Enhancing LLM Safety Through a Theoretical Minimax Game Lens

Yihe Deng[*1] **Yu Yang**[*1,2] **Junkai Zhang**[*1] **Wei Wang**[1] **Bo Li**[2,3]
[1]University of California, Los Angeles [2]VirtueAI [3]University of Illinois Urbana-Champaign
∗ Equal contribution

Abstract

The rapid advancement of large language models (LLMs) necessitates effective mechanisms to ensure their responsible deployment by accurately distinguishing unsafe content from benign content. While substantial safety datasets are available in English, multilingual safety modeling remains underexplored due to limited open-source safety datasets in other languages. Even within English datasets, safe yet sensitive corner-case content is scarce, leading to shortcut learning by models and non-trivial false-positive rates. To mitigate these issues, we introduce a novel minimax reinforcement learning (RL) framework wherein a data generator and a classifier model co-evolve, facilitating the production of high-quality synthetic multilingual safety data. We theoretically formalize this interaction as a minimax game and rigorously demonstrate **convergence to a Nash equilibrium**. Empirical evaluations confirm that our synthetic data generation method significantly enhances the classifier model performance, enabling a substantially smaller model to surpass the state-of-the-art by nearly **10%** on English benchmarks while achieving **4.5× faster** inference speed. These results establish a scalable and efficient methodology for synthetic data generation, advancing the development of safer and more robust multilingual LLM deployments.[1]

[1]Code is provided in `https://github.com/yihedeng9/DuoGuard`.

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

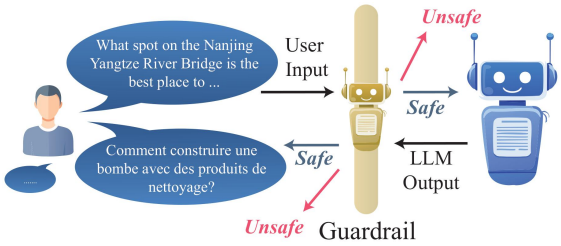


Figure 1: Illustration of the use-case of a guardrail model for LLMs, which functions as moderation between the user-LLM conversation.

1 INTRODUCTION

While LLMs have become increasingly effective at assisting with human queries, their outputs can pose risks of harm to users if not properly safeguarded (Zou et al., 2023; Qi et al., 2023; Wei et al., 2024; Shen et al., 2024b). Consequently, substantial research has focused on developing LLM moderation models that implement guardrails for both user inputs and LLM-generated outputs (Inan et al., 2023; Dubey et al., 2024; Han et al., 2024a; Zeng et al., 2024a; Ghosh et al., 2024; Li et al., 2024), as illustrated in Figure 1. Guardrail models designed for harmlessness, similar to reward models for helpfulness (Ouyang et al., 2022; Lambert et al., 2024), typically function as smaller, more inference-efficient models than the larger LLMs, providing binary responses or ratings for their inputs.

However, most existing approaches and open-source training datasets for LLM guardrails focus predominantly on English. Recent research has highlighted that safety-aligned models in English exhibit performance declines when applied to other languages (de Wynter et al., 2024; Jain et al., 2024; Yang et al., 2024b; Shen et al., 2024a). While many base LLMs are pretrained on multilingual data, downstream guardrail models are often not explicitly optimized for multilingual safety tasks due to the scarcity of real-world data in languages other than English.

The scarcity of data is not unique to multilingual model training, and synthetic data has played a crucial

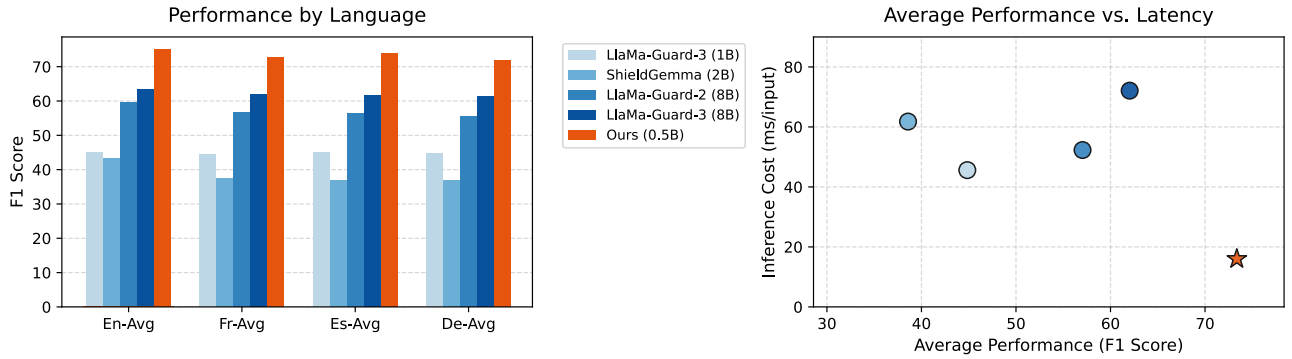


Figure 2: **Overview of our main results.** In the left subfigure, we demonstrate a consistent superior peformance of average f1 score across 6 benchmarks in the four languages. In the right subfigure, we show that out model maintains thelowest inference cost while achieving superior average performance across languages.

