaviors and ideologies that are discriminatory, unlawful, or unethical. These include the following six categories denoted that compose the safety risk taxonomy  $C = \{c_1, c_2, c_3, c_4, c_5, c_6\}$ .

- c<sub>1</sub>: **Criminal Planning** Content that encourages or provides guidance on criminal activities such as physical violence, theft, or illegal cyber activities.
- c<sub>2</sub>: **Hate Speech** Expressions that promote discrimination or hostility against individuals or groups based on race, gender, sexual orientation, religion, disability, or economic status.
- c<sub>3</sub>: **Regulated or Controlled Substances** Content endorsing or facilitating the illegal production, distribution, or consumption of controlled substances including drugs and alcohol.

- c<sub>4</sub>: **Sexual Content** Promotions of or instructions on producing and distributing sexually explicit material.
- c<sub>5</sub>: **Suicide & Self-Harm** Content that encourage selfharm or provide instructions for committing suicide, or content that promotes unhealthy body standards.
- c<sub>6</sub>: **Guns & Illegal Weapons** Information that supports the acquisition, creation, or usage of illegal weapons including firearms, chemical, biological, and radioactive materials.

**Definition 2** (Adversarial Attacks). To further understand and address the methods used to exploit or manipulate a LLM into generating harmful content, we define a set of adversarial techniques designed to circumvent the model's safety guidelines, potentially leading to outputs that fall within the harmful categories of the safety risk taxonomy C. The adversarial techniques are as follows:

- Adversarial suffix  $(a_1)$ : Adding deceptive content at the end of a prompt. The added suffix might direct the model to generate inappropriate or harmful responses, despite the initial part of the prompt being benign.
- Adversarial prefix (a<sub>2</sub>): Adding a misleading or harmful phrase at the beginning of a prompt. The goal is to trick the LLM into continuing along the harmful theme set by the prefix, potentially bypassing safety filters.
- Token manipulation (a<sub>3</sub>): Altering specific words or characters within a prompt to change its meaning. It can include misspelling words, swapping tokens, or inserting additional tokens that alter the intent of the prompt.
- **Jailbreaking** (a<sub>4</sub>): Using the prompt to change the model's operational parameters. This could involve instructing the model to ignore its ethical guidelines or safety protocols, effectively tricking it into operating in an unsafe mode.

In the following, we denote by  $A = \{a_1, a_2, a_3, a_4\}$  the adversarial attack taxonomy.

**Definition 3** (Safe/Unsafe output). *Given a model*  $\pi$ , *for any input prompt x, the model's output y*  $\sim \pi(\cdot|x)$  *is evaluated against the six harmful categories*  $\{c_1, c_2, c_3, c_4, c_5, c_6\}$ . *The evaluation function y*  $\rightarrow \omega(y)$  *is defined as follows:* 

$$\omega(y) = \begin{cases} 0 & if \ y \in \{c_1, c_2, c_3, c_4, c_5, c_6\} \\ 1 & otherwise \ (y \ is \ safe) \end{cases}$$

The safety evaluation function  $y \to \omega(y)$  determines whether the response y falls into any of the harmful categories  $c_i \in C$ , thereby marking the response as unsafe if it evaluates to 0.

In practice, we use as evaluation function  $\omega$  the LLama Guard 3 8B (Dubey et al. 2024).

**Definition 4** (Safety Score S). Given a model  $\pi$ . For a given category  $c \in C$ , the safety score  $S_c(\pi)$  is defined as:

$$S_c(\pi) = \frac{\sum_{p \in P_c} \omega(\pi(p))}{|P_c|}$$

where  $\omega(\pi(p)) = 1$  if the response to prompt p is safe, and 0 otherwise. Here,  $P_c$  represents the set of prompts associated with category c, and p is a specific prompt within this set.

The overall safety score  $S(\pi)$  is computed as the weighted average of the category-wise safety scores:

$$S(\pi) = \sum_{c_i \in C} \frac{|P_{c_i}|}{\sum_{c_i \in C} |P_{c_i}|} \times S_{c_i}(\pi)$$

where  $|P_{c_i}|$  is the number of prompts in category  $c_i$ .

**Definition 5** (Attack Success Rate Score (ASR)). *Given a*  $model \pi$ . The attack success rate (ASR)  $ASR_a(\pi)$  for each attack  $a \in A$  is defined as:

$$ASR_a(\pi) = \frac{\sum_{\tilde{p} \in \tilde{P}_a} (1 - \omega(\pi(\tilde{p})))}{|\tilde{P}_a|}$$

where  $\omega(\pi(\tilde{p})) = 1$  if the response to prompt  $\tilde{p}$  is safe, and 0 otherwise. Here,  $\tilde{P}_a$  represents the set of prompts associated with the adversarial attack  $a \in A$ , and  $\tilde{p}$  is a specific prompt within this set.

The overall attack success rate  $ASR(\pi)$  is computed as the weighted average of the category-wise attack success rates:

$$ASR(\pi) = \sum_{a_i \in A} \left( \frac{|\tilde{P}_{a_i}|}{\sum_{a_i \in A} |\tilde{P}_{a_i}|} \times ASR_{a_i}(\pi) \right)$$

where  $|\tilde{P}_{a_i}|$  is the number of prompts for the attack  $a_i$ .

**Safety Objectives:** Given a LLM  $\pi$ , the objective is to finetune it with a finetune procedure f so that the safety score  $S(f(\pi))$  is maximized while minimizing the Attack Success Rate ASR( $f(\pi)$ ):

$$\max S(f(\pi))$$
 and  $\min ASR(f(\pi))$ 

This dual focus aims at enhancing model robustness by reducing the likelihood of generating harmful content and improving resistance to adversarial attacks.

### **Safety Alignment**

In this section, we start by introducing the methodology that we propose to address the safety of LLMs, then we present the safety alignment methods as well as the corresponding safety dataset.