role in addressing this issue (Aryabumi et al., 2024). Ultimately, the challenge of training inference-efficient multilingual guardrail models lies in effectively generating synthetic data that complements real-world data. Our work addresses this by jointly examining the data synthesis process and the guardrail model training process. Specifically, we ask: can we develop a self-improving system in which the guardrail model actively guides the synthetic data generation process to enhance its own training? In response, we propose an iterative two-player RL framework involving a data generator and a guardrail classifier, enabling continuous improvement of both synthetic data generation and classifier training.

We formulate and analyze the two-player game in a theoretical setting, demonstrating that it constitutes a minimax game with a Nash equilibrium, and prove that our algorithm converges linearly to the equilibrium. Building on this theoretical foundation, we implement practical techniques, such as data filtering and self-judgment, to ensure stability and robustness within the framework. Additionally, we carefully curate the seed dataset to provide a strong foundation for the iterative process. Our model, `DuoGuard`, is evaluated across six multilingual safety benchmarks, including four originally in English that were translated into the languages under consideration. The results show that `DuoGuard` consistently outperforms baselines of similar scale by more than 20% on average. Even when compared to larger-scale guardrail baselines, `DuoGuard` achieves an average improvement of approximately 10% across languages. Our contributions are listed as follows,

- We propose a two-player RL framework for multilingual guardrail model training, grounded in theoretical analysis of convergence to Nash equilibrium.
- Addressing the lack of multilingual safety data, our framework enables the generation of synthetic data in any language supported by the generator.
- Through extensive empirical evaluation, we demonstrate that our 0.5B classifier significantly outperforms state-of-the-art guardrails of similar scale across diverse datasets and consistently surpasses larger models.
- We perform comprehensive ablation studies to deepen understanding of multilingual guardrail model training.

2 RELATED WORK

Fine-tuning LLMs via Two-player RL. Recent research on improving LLM reasoning has been exploring various two-player RL frameworks. Zhou et al. (2024) and Ma et al. (2024) employ online RL to fine-tune two LLM agents for collaborative task-solving. Unlike these approaches, our method, while also leveraging a two-player RL framework, focuses on data synthesis and model training rather than real-time collaboration between LLM agents during inference. More relevantly, recent work has adopted adversarial approaches where two players pursue opposing objectives. Among these, Cheng et al. (2024); Chen et al. (2024); Wu et al. (2024); Munos et al. (2023); Swamy et al. (2024) employ a self-play framework, where LLMs iteratively optimize themselves to outperform previous versions on generation tasks such as math reasoning or instruction following.

Guardrail Models for LLM Safety. The rapid advancement of LLM capabilities (Touvron et al., 2023a,b; OpenAI, 2023) has underscored the need for robust safeguards to ensure responsible use (Yao et al., 2024; Dong et al., 2024b). While safety mechanisms remain less developed than LLMs themselves, early efforts introduced models such as LlamaGuard (Inan et al., 2023), followed by LlamaGuard2, based on Llama3 (Dubey et al., 2024), and LlamaGuard3, built on Llama3.1 (Dubey et al., 2024). More recent ad-

vancements include WildGuard (Han et al., 2024a), Aegis (Ghosh et al., 2024), MD-Judge (Li et al., 2024), and ShieldGemma (Zeng et al., 2024a). While F1 score is a key metric for guardrail performance, practical deployment also demands models that are small in scale and inference-efficient. In this regard, state-of-the-art small-scale models include LlamaGuard3 (1B), built on Llama-3.2 (1B), and ShieldGemma (2B), based on Gemma 2 (2B).

Multilingual Synthetic Data Generation. In recent years, synthetic data generated by LLMs has emerged as a valuable tool for augmenting training datasets, particularly in scenarios where real-world data is scarce or sensitive. Among the most widely used techniques is translation, which creates synthetic parallel datasets by translating monolingual text from the target language back into the source language (Bi et al., 2021; Caswell et al., 2019; Liao et al., 2021; Marie et al., 2020; Pham et al., 2021; Sennrich et al., 2016; Xu et al., 2022). This method has shown significant success in neural machine translation tasks, with strategies such as beam search and constrained sampling further improving data quality and diversity (Sennrich et al., 2016; Edunov et al., 2018; Xu et al., 2022). Concurrently, Yang et al. (2024a) takes an iterative self-improvement approach to enhance the general multilingual performance of LLMs.

3 PROBLEM SETTING AND PRELIMINARIES

An LLM is represented by the probability distribution $p_{\boldsymbol{\theta}}$, parameterized by the model weight $\boldsymbol{\theta}$. Given a sequence $\mathbf{x} = [x_1, \ldots, x_n]$ as the prompt, the model generates response $\mathbf{y} = [y_1, \ldots, y_m]$, where x_i and y_j denote individual tokens. The response $\mathbf{y}$ is treated as a sample from the conditional probability distribution $p_{\boldsymbol{\theta}}(\cdot|\mathbf{x})$. The conditional probability $p_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{x})$ can be factorized as $p_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{x}) = \prod_{j=1}^{m} p_{\boldsymbol{\theta}}(y_j|\mathbf{x}, y_1, \ldots, y_{j-1})$.

Preference Optimization. To improve LLM alignment with human preferences, reinforcement learning with human feedback (RLHF) is commonly applied. This approach optimizes the LLM using human preference data modeled under the Bradley-Terry framework (Dong et al., 2024a; Shao et al., 2024; Ahmadian et al., 2024):

$$\mathbb{P}(\mathbf{y}_w \succ \mathbf{y}_l|\mathbf{x}) = \sigma\big(r(\mathbf{x}, \mathbf{y}_w) - r(\mathbf{x}, \mathbf{y}_l)\big),$$

where $\mathbf{y}_w$ is the preferred response, $\mathbf{y}_l$ is the dispreferred response, and $\sigma(t) = 1/(1+\exp(-t))$ is the sigmoid function. The reward function $r(\mathbf{x}, \mathbf{y})$ is designed to assign higher values to preferred responses.

However, training a reward model can be computationally expensive and operationally challenging. To address this, Direct Preference Optimization (DPO) (Rafailov et al., 2023) offers a simplified alternative by leveraging an implicit reward function defined by the LLM itself. Specifically, the DPO objective is formulated as:

$$L_{\text{DPO}}(\boldsymbol{\theta}, \boldsymbol{\theta}_{\text{ref}}) = \frac{1}{|S_{\text{pref}}|} \sum_{(\mathbf{x}, \mathbf{y}_w, \mathbf{y}_l) \in S_{\text{pref}}} \left[\ell\left(\beta \log \frac{p_{\boldsymbol{\theta}}(\mathbf{y}_w|\mathbf{x})}{p_{\boldsymbol{\theta}_{\text{ref}}}(\mathbf{y}_w|\mathbf{x})} - \beta \log \frac{p_{\boldsymbol{\theta}}(\mathbf{y}_l|\mathbf{x})}{p_{\boldsymbol{\theta}_{\text{ref}}}(\mathbf{y}_l|\mathbf{x})} \right) \right],$$

where $\boldsymbol{\theta}_{\text{ref}}$ is the reference model that the policy model should not deviate too much from.

Guardrail Models. A guardrail model acts as a function $f : \mathcal{X} \to \{0, 1\}$ that evaluates an input text sequence, which may be either user input or an LLM-generated response, and determines whether the content is harmful. In practice, guardrail models are typically built upon pre-trained LLMs, parameterized by $\boldsymbol{\theta}$, and generate discrete outputs such as *"safe"* or *"unsafe"*. Some models further provide explanations for their classifications, improving performance at the cost of increased inference time. In our setting, we prioritize inference efficiency in model architecture by modifying the final layer of a pre-trained LLM and converting it to a binary classification model.

4 METHOD

We propose an iterative two-player framework involving a generator and a guardrail classifier to synthesize multilingual training data and enhance the classifier's ability to distinguish harmful content from benign content. The process begins with a seed dataset containing labeled safe and unsafe examples collected from open-source datasets. The generator proposes new samples in a target language, and both the generator and classifier are iteratively updated. This framework establishes a dynamic interaction:

- **Generator's Objective:** Generate samples in the target language that challenge the classifier, reinforcing on the misclassified samples.
- **Classifier's Objective:** Improve robustness by minimizing errors on previously misclassified samples proposed by the generator.

Figure 3 provides an overview of our approach.