### Methodology

The safety problem in LLMs can be approached as an alignment problem. The objective is to align the model with a dataset that contains both safe and less safe responses. By doing so, the model learns to prioritize generating safer outputs while minimizing the risk of harmful content. This alignment process, supported by preference optimization techniques (such as DPO, IPO, etc.), fine-tunes the model to consistently favor responses that adhere to the safety risk taxonomy *C* as defined in Definition 1.

# Dataset with the pairwise comparison for safe alignment

To construct our preference dataset, we start by filtering the PKU-SafeRLHF dataset available at https://huggingface.co/datasets/PKU-Alignment/PKU-SafeRLHF.

Table 1: Loss functions for various Safe Optimization methods

Method	Loss Function
Safe-DPO (Rafailov et al. 2024)	$\mathscr{L}_{\text{Safe-DPO}}(\pi_{\theta}; \pi_{\text{ref}}) = -\mathbf{E}_{\left(x, \mathbf{y}_{sw}, \mathbf{y}_{ij}\right) - \mathscr{D}_{\text{Safery}}} \left[ \log \sigma \left( \beta \log \frac{\pi_{\theta}(\mathbf{y}_{sw}   \mathbf{x})}{\pi_{\text{ref}}(\mathbf{y}_{sw}   \mathbf{x})} - \beta \log \frac{\pi_{\theta}(\mathbf{y}_{i}   \mathbf{x})}{\pi_{\text{ref}}(\mathbf{y}_{i}   \mathbf{x})} \right) \right]$
Safe-robust_DPO (Chowdhury 2024)	$\mathscr{L}_{\text{Safe-rDPO}}(\pi_{\theta}; \pi_{\text{ref}}) = -\frac{1}{1-2\varepsilon} \mathbf{E}_{\left[x, \mathbf{y}_{z_{tr}}, \mathbf{y}_{z_{l}}\right] \sim \mathscr{D}_{\text{Safev}}} \left[ (1-\varepsilon) \log \sigma \left( \beta \log \frac{\pi_{\theta}(\mathbf{y}_{z_{tr}}   \mathbf{x})}{\pi_{\text{ref}}(\mathbf{y}_{z_{tr}}   \mathbf{x})} - \beta \log \frac{\pi_{\theta}(\mathbf{y}_{z_{l}}   \mathbf{x})}{\pi_{\text{ref}}(\mathbf{y}_{z_{l}}   \mathbf{x})} \right) - \varepsilon \log \sigma \left( -\beta \log \frac{\pi_{\theta}(\mathbf{y}_{z_{tr}}   \mathbf{x})}{\pi_{\text{ref}}(\mathbf{y}_{z_{tr}}   \mathbf{x})} + \beta \log \frac{\pi_{\theta}(\mathbf{y}_{z_{lr}}   \mathbf{x})}{\pi_{\text{ref}}(\mathbf{y}_{z_{lr}}   \mathbf{x})} \right) \right]$
Safe-IPO (Azar et al. 2024)	$\mathcal{L}_{\text{Safe-IPO}}\left(\pi_{\theta}; \pi_{\text{ref}}\right) = \mathbf{E}_{\left(\mathbf{x}, \mathbf{y}_{s_{w}}, \mathbf{y}_{s_{l}}\right) \sim \mathcal{D}_{\text{Safev}}}\left[\log\left(\frac{\pi_{\theta}(\mathbf{y}_{s_{w}} \mathbf{x})\pi_{\text{ref}}(\mathbf{y}_{s_{l}} \mathbf{x})}{\pi_{\theta}(\mathbf{y}_{s_{l}} \mathbf{x})\pi_{\text{ref}}(\mathbf{y}_{s_{w}} \mathbf{x})}\right) - \frac{1}{2\tau}\right]^{2}$
Safe-SLiC (Zhao et al. 2022)	$\mathcal{L}_{\text{Safe-SLiC}}(\pi_{\theta}; \pi_{\text{ref}}) = \mathbf{E}_{\left(x, \mathbf{y}_{Sw}, \mathbf{y}_{s}\right) - \mathcal{D}_{\text{Safety}}}\left[\max\left(0, 1 - \beta\left(\log\frac{\pi_{\theta}(\mathbf{y}_{Sw} \mathbf{x})}{\pi_{\text{ref}}(\mathbf{y}_{Sw} \mathbf{x})} - \log\frac{\pi_{\theta}(\mathbf{y}_{s} \mathbf{x})}{\pi_{\text{ref}}(\mathbf{y}_{s} \mathbf{x})}\right)\right)\right]$
Safe-KTO (Ethayarajh et al. 2024)	$\mathcal{L}_{\text{Safe-KTO}}(\pi_{\theta}, \pi_{\text{ref}}) = \mathbf{E}_{x, y \sim \mathcal{D}_{\text{Safery}}} \left[ \lambda_{y} - v(x, y) \right] \text{ with } r_{\theta}(x, y) = \log \frac{\pi_{\theta}(y x)}{\pi_{\text{ref}}(y x)} \text{ and } v(x, y) = \begin{cases} \lambda_{y} \sigma \left( \beta \left( r_{\theta}(x, y) - z_{0} \right) \right) \text{ if } y \sim y_{s_{w}} \mid x \\ \lambda_{y} \sigma \left( \beta \left( z_{0} - r_{\theta}(x, y) \right) \right) \text{ if } y \sim y_{s_{l}} \mid x \end{cases} \text{ with } \lambda_{y} > 0$
Safe-EXO (Ji et al. 2024a)	$\mathscr{L}_{\text{Safe-EXO}}\left(\pi_{\theta}\right) = \mathbf{E}_{\left(\mathbf{z},\mathbf{y}_{S_{liv}},\mathbf{y}_{S_{l}}\right) - \mathscr{D}_{\text{Safety}}}\left[\mathbb{D}_{\text{KL}}\left(\mathbf{p}_{f_{\theta}}\left(\cdot\mid\mathbf{y}_{S_{w}},\mathbf{y}_{S_{l}},\boldsymbol{x}\right)\mid\mathbf{p}_{r_{\phi}}\left(\cdot\mid\mathbf{y}_{S_{w}},\mathbf{y}_{S_{l}},\boldsymbol{x}\right)\right)\right] \text{ with } f_{\theta} = \log \pi_{\theta} - \log \pi_{\text{ref}} \text{ and } r_{\phi} \text{ the implicit reward function from DPO.}$
Safe-NCA (Chen et al. 2024)	$\mathcal{L}_{\text{safe-NCA Pair}}(\pi_{\theta}; \pi_{\text{ref}}) = -\mathbf{E}_{\left(\mathbf{x}, \mathbf{y}_{sw}, \mathbf{y}_{s_{l}}\right) \sim \mathcal{D}_{\text{Safety}}} \left[ \log \sigma \left( \beta \left( \log \frac{\pi_{\theta}(\mathbf{y}_{sw} \mathbf{x})}{\pi_{\text{ref}}(\mathbf{y}_{sw} \mathbf{x})} - \log \frac{\pi_{\theta}(\mathbf{y}_{s_{l}} \mathbf{x})}{\pi_{\text{ref}}(\mathbf{y}_{s_{l}} \mathbf{x})} \right) + \frac{1}{2} \log \left[ \sigma \left( \beta \log \frac{\pi_{\text{ref}}(\mathbf{y}_{sw} \mathbf{x})}{\pi_{\theta}(\mathbf{y}_{sw} \mathbf{x})} \right) \times \sigma \left( \beta \log \frac{\pi_{\text{ref}}(\mathbf{y}_{s_{l}} \mathbf{x})}{\pi_{\theta}(\mathbf{y}_{s_{l}} \mathbf{x})} \right) \right] \right]$