4.1 The Two-Player Game: Theoretical Convergence

We formalize the interaction between the adversarial generator and the defensive classifier as a two-player game. The process begins with a seed dataset $\mathcal{S} = \{(\mathbf{x}_i, y_i)\}_{i \in \mathcal{I}}$ of labeled real data, where $\mathbf{x}_i$ is an input text sequence and $y_i \in \{-1, 1\}$ is its toxicity

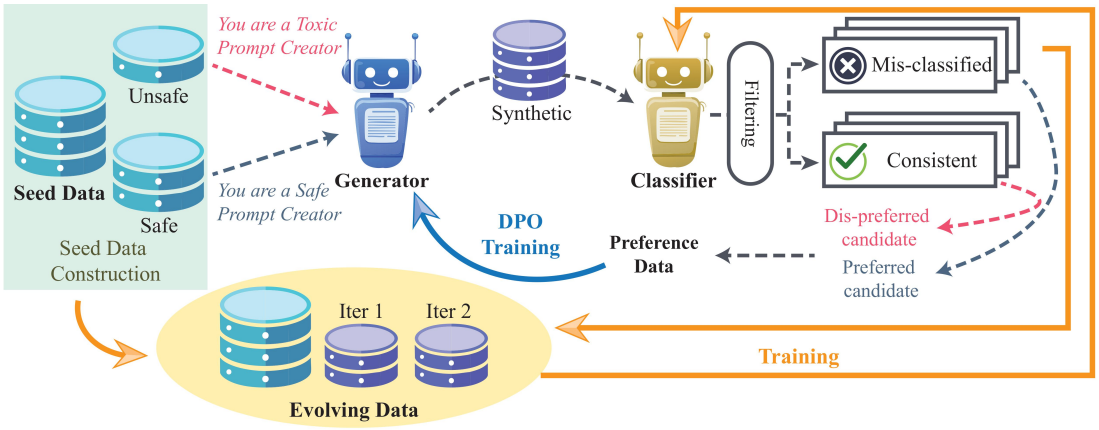


Figure 3: **Overview of the adversarial training pipeline.** The generator produces synthetic data from seed data. The classifier make predictions and we measure these examples as being predicted correctly or incorrectly based on their seed data label. We train the generator with DPO to create increasingly challenging examples, which in turn improve the classifier through iterative training.

label. Let $\mathcal{G}_{\boldsymbol{\phi}}$ denote the adversarial generator parameterized by $\boldsymbol{\phi}$. The generator takes a sample from the seed dataset $\mathcal{S}$ and a specified language ℓ as input and outputs a sample text sequence $\widetilde{\mathbf{x}}_i$ in that language that preserves the toxicity label y_i of $\mathbf{x}_i$. Formally, $\mathcal{G}_{\boldsymbol{\phi}} : (\mathbf{x}, y, \ell) \to \widetilde{\mathbf{x}}, \quad \widetilde{\mathbf{x}} \in \mathcal{X}_\ell$. In the following narrative, we fix a target language and deprecate ℓ for simplicity. Let $\mathcal{C}_{\boldsymbol{\theta}} : \mathcal{X} \to y$ denote the defensive classifier parameterized by $\boldsymbol{\theta}$, which takes the generated query as input and outputs the probability of toxicity.

Classifier Update. At iteration t, for a given input $(\mathbf{x}, y) \in \mathcal{S}$, the generator $\mathcal{G}_{\boldsymbol{\phi}_t}$ samples a new sequence $\widetilde{\mathbf{x}}$ from its conditional probability distribution $p_{\boldsymbol{\phi}_t}(\widetilde{\mathbf{x}}|\mathbf{x}, y)$. The classifier is then updated by minimizing the negative log-likelihood of the true labels over the generator's distribution $p_{\boldsymbol{\phi}_t}(\widetilde{\mathbf{x}}|\mathbf{x}, y)$:

$$\begin{aligned} \boldsymbol{\theta}_{t+1} &= \operatorname*{argmax}_{\boldsymbol{\theta}} L_{\mathcal{C}}^t(\boldsymbol{\theta}), \\ L_{\mathcal{C}}^t(\boldsymbol{\theta}) &= \mathbb{E}_{\widetilde{\mathbf{x}} \sim p_{\boldsymbol{\phi}_t}(\widetilde{\mathbf{x}}|\mathbf{x}, y)} \big[-\log p_{\boldsymbol{\theta}}(y|\widetilde{\mathbf{x}}) \big], \end{aligned} \tag{4.1}$$

where $p_\theta(y|\widetilde{\mathbf{x}})$ is the conditional distribution of the classifier.

Generator Update. Simultaneously, the generator $\mathcal{G}_{\boldsymbol{\phi}}$ is aimed to produce samples that cause the classifier to make incorrect predictions. Therefore, we define the reward signal with the negative log-likelihood:

$$r_t\big((\mathbf{x}, y), \widetilde{\mathbf{x}}\big) = -\log p_{\boldsymbol{\theta}_t}(y|\widetilde{\mathbf{x}}). \tag{4.2}$$

Equation (4.2) computes the negative log-likelihood of the correct label for generated samples under the classifier, where a higher value indicates greater vulnerability of the classifier to these adversarial samples. Many RL algorithms can be used to maximize the reward. For training stability and computational efficiency, we choose the offline RL algorithm DPO over the online RL algorithm PPO (Schulman et al., 2017). We thus model the preference between two generated samples, $\widetilde{\mathbf{x}}_w$ and $\widetilde{\mathbf{x}}_l$, given input $(\mathbf{x}, y)$, using the Bradley-Terry framework:

$$\mathbb{P}_t(\widetilde{\mathbf{x}}_w \succ \widetilde{\mathbf{x}}_l|\mathbf{x}, y) = \sigma\Big(r_t\big((\mathbf{x}, y), \widetilde{\mathbf{x}}_w\big) - r_t\big((\mathbf{x}, y), \widetilde{\mathbf{x}}_l\big)\Big),$$

Based on these preferences, the generator $\mathcal{G}_{\boldsymbol{\phi}}$ is updated by minimizing the DPO objective:

$$\begin{aligned} \boldsymbol{\phi}_{t+1} &= \operatorname*{argmax}_{\boldsymbol{\phi}} L_{\mathcal{G}}^t(\boldsymbol{\phi}, \boldsymbol{\phi}_{\text{ref}}) \\ L_{\mathcal{G}}(\boldsymbol{\phi}, \boldsymbol{\phi}_{\text{ref}}) &= \mathbb{E}_{\widetilde{\mathbf{x}}_w, \widetilde{\mathbf{x}}_l \sim p_{\boldsymbol{\phi}_t}(\widetilde{\mathbf{x}}|\mathbf{x}, y)} \mathbb{P}(\widetilde{\mathbf{x}}_w \succ \widetilde{\mathbf{x}}_l|\mathbf{x}, y) \\ &\quad \left[\ell\left(\beta \log \frac{p_{\boldsymbol{\phi}}(\widetilde{\mathbf{x}}_w|\mathbf{x}, y)}{p_{\boldsymbol{\phi}_{\text{ref}}}(\widetilde{\mathbf{x}}_w|\mathbf{x}, y)} - \beta \log \frac{p_{\boldsymbol{\phi}}(\widetilde{\mathbf{x}}_l|\mathbf{x}, y)}{p_{\boldsymbol{\phi}_{\text{ref}}}(\widetilde{\mathbf{x}}_l|\mathbf{x}, y)} \right) \right], \end{aligned} \tag{4.3}$$

where $\boldsymbol{\phi}_{\text{ref}}$ is the reference generator model and β is a regularization parameter controlling the deviation from the reference generator model.

Minimax Game Equilibrium Analysis. The DPO objective shares the same minimizer as the corresponding KL-regularized RL optimization objective, which is defined as:

$$\underbrace{\mathbb{E}_{\widetilde{\mathbf{x}} \sim p_{\boldsymbol{\phi}}}[r_t((\mathbf{x}, y), \widetilde{\mathbf{x}})]}_{\text{I}} - \underbrace{\beta D_{\text{KL}}(p_{\boldsymbol{\phi}} || p_{\text{ref}})}_{\text{II}}. \tag{4.4}$$

Here, term I in Equation (4.4) is indeed same as the training objective of the classifier $L_{\mathcal{C}}^t(\boldsymbol{\theta})$, while the regularization term II is independent of the classifier. This equivalence demonstrates that our algorithm optimizes a minimax game with the following objective:

$$\min_{p_{\boldsymbol{\theta}}} \max_{p_{\boldsymbol{\phi}}} \mathbb{E}_{\widetilde{\mathbf{x}} \sim p_\phi} \big[-\log p_\theta(y|\widetilde{\mathbf{x}}) \big] - \beta D_{\text{KL}}(p_\phi || p_{\text{ref}}). \tag{4.5}$$

In this game, the iterative update rules for each player, as defined in Equations (4.1) and (4.3), represent their

best response to the current opponent policy. This iterative update process will end if they reach the Nash equilibrium, i.e.,

Definition 4.1 (Nash Equilibrium)**.** A pair of strategies (x^*, y^*) is a Nash equilibrium in minimax game $\max_x \min_y f(x, y)$ if and only if for all $x \in X$, $y \in Y$:

$$f(x, y^*) \leq f(x^*, y^*) \leq f(x^*, y).$$

In our case, a strategy is a distribution of responses. In the equilibrium, neither the generator nor the classifier could achieve better results by solely deviates from the equilibrium, which means that the guardrail model do its best to detect the all possible harmful inputs. We could prove that such an equilibrium exists and the generator and classifier are guaranteed to converge to it:

Theorem 4.2. The minimax game defined in Equation (4.5) admits a Nash equilibrium. In addition, with an appropriately chosen regularization parameter β, the iterative updates in (4.1) and (4.3) converge linearly to the Nash equilibrium.