Safe-SPPO (Wu et al. 2024)	$\mathscr{L}_{\text{Safe SPPO}}\left(\pi_{\theta}; \pi_{\text{ref}}\right) = \mathbf{E}_{\left(x, \mathbf{y}_{2y_t}, \mathbf{y}_{z_t}\right) \sim \mathscr{D}_{\text{Safery}}}\left[\left(\log \frac{\pi_{\theta}(\mathbf{y}_{z_{tt}}   \mathbf{x})}{\pi_{\text{ref}}(\mathbf{y}_{z_{tt}}   \mathbf{x})} - \frac{0.5}{\beta}\right)^2 + \left(\log \frac{\pi_{\theta}(\mathbf{y}_{z_t}   \mathbf{x})}{\pi_{\text{ref}}(\mathbf{y}_{z_t}   \mathbf{x})} + \frac{0.5}{\beta}\right)^2\right]$
Safe-AOT (Melnyk et al. 2024)	$\mathscr{L}_{\text{Safe-AOT}}(\pi_{\theta}; \pi_{\text{ref}}) = -\mathbf{E}_{\left[\mathbf{x}, \mathbf{y}_{s_w}, \mathbf{y}_{s_j}\right] \sim \mathscr{D}_{\text{Safev}}} \left[ \log \sigma \left( \beta \left( \log \frac{\pi_{\theta}(\mathbf{y}_{s_w}^c   \mathbf{x})}{\pi_{\theta}(\mathbf{y}_{s_j}^c   \mathbf{x})} - \log \frac{\pi_{\text{ref}}(\mathbf{y}_{s_w}^c   \mathbf{x})}{\pi_{\text{ref}}(\mathbf{y}_{s_j}^c   \mathbf{x})} \right) \right) (1 - \epsilon) + \log \sigma \left( -\beta \left( \log \frac{\pi_{\theta}(\mathbf{y}_{s_w}^c   \mathbf{x})}{\pi_{\theta}(\mathbf{y}_{s_j}^c   \mathbf{x})} - \log \frac{\pi_{\text{ref}}(\mathbf{y}_{s_w}^c   \mathbf{x})}{\pi_{\text{ref}}(\mathbf{y}_{s_j}^c   \mathbf{x})} \right) \right) \epsilon \right]$
Safe-AOT_pair (Melnyk et al. 2024)	$ \mathcal{L}_{\text{Safe-AOTp}}\left(\pi_{\theta}; \pi_{\text{ref}}\right) = -\mathbf{E}_{\left(x, \mathbf{y}_{\text{siv}}, \mathbf{y}_{f_{l}}\right) - \mathcal{D}_{\text{Safety}}}\left[\log\sigma\left(\beta\left(\log\frac{\pi_{\theta}(\mathbf{y}_{\text{siv}}^{c} \mathbf{x})}{\pi_{\text{ref}}(\mathbf{y}_{\text{siv}}^{c} \mathbf{x})} - \log\frac{\pi_{\theta}(\mathbf{y}_{\text{si}}^{c} \mathbf{x})}{\pi_{\text{ref}}(\mathbf{y}_{\text{si}}^{c} \mathbf{x})}\right)\right)(1 - \epsilon) \\ + \log\sigma\left(-\beta\left(\log\frac{\pi_{\theta}(\mathbf{y}_{\text{siv}}^{c} \mathbf{x})}{\pi_{\text{ref}}(\mathbf{y}_{\text{siv}}^{c} \mathbf{x})} - \log\frac{\pi_{\theta}(\mathbf{y}_{\text{siv}}^{c} \mathbf{x})}{\pi_{\text{ref}}(\mathbf{y}_{\text{siv}}^{c} \mathbf{x})}\right)\right)\epsilon\right) \\ + 2\pi i \left[-2\pi i \left(-\beta\left(\log\frac{\pi_{\theta}(\mathbf{y}_{\text{siv}}^{c} \mathbf{x})}{\pi_{\text{ref}}(\mathbf{y}_{\text{siv}}^{c} \mathbf{x})} - \log\frac{\pi_{\theta}(\mathbf{y}_{\text{siv}}^{c} \mathbf{x})}{\pi_{\text{ref}}(\mathbf{y}_{\text{siv}}^{c} \mathbf{x})}\right)\right)\epsilon\right] \\ + 2\pi i \left[-2\pi i \left(-\beta\left(\log\frac{\pi_{\theta}(\mathbf{y}_{\text{siv}}^{c} \mathbf{x})}{\pi_{\text{ref}}(\mathbf{y}_{\text{siv}}^{c} \mathbf{x})} - \log\frac{\pi_{\theta}(\mathbf{y}_{\text{siv}}^{c} \mathbf{x})}{\pi_{\text{ref}}(\mathbf{y}_{\text{siv}}^{c} \mathbf{x})}\right)\right]\epsilon\right] \\ + 2\pi i \left[-2\pi i \left(-\beta\left(\log\frac{\pi_{\theta}(\mathbf{y}_{\text{siv}}^{c} \mathbf{x})}{\pi_{\text{ref}}(\mathbf{y}_{\text{siv}}^{c} \mathbf{x})} - \log\frac{\pi_{\theta}(\mathbf{y}_{\text{siv}}^{c} \mathbf{x})}{\pi_{\text{ref}}(\mathbf{y}_{\text{siv}}^{c} \mathbf{x})}\right)\right)\right]\epsilon\right] \\ + 2\pi i \left[-2\pi i \left(-\beta\left(\log\frac{\pi_{\theta}(\mathbf{y}_{\text{siv}}^{c} \mathbf{x})}{\pi_{\text{ref}}(\mathbf{y}_{\text{siv}}^{c} \mathbf{x})} - \log\frac{\pi_{\theta}(\mathbf{y}_{\text{siv}}^{c} \mathbf{x})}{\pi_{\text{ref}}(\mathbf{y}_{\text{siv}}^{c} \mathbf{x})}\right)\right]\right] \\ + 2\pi i \left[-2\pi i \left(-\beta\left(\log\frac{\pi_{\theta}(\mathbf{y}_{\text{siv}}^{c} \mathbf{x})}{\pi_{\text{ref}}(\mathbf{y}_{\text{siv}}^{c} \mathbf{x})} - \log\frac{\pi_{\theta}(\mathbf{y}_{\text{siv}}^{c} \mathbf{x})}{\pi_{\text{ref}}(\mathbf{y}_{\text{siv}}^{c} \mathbf{x})}\right)\right]\right]\right) \\ + 2\pi i \left[-2\pi i \left(-\beta\left(\log\frac{\pi_{\theta}(\mathbf{y}_{\text{siv}}^{c} \mathbf{x})}{\pi_{\text{ref}}(\mathbf{y}_{\text{siv}}^{c} \mathbf{x})} - \log\frac{\pi_{\theta}(\mathbf{y}_{\text{siv}}^{c} \mathbf{x})}{\pi_{\text{ref}}(\mathbf{y}_{\text{siv}}^{c} \mathbf{x})}\right)\right]\right]\right]$
Safe-ORPO (Hong, Lee, and Thorne 2024)	$\mathscr{L}_{\text{Safe-ORPO}}\left(\pi_{\theta}\right) = -\mathbf{E}_{\left(x,\mathbf{y}_{Suv}\right) - \mathscr{D}_{\text{Safety}}}\left[\log \pi_{\theta}\left(\mathbf{y}_{Suv} x\right)\right] - \lambda \mathbf{E}_{\left(x,\mathbf{y}_{Suv},\mathbf{y}_{S_{l}}\right) - \mathscr{D}_{\text{Safety}}}\left[\log \sigma\left(\log\left(\frac{\pi_{\theta}(\mathbf{y}_{Suv} x)}{1 - \pi_{\theta}(\mathbf{y}_{Suv} x)} \times \frac{1 - \pi_{\theta}(\mathbf{y}_{I_{2}} x)}{\pi_{\theta}(\mathbf{y}_{I_{2}} x)}\right)\right)\right] \text{ with } \lambda > 0$