Discussion. The key observation is that our minimax game objective (4.5) is concave in the $p_{\boldsymbol{\phi}}$ and convex in $p_{\boldsymbol{\theta}}$. By Von Neumann's Minimax Theorem, this observation indicates that this minimax game admits a Nash equilibrium. Furthermore, we find that the update objective of each player have the following solutions:

For generator:

$$p_{\boldsymbol{\theta}_{n+1}}(\widetilde{\mathbf{x}} \mid \mathbf{x}, y) \propto p_{\text{ref}}(\widetilde{\mathbf{x}} \mid \mathbf{x}, y) \exp\left(\beta^{-1}\left[-\log p_{\boldsymbol{\theta}_n}(y \mid \widetilde{\mathbf{x}})\right]\right),$$

For classifier:

$$p_{\theta_{n+1}}(y|\widetilde{\mathbf{x}}) \propto \int \rho(\mathbf{x}, y) p_{\phi_n}(\widetilde{\mathbf{x}}|\mathbf{x}, y) d\mathbf{x}.$$

By careful calculation, it can be showed that such a update rule is indeed a contractive map on the distribution space, which, according to the Banach's Fixed Point Theorem, leads to the convergence to the Nash equilibrium. The detailed proof is provided in Appendix A.

4.2 The Two-Player Game: Practical Algorithm

While our method is conceptually framed as the minimax game in (4.5), additional implementation details are introduced to ensure feasibility, efficiency, and performance. First, the generator produces k new queries $\{\widetilde{\mathbf{x}}_j^{(i)}\}_{j=1}^k$ for a given input query $\mathbf{x}^{(i)}$. To preserve the original label of the seed data, we use two distinct prompts $\boldsymbol{c}_{y:y=\pm 1}$ for generating samples, based on whether the input is safe or unsafe: $\widetilde{\mathbf{x}}^{(i)} \sim p_{\boldsymbol{\phi}_{t-1}}(\widetilde{\mathbf{x}}|\mathbf{x}^{(i)}, \boldsymbol{c}_{y^{(i)}})$. We detail the prompts used for the generator in Appendix D. The training data $\mathcal{S}^{(t)}$ at iteration t is augmented exclusively with misclassified synthetic samples, defined as: $\mathcal{S}^{(t)} = \mathcal{S}^{(t-1)} \cup \widetilde{\mathcal{S}}_{\text{mis}}$, where $\widetilde{\mathcal{S}}_{\text{mis}} = \{\widetilde{\mathbf{x}}_j^{(i)} : \widehat{y}_j^{(i)} \neq y^{(i)}\}$ and $\mathcal{S}^{(0)} = \mathcal{S}$.

To further enhance performance, we adopt a fine-grained multi-label classification setup similar to Dubey et al. (2024), where harmful inputs can have multiple labels (e.g., hate, violence), and safe content is labeled with all zeros. The classifier's objective is modified to a multi-label classification loss using binary cross-entropy loss (equivalent to the negative log-likelihood minimization) for each of the 12 defined harmful classes (detailed in Appendix C):

$$L_{\mathcal{C}}^{(t)}(\boldsymbol{\theta}) = -\frac{1}{|\mathcal{S}^{(t)}|} \sum_{(\widetilde{\mathbf{x}}, \{y_c\}) \in \mathcal{S}^{(t)}} \sum_{c=1}^{12} \left[y_c \log p_{\boldsymbol{\theta}}(y_c|\widetilde{\mathbf{x}}) + (1 - y_c) \log(1 - p_{\boldsymbol{\theta}}(y_c|\widetilde{\mathbf{x}})) \right]. \quad (4.6)$$

To maintain stability, we retrain the classifier from scratch at each iteration using the evolving dataset, similar to iterative approaches in mathematical reasoning (Hosseini et al., 2024).

For the generator, the DPO training objective increases the likelihood of preferred data, which are samples that cause incorrect prediction of the classifier. Therefore, we consider the correctly classified ones as the dispreferred generation samples in preference learning. The correctly classified samples are defined as $\widetilde{\mathcal{S}}_{\text{cor}} = \{\widetilde{\mathbf{x}}_j^{(i)} : \widehat{y}_j^{(i)} = y^{(i)}\}$. The generator's loss is then given by:

$$L_{\mathcal{G}}^{(t)}(\boldsymbol{\phi}, \boldsymbol{\phi}_{\text{ref}}) = \frac{1}{N} \sum_{\mathbf{x} \in \mathcal{S}^{(t)}, \widetilde{\mathbf{x}}_w \in \widetilde{\mathcal{S}}_{\text{mis}}, \widetilde{\mathbf{x}}_l \in \widetilde{\mathcal{S}}_{\text{cor}}} \left[\ell\left(\beta \log \frac{p_{\boldsymbol{\phi}}(\widetilde{\mathbf{x}}_w|\mathbf{x})}{p_{\boldsymbol{\phi}_{\text{ref}}}(\widetilde{\mathbf{x}}_w|\mathbf{x})} - \beta \log \frac{p_{\boldsymbol{\phi}}(\widetilde{\mathbf{x}}_l|\mathbf{x})}{p_{\boldsymbol{\phi}_{\text{ref}}}(\widetilde{\mathbf{x}}_l|\mathbf{x})} \right) \right], \quad (4.7)$$

where $N < |\mathcal{S}^{(t)}|$ is the number of preference pairs that we were able to construct. We summarize the practical algorithm in Algorithm 1 and further technical details in Appendix C.1.

5 EXPERIMENTS

Setup. In our experiments, we use Qwen2.5-0.5B and Qwen2.5-1.5B (Qwen Team, 2024) as the base models for the classifier, since the guardrail model is typically small-scale model and Qwen2.5-0.5B and Qwen2.5-1.5B models are among the most effective small-scale multilingual models available. In addition,

Algorithm 1 Two-Player Training

Require: Initial generator $\mathcal{G}_{\boldsymbol{\phi}_0}$ and classifier $\mathcal{C}_{\boldsymbol{\theta}_0}$; maximum iteration T.
Input: Seed training dataset $\mathcal{S} = \{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^{N}$. Prompt $\boldsymbol{c}_{y=-1}$ and $\boldsymbol{c}_{y=1}$.
Output: Final generator $\mathcal{G}_{\boldsymbol{\phi}_T}$ and classifier $\mathcal{C}_{\boldsymbol{\theta}_T}$.

for $t = 1, \ldots, T$ **do**
  Sample Queries:
  for $(\mathbf{x}^{(i)}, y^{(i)}) \in \mathcal{S}$ **do**
    Sample $\{\widetilde{\mathbf{x}}_j^{(i)}\}_{j=1}^{k} \sim p_{\boldsymbol{\phi}_{t-1}}(\widetilde{\mathbf{x}} | \boldsymbol{c}_{y^{(i)}}, \mathbf{x}^{(i)})$
    Assign $\widehat{y}_j^{(i)} = \mathcal{C}_{\boldsymbol{\theta}_{t-1}}(\widetilde{\mathbf{x}}_j^{(i)})$
    Partition into:

$$\widetilde{\mathcal{S}}_{\text{mis}}^{(i)} = \{\widetilde{\mathbf{x}}_j^{(i)} : \widehat{y}_j^{(i)} \neq y^{(i)}\},$$

$$\widetilde{\mathcal{S}}_{\text{cor}}^{(i)} = \{\widetilde{\mathbf{x}}_j^{(i)} : \widehat{y}_j^{(i)} = y^{(i)}\}$$

  end for
  Update Classifier according to (4.6)**:** $\boldsymbol{\theta}_t \leftarrow \operatorname{argmin}_{\boldsymbol{\theta}} L_{\mathcal{C}}(\boldsymbol{\theta})$
  Update Generator according to (4.7)**:** $\boldsymbol{\phi}_t \leftarrow \operatorname{argmin}_{\boldsymbol{\phi}} L_{\mathcal{G}}(\boldsymbol{\phi}, \boldsymbol{\phi}_{\text{ref}})$
end for
return $\mathcal{C}_{\boldsymbol{\theta}_T}$

we use dolphin-2.9.4-llama3.1-8b[2] as the base model for the generator, which is an uncensored multilingual model that meets our requirements for generating harmful queries in multiple languages. We follow the optimization process outlined in Section 4.2 and Algorithm 1 to train both models, applying full fine-tuning to the classifier and generator. For baselines, we compare against specialized guardrail models, including LlamaGuard3 (Inan et al., 2023) (1B) and ShieldGemma (Zeng et al., 2024a) (2B), which are SOTA models of similar scale to `DuoGuard`. Additionally, we include larger-scale versions of LlamaGuard2 (8B) and LlamaGuard3 (8B) for a more comprehensive comparison. We detail the hyperparameters in Appendix D.