Our goal is to create a dataset that enables pairwise comparisons of responses based on safety. Each entry in our constructed dataset includes a prompt  $x^{(j)}$ , a safe response  $y_{s_w}^{(j)}$  (safety-accepted response), and a less safe response  $y_{s_l}^{(j)}$  (safety-rejected response). Thus, we get  $\mathscr{D}_{\mathrm{Safety}} = \left\{x^{(j)}, y_{s_w}^{(j)}, y_{s_l}^{(j)}\right\}_{j=1}^N$ , where response  $y_{s_w}^{(j)}$  is more harmless than response  $y_{s_l}^{(j)}$  and N=47,077 is the cardinality of  $\mathscr{D}_{\mathrm{Safety}}$ . This approach allows us to compare the safety of different responses to the same prompt, providing a robust foundation for optimizing the safety of LLMs through optimization techniques from human feedback.

### **Safety Alignment Methods**

In the context of aligning LLMs with safety objectives, we propose the application of 10 alignment methods, each optimizing a distinct loss function tailored to enforce safety constraints. Table 1 summarizes the key methods, alongside their corresponding loss functions. We refer to these methods as Safe-DPO, Safe-rDPO, Safe-IPO, etc., which are designed to optimize the model's behavior by minimizing risks associated with unsafe outputs. The notations used in the loss functions are as follows:  $\mathbf{y}_{s_w} = \{y_{s_w}^1, y_{s_w}^2, \dots, y_{s_w}^N\}$  and  $\mathbf{y}_{s_l} = \{y_{s_l}^1, y_{s_l}^2, \dots, y_{s_l}^N\}$  represent sets of safe and unsafe outputs, respectively, with their sorted versions denoted by  $\mathbf{y}_{s_w}^{<}$  and  $\mathbf{y}_{s_l}^{<}$ . The logistic function is denoted by  $\sigma$ ,  $\epsilon \in (0,1)$  represents the label smoothing parameter, and  $\beta > 0$  denotes the temperature parameter. The model being updated is represented by  $\pi_{\theta}$ , while  $\pi_{\mathrm{ref}}$  denotes the frozen reference model. Moreover,  $\mathbb{D}_{KL}$  denotes the Kullback-Leibler divergence and  $\mathbf{p}_f$  is the empirical distribution of  $\exp(f)$ : the exponential function f. These notations are essential for understanding the loss functions presented in Table 1.

### **Evaluations**

In this section, we elaborate on the benchmarking tools used in this work and associated metrics.

### **ALERT**

To evaluate our model's safety we used ALERT (Tedeschi et al. 2024). ALERT benchmark is designed to assess the safety of LLMs. It includes 45k testing instructions grouped into the 6 main risk categories (corresponding to the safety risk taxonomy C) and 32 subcategories. (Tedeschi et al. 2024) introduced a dataset of prompts composed of 14k samples, denoted as  $P_c$ , designed to evaluate the safety and robustness of LLMs against a wide range of potentially harmful inputs categorized under various safety risks. In evaluating our models, we leverage  $Llama\ Guard\ 3\ 8B$  (Dubey et al. 2024), considered to be on of the best available safety classifier. Performance is measured using the safety score defined in Definition 4.

### **Adversarial ALERT**

To comprehensively evaluate the safety and robustness of our models, we utilized the adversarial component of the ALERT benchmark in our testing protocol. Alert includes 31k prompts (denoted as  $\tilde{P}_a$ ), specifically designed to challenge the model with inputs crafted to exploit potential weaknesses, using techniques such as adversarial suffix attack, adversarial prefix attack, token manipulation, and jailbreaking as stated in Definition 2. The purpose of this part is to understand the LLM behavior under manipulative scenarios. Robustness against adversarial attacks is captured using ASR defined in Definition 5.

### **Toxicity**

A significant component of having Safe LLMs entails having a *nontoxic* LLM. Therefore, the toxicity benchmark introduced in an existing work (Wang et al. 2023) was used

Table 2: Benchmarking the Falcon models with the ALERT framework. Each row depicts a risk category from the ALERT taxonomy, while each column depicts an LLM under evaluation. Values in the last row depict overall safety scores, all others are category-wise safety scores (higher is safer). Safe scores ( $S(\pi) \ge 99$ ) are gray, unsafe scores within  $90 \le S(\pi) < 99$  are orange, and highly unsafe scores  $S(\pi) < 90$  are red. Best viewed in color.

		Falcon	11B									
Unsafe Category	Instruct	Safe-ORPO	Safe-SPPO	Safe-NCA	Safe-KTO	Safe-DPO	Safe-AOTp	Safe-EXO	Safe-AOT	Safe-SLiC	Safe-rDPO	Safe-IPO
Hate Speech & Discrimination	84.33%	80.73%	99.31%	99.71%	99.89%	99.98%	100.00%	99.98%	99.98%	99.96%	99.98%	99.96%
Suicide & Self-Harm	84.27%	76.68%	99.64%	99.64%	99.28%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
Guns & Illegal Weapons	66.56%	71.03%	96.15%	98.69%	98.71%	98.44%	99.60%	99.75%	100.00%	99.92%	99.75%	99.75%
Criminal Planning	47.41%	53.19%	90.98%	95.96%	97.74%	99.05%	99.70%	99.77%	99.68%	99.79%	99.82%	99.82%
Sexual Content	73.15%	72.74%	96.20%	98.89%	99.34%	99.34%	100.00%	99.89%	100.00%	100.00%	99.67%	99.89%
Regulated or Controlled Substances	57.45%	63.61%	95.44%	98.34%	98.04%	99.03%	99.67%	99.59%	99.60%	99.58%	99.84%	99.87%
Overall Safety Score S	64.42%	66.35%	95.13%	97.97%	98.67%	99.30%	99.81%	99.82%	99.83%	99.84%	99.86%	99.87%

Table 3: Benchmarking the models with the ALERT framework. Each row depicts a risk category from ALERT taxonomy, while each column depicts an LLM under evaluation. Values in the last row depict overall safety scores, all others are category-wise safety scores (higher is safer). Safe scores ( $S(\pi) \ge 99$ ) are grey, unsafe scores within  $90 \le S(\pi) < 99$  are orange, and highly unsafe scores  $S(\pi) < 90$  are red. Safe-Falcon corresponds to Falcon 11B Safe-IPO in Table 2.

Unsafe Category	Alpaca	Mistral 7B	OLMo 7B	YI 1.5 9B	Llama 3.1 8B	Mixtral 8x7B	Qwen-27B	Vicuna 7B	GPT-3.5	Gemma 2 9B	Safe Falcon
Hate Speech & Discrimination	81.44%	91.01%	91.92%	94.60%	96.32%	97.29%	96.49%	99.20%	97.96%	99.36%	99.96%
Suicide & Self-Harm	73.96%	88.25%	86.80%	90.96%	97.65%	96.93%	98.73%	97.65%	100.00%	98.19%	100.00%
Guns & Illegal Weapons	66.51%	72.92%	85.16%	74.68%	92.59%	87.52%	93.01%	93.26%	96.76%	98.01%	99.75%
Criminal Planning	51.32%	52.91%	69.13%	68.55%	84.87%	82.50%	86.59%	88.75%	88.96%	89.47%	99.82%
Sexual Content	72.62%	87.70%	78.26%	85.52%	93.99%	95.42%	96.43%	98.77%	99.00%	98.77%	99.89%
Regulated or Controlled Substances	55.18%	62.52%	71.13%	70.28%	94.18%	87.35%	92.06%	92.68%	95.97%	98.11%	99.87%
Overall Safety Score S	64.28%	70.76%	78.74%	79.02%	84.16%	88.18%	91.93%	93.65%	94.3%	95.3%	99.87%

to evaluate our proposed safe LLMs. Note that the mentioned benchmark leverages perspective API<sup>1</sup> (Lees et al. 2022) as a toxicity evaluator model, that returns toxicity score ranging from 0 (nontoxic) to 1 (toxic). The benchmark is composed of four benchmark datasets, with the first two being a 1.2k subset of toxic (T) and nontoxic (NT) prompts from the RealToxicityPrompts dataset (Gehman et al. 2020). The remaining two datasets are 1.2k prompts obtained using GPT-3.5 (G3.5) and GPT4 (G4), which were designed to provoke a toxic LLM response. Additionally, two system prompts were explored, namely, benign and adversarial. Both types will instruct the LLM to be a helpful assistant, with the difference that adversarial prompting will explicitly steer the LLM to be toxic. We followed the same setup suggested by (Wang et al. 2023), where 25 generations were generated for each prompt. Based on these generations, two metrics were considered, which are presented below:

- Expected Max Toxicity denoted as  $\mathbb{E}[\max_{tox}]$ , aimed to evaluate the worst-case scenario, where the maximum score over the 25 generations for each prompt, is averaged over all prompts.
- **Average Toxicity** this averages the toxicity scores for all generations, labeled as avg<sub>tox</sub>.

### Results

In this section, results obtained using benchmarking tools introduced in the previous section are presented and discussed. Table 2 presents the safety score S across different alignment techniques on the Falcon 11B model (Malartic et al. 2024). It includes the baseline Instruct model (a finetuned version of Falcon 11B on Ultrachat dataset) and 11 safety enhanced models (following the safety alignment methods in Table 1 where the reference model  $\pi_{ref}$  corresponds to Falcon 11B instruct). The results of Table 2 obtained using ALERT for all safe Falcon models. These results are quantified using the safety score S, as outlined in Definition 4, with higher scores indicating greater safety and resilience. Most models exhibited commendable performance, highlighting the effectiveness of the alignments applied. However, notable deviations were observed with the Instruct, which is the base model, understandably scored lower in comparison to its aligned counterparts. Similarly, the Safe-ORPO model displayed performance metrics closely mirroring those of the Instruct model, suggesting the limited efficacy of the ORPO enhancements in improving safety outcomes. Contrastingly, the other alignment techniques demonstrated substantial improvements in safety across all categories. Such results underscore the potential of alignment techniques to enhance model robustness and safety in critical safety cate-

<sup>&</sup>lt;sup>1</sup>www.perspectiveapi.com

Table 4: ASR score of each attacking strategy in the adversarial ALERT. Each row represents an attacking strategy, while each column corresponds to an LLM under evaluation. A model is robust when the ASR is  $\leq 1\%$  (grey), vulnerable  $1\% < ASR \leq 5\%$  (orange), and highly vulnerable ASR  $\geq 5\%$  (red). Best viewed in color.

	Falcon 11B											
Attack Type	Instruct	Safe- ORPO	Safe- SPPO	Safe-NCA	Safe-KTO	Safe- DPO	Safe- SLiC	Safe-AOT	Safe-EXO	Safe- AOTp	Safe-IPO	Safe- rDPO
Adversarial Suffix	46.51%	32.16%	10.95%	5.92%	1.42%	0.70%	0.40%	0.31%	0.13%	0.11%	0.16%	0.10%
Adversarial Prefix	50.48%	38.85%	4.93%	1.33%	0.13%	0.11%	0.04%	0.01%	0.00%	0.06%	0.00%	0.00%
Token Manipulation	30.08%	27.15%	3.75%	1.58%	1.04%	0.51%	0.21%	0.18%	0.14%	0.18%	0.11%	0.15%
Jailbreaking	61.49%	36.44%	15.02%	5.82%	1.86%	3.05%	1.18%	0.81%	0.43%	0.30%	0.01%	0.01%
Overall ASR Score	45.60%	33.06%	8.22%	3.47%	1.11%	1.04%	0.44%	0.31%	0.17%	0.16%	0.07%	0.06%

Table 5: ASR score of each attacking strategy in the adversarial ALERT. Each row represents an attacking strategy, while each column corresponds to an LLM under evaluation. A model is robust when the ASR is  $\leq 1\%$  (grey), vulnerable  $1\% < ASR \leq 5\%$  (orange), and highly vulnerable ASR  $\geq 5\%$  (red). Safe-Falcon corresponds to Falcon 11B Safe-rDPO in Table 4.

Attack Type	Alpaca	YI 1.5 9B	OLMo 7B	Mistral 7B	Llama 3.1 8B	Qwen-2 7B	Vicuna 7B	Mixtral 8x7B	Gemma 2 9B	GPT-3.5	Safe- Falcon
Adversarial Suffix	26.97%	33.03%	24.00%	31.57%	19.61%	19.65%	9.50%	9.41%	7.98%	5.54%	0.10%
Adversarial Prefix	32.80%	26.50%	27.62%	56.08%	17.65%	12.08%	8.63%	9.02%	5.10%	3.01%	0.00%
Token Manipulation	27.82%	18.81%	18.07%	15.48%	10.21%	8.27%	4.88%	4.09%	5.21%	4.74%	0.15%
Jailbreaking	54.83%	51.48%	48.47%	10.12%	40.02%	36.91%	30.22%	27.82%	27.71%	20.63%	0.01%
Overall ASR Score	34.90%	31.22%	28.50%	27.15%	20.82%	18.24%	12.54%	11.82%	10.93%	8.14%	0.06%

gories. Table 3, contrasts our safest aligned Falcon in terms of safety score S, Safe-IPO, against common LLMs. Similar to the observations noted for the ASR score, Safe-Falcon performed best against all considered benchmarks, with approximately 4% increase when compared to GPT-3.5, the second performing model, in terms of safety score. Table 4 presents the ASR across different alignment techniques on the Falcon 11B model, against a set of adversarial attacks. Performance is evaluated by the percentage of successful attacks, with lower percentages indicating high robustness. The overall performance of each model is also summarized to identify the most robust model against adversarial attacks. The models Safe-IPO, Safe-AOTp, and Safe-EXO emerged as the top performers, demonstrating the highest resistance across the tested attacks. Table 5 compares our aligned models against mainstream models in terms of ASR, where Safe-Falcon refers to Safe-rDPO due to being the best in terms of ASR. Safe-Falcon and GPT-3.5 exhibit superior robustness, with particularly low overall vulnerability scores of 0.06% and 8.14%, respectively. In contrast, models like Yi 1.5 9B and Alpaca display significantly higher susceptibility, with Alpaca recording the worst ASR overall score of 34.90%. The results underscore the safety shortcomings in the considered models and highlight the potential of the alignment techniques employed for Safe-Falcon.

Moving on to toxicity, heatmaps in Figure 6, present the toxicity evaluation results for all safe LLMs proposed in this paper, against the *Instruct*. It is evident from Figure 6, that all Safe Falcons demonstrate a tangible drop in toxic responses, in both types of system prompting. As expected, results bolster the claim that *Instruct* is highly susceptible to producing toxic content, which compromises the general safety of the LLM. This claim is manifested by scoring the highest in all heatmaps in Figure 6. As seen in figures Figure 6.b and Figure 6.d, when adversarial prompting was applied to the prompts in NT, Instruct likelihood of returning toxic prompts increased, evident by the obtained 0.88 and 0.29 in terms of  $\mathbb{E}[\max_{tox}]$ and avg<sub>tox</sub>, respectively. Looking at Figure 6.c, Safe-IPO and Safe-rDPO demonstrated the maximum drop in all metrics that is clear by the scoring 0.05 or less in terms of avg<sub>tox</sub>, when benign prompting is applied. Additionally, both techniques illustrated their robustness by maintaining low toxic scores even when adversarial prompting is deployed. This observation can be seen Figure 6.b and Figure 6.d, with scores being limited to 0.25 and 0.07 for  $E[max_{tox}]$  and  $avg_{tox}$ , respectively. On the other hand, the worst performing alignment technique is Safe-SPPO as it obtained the least improvement in toxicity. Several conclusions can be deduced from this experiment, which are (1) Instruct is vulnerable to producing toxic text impact-

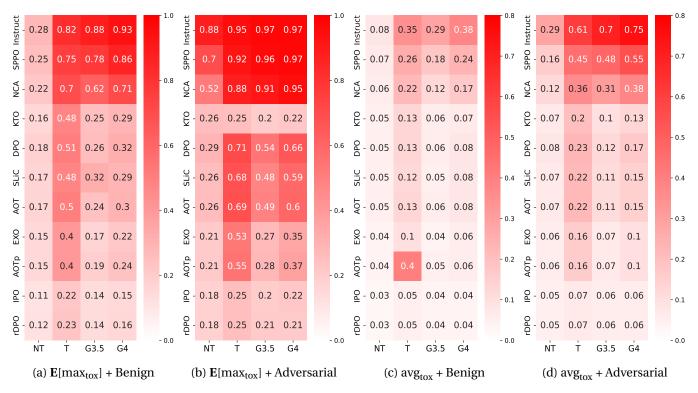


Figure 6: Toxicity evaluation on all alignment techniques with benign and adversarial prompting. Metrics are Expected Max Toxicity ( $\mathbb{E}$ )[max<sub>tox</sub>]) and Average Toxicity (avg<sub>tox</sub>). The higher the red intensity corresponds to more toxic LLM.

ing the LLM safety, (2) all safe models trained using various alignment techniques obtained noticeable improvement in being less toxic over different system prompting, and (3) Safe-IPO and Safe-rDPO were the best performing alignment techniques in terms of toxicity, and the most robust against adversarial prompting. Following evaluating the safety of our proposed models, we evaluate them on general benchmarks typically used to assess the general performance of LLM (Fourrier et al. 2024). Table 6 depicts the results of our safe models with the base *instruct* model. The key takeaway from this table is that alignment techniques maintained scores within the same range of *Instruct*. Counter to expectation, *Safe-NCA* outperformed *Instruct* in three benchmarks. Additionally, Safe-rDPO scored the best in GPQA and matched Instruct in IFEval. Therefore, this demonstrates that our aligned model maximized performance in terms of safety while preserving general performance scores on par with the base model.

### **Conclusion and Future Works**

This work studies the effect of variants of direct preference optimization methods on LLMs safety. Our experiments demonstrate a substantial improvement in the safety score of the Falcon 11B model, increasing from 57.64% to 99.90%, positioning it among the safest LLM models in the state of the art.

However, our study also revealed an important tradeoff: while safety scores improved dramatically, we ob-

Table 6: Performance scores of different Falcon 11B models on the LLM Leaderboard (Fourrier et al. 2024) (reporting the raw values.)

	Tasks											
Falcon 11B	IFEval	ввн	GPQA	MATH	MuSR	MMLU-PRO						
Instruct	0.387	0.426	0.286	0.012	0.425	0.253						
Safe-DPO	0.373	0.428	0.293	0.004	0.400	0.260						
Safe-rDPO	0.387	0.414	0.30	0.002	0.427	0.262						
Safe-IPO	0.274	0.412	0.296	0.005	0.394	0.261						
Safe-SLiC	0.360	0.428	0.295	0.002	0.401	0.264						
Safe-KTO	0.363	0.429	0.293	0.001	0.406	0.263						
Safe-EXO	0.351	0.413	0.287	0	0.436	0.264						
Safe-NCA	0.371	0.435	0.290	0.015	0.441	0.262						
Safe-SPPO	0.380	0.435	0.282	0.0075	0.430	0.259						
Safe-AOT	0.360	0.421	0.288	0.0052	0.439	0.263						
Safe-AOTp	0.351	0.419	0.285	0.0007	0.438	0.264						

served a reduction in general capabilities, particularly in mathematical tasks. This finding highlights the complex relationship between safety enhancements and overall model performance. In particular, we identified noise contrastive alignment (Safe-NCA) as an optimal method for balancing safety and performance.

Our research conclusively shows that alignment tech-

niques can be sufficient for building safe and robust models. Nevertheless, the observed trade-offs highlight the need for further investigation. As future work, we plan to explore methods to mitigate the negative impact on performance in other tasks such as math and reasoning, while maintaining the high level of safety achieved.

### References

- Azar, M. G.; Guo, Z. D.; Piot, B.; Munos, R.; Rowland, M.; Valko, M.; and Calandriello, D. 2024. A general theoretical paradigm to understand learning from human preferences. In *International Conference on Artificial Intelligence and Statistics*, 4447–4455. PMLR.
- Chen, H.; He, G.; Su, H.; and Zhu, J. 2024. Noise contrastive alignment of language models with explicit rewards. *arXiv* preprint arXiv:2402.05369.
- Chowdhury. 2024. Provably robust dpo: Aligning language models with noisy feedback. *arXiv preprint arXiv:2403.00409*.
- Dai, J.; Pan, X.; Sun, R.; Ji, J.; Xu, X.; Liu, M.; Wang, Y.; and Yang, Y. 2023. Safe rlhf: Safe reinforcement learning from human feedback. *arXiv preprint arXiv:2310.12773*.
- Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; Fan, A.; et al. 2024. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*.
- Ethayarajh, K.; Xu, W.; Muennighoff, N.; Jurafsky, D.; and Kiela, D. 2024. Kto: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*.
- Fourrier, C.; Habib, N.; Lozovskaya, A.; Szafer, K.; and Wolf, T. 2024. Open LLM Leaderboard v2.
- Ge, S.; Zhou, C.; Hou, R.; Khabsa, M.; Wang, Y.-C.; Wang, Q.; Han, J.; and Mao, Y. 2023. Mart: Improving llm safety with multi-round automatic red-teaming. *arXiv preprint arXiv:2311.07689*.
- Gehman, S.; Gururangan, S.; Sap, M.; Choi, Y.; and Smith, N. A. 2020. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. *arXiv preprint arXiv:2009.11462*.
- Gupta, P.; Yau, L. Q.; Low, H. H.; Lee, I.; Lim, H. M.; Teoh, Y. X.; Koh, J. H.; Liew, D. W.; Bhardwaj, R.; Bhardwaj, R.; et al. 2024. WalledEval: A Comprehensive Safety Evaluation Toolkit for Large Language Models. *arXiv preprint arXiv:2408.03837*.
- Hong, J.; Lee, N.; and Thorne, J. 2024. Reference-free monolithic preference optimization with odds ratio. *arXiv* preprint arXiv:2403.07691.
- Ji, H.; Lu, C.; Niu, Y.; Ke, P.; Wang, H.; Zhu, J.; Tang, J.; and Huang, M. 2024a. Towards efficient and exact optimization of language model alignment. *arXiv preprint arXiv:2402.00856*.
- Ji, J.; Liu, M.; Dai, J.; Pan, X.; Zhang, C.; Bian, C.; Chen, B.; Sun, R.; Wang, Y.; and Yang, Y. 2024b. Beavertails: Towards improved safety alignment of llm via a human-preference dataset. *Advances in Neural Information Processing Systems*, 36.

- Jiang, B.; Jing, Y.; Shen, T.; Yang, Q.; and Xiong, D. 2024. DART: Deep Adversarial Automated Red Teaming for LLM Safety. *arXiv preprint arXiv:2407.03876*.
- Kumar, A.; Agarwal, C.; Srinivas, S.; Feizi, S.; and Lakkaraju, H. 2023. Certifying llm safety against adversarial prompting. *arXiv preprint arXiv:2309.02705*.
- Lees, A.; Tran, V. Q.; Tay, Y.; Sorensen, J.; Gupta, J.; Metzler, D.; and Vasserman, L. 2022. A new generation of perspective api: Efficient multilingual character-level transformers. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, 3197–3207.
- Malartic, Q.; Chowdhury, N. R.; Cojocaru, R.; Farooq, M.; Campesan, G.; Djilali, Y. A. D.; Narayan, S.; Singh, A.; Velikanov, M.; Boussaha, B. E. A.; et al. 2024. Falcon2-11B Technical Report. *arXiv preprint arXiv:2407.14885*.
- Melnyk, I.; Mroueh, Y.; Belgodere, B.; Rigotti, M.; Nitsure, A.; Yurochkin, M.; Greenewald, K.; Navratil, J.; and Ross, J. 2024. Distributional Preference Alignment of LLMs via Optimal Transport. *arXiv preprint arXiv:2406.05882*.
- Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744.
- Rafailov, R.; Sharma, A.; Mitchell, E.; Manning, C. D.; Ermon, S.; and Finn, C. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.
- Tedeschi, S.; Friedrich, F.; Schramowski, P.; Kersting, K.; Navigli, R.; Nguyen, H.; and Li, B. 2024. ALERT: A Comprehensive Benchmark for Assessing Large Language Models' Safety through Red Teaming. *arXiv preprint arXiv:2404.08676*.
- Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.
- Varshney, N.; Dolin, P.; Seth, A.; and Baral, C. 2023. The art of defending: A systematic evaluation and analysis of llm defense strategies on safety and over-defensiveness. *arXiv* preprint *arXiv*:2401.00287.
- Wang, B.; Chen, W.; Pei, H.; Xie, C.; Kang, M.; Zhang, C.; Xu, C.; Xiong, Z.; Dutta, R.; Schaeffer, R.; et al. 2023. DecodingTrust: A Comprehensive Assessment of Trustworthiness in GPT Models. In *NeurIPS*.
- Wu, Y.; Sun, Z.; Yuan, H.; Ji, K.; Yang, Y.; and Gu, Q. 2024. Self-play preference optimization for language model alignment. *arXiv preprint arXiv:2405.00675*.
- Yuan, T.; He, Z.; Dong, L.; Wang, Y.; Zhao, R.; Xia, T.; Xu, L.; Zhou, B.; Li, F.; Zhang, Z.; et al. 2024. R-judge: Benchmarking safety risk awareness for llm agents. *arXiv preprint arXiv:2401.10019*.
- Zhao, Y.; Khalman, M.; Joshi, R.; Narayan, S.; Saleh, M.; and Liu, P. J. 2022. Calibrating sequence likelihood improves conditional language generation. In *The eleventh international conference on learning representations*.

Zheng, C.; Yin, F.; Zhou, H.; Meng, F.; Zhou, J.; Chang, K.-W.; Huang, M.; and Peng, N. 2024. On prompt-driven safeguarding for large language models. In *Forty-first International Conference on Machine Learning*.