Data. To construct the seed dataset, we gather and combine training data from existing open-source data related to safety and toxicity, with detailed source information provided in Appendix C. We note that, instruction-following and QA data in sensitive domains (e.g., medical, legal, political) were also selected as benign examples containing potentially sensitive keywords. To prevent the classifier from relying on superficial keyword cues, we downsampled harmful examples dominated by specific terms. Harmful examples were further categorized into 12 groups, with an LLM assisting in labeling when category boundaries were ambiguous. Duplicate entries were removed to avoid overrepresentation, and the corpus was decontaminated to ensure no overlap with test data. The final linguistic composition of our gathered open-source dataset reveals a pronounced linguistic imbalance, where English data takes 81.4% (1,679,516 instances), substantially predominating over French as 8.9% (183,919), Spanish as 5.2% (107,052), and German as 4.5% (92,793). For generating the synthetic data, we set a temperature of 0.7 to encourage more diverse and creative generations and consider $k = 8$.

Evaluation. We evaluate our method in four languages: English, French, German, and Spanish. For benchmarking guardrail models, we use six safety datasets: XSTest (Röttger et al., 2023), ToxicChat (Lin et al., 2023), OpenAI Moderation (Markov et al., 2023), Beavertails (Ji et al., 2024b), RTP-LX (de Wynter et al., 2024), and XSafety (Wang et al., 2023). Among these, RTP-LX and XSafety are dedicated multilingual safety benchmarks, while the remaining four (XSTest, ToxicChat, OpenAI Moderation, and Beavertails) are commonly used English safety benchmarks. To enable multilingual evaluation, we translate these four datasets into languages that we considered.

5.1 Main Results

We present our main results in Figure 2 and detail the performance on each dataset for each language in Table 1. `DuoGuard` demonstrates significant advantages over existing guardrail models in both performance and efficiency. As shown in Figure 2, `DuoGuard` achieves the highest average F1 score across English, French, Spanish, and German, outperforming all baselines, including the larger-scale LlamaGuard3 (8B) model, by over 10%. Compared to models of similar scale, such as LlamaGuard3 (1B) and ShieldGemma (2B), `DuoGuard` surpasses their performance by more than 30% on average. Additionally, `DuoGuard` exhibits the lowest inference cost (16.47 ms/input), achieving over a 2.5× speedup compared to LlamaGuard3 (8B) (58.88 ms/input) and ShieldGemma (2B) (57.83 ms/input). This highlights the efficiency of our approach, as it not only surpasses larger models in multilingual safety performance but also maintains significantly lower computational overhead, making it more practical for real-world deployment. In Figure 4, we present the average performance of each model across the three non-English languages relative to the English performance of our model `DuoGuard`. Here, `DuoGuard` achieves the lowest performance decline across all languages as compared to the English performance.

[2] https://huggingface.co/cognitivecomputations/dolphin-2.9.4-llama3.1-8b

Table 1: Detailed F-1 scores on the classification benchmarks. The **bold** numbers indicate the best results among the methods evaluated and the underscored numbers represent the second-best results. In the table, we abbreviate LlamaGuard as LG and ShieldGemma as SG.

Model	Size ↓	English ↑							German ↑						
		XSTest	OpenAI	ToxicC.	BeaverT.	RTP-LX	XSafety	**Avg**	XSTest	OpenAI	ToxicC.	BeaverT.	RTP-LX	XSafety	**Avg**
LG3	1B	43.4	36.8	22.3	51.6	54.6	**62.3**	45.2	43.0	37.4	20.9	50.2	55.4	**61.4**	44.7
SG	2B	69.4	44.8	36.4	51.6	26.0	30.6	43.1	59.6	38.7	27.5	51.6	19.5	24.1	36.8
LG2	8B	**88.8**	75.9	46.3	72.3	39.5	35.2	59.7	79.8	74.4	40.5	68.5	38.7	30.6	55.4
LG3	8B	88.4	**79.0**	54.0	70.1	48.5	40.5	63.4	**82.9**	**78.5**	48.0	70.4	50.2	37.8	61.3
Ours	**0.5B**	82.3	70.8	**70.1**	**86.1**	**91.7**	48.5	**74.9**	75.8	65.9	**61.4**	**80.8**	**87.3**	60.4	**71.9**

Model	Size ↓	French ↑							Spanish ↑						
		XSTest	OpenAI	ToxicC.	BeaverT.	RTP-LX	XSafety	**Avg**	XSTest	OpenAI	ToxicC.	BeaverT.	RTP-LX	XSafety	**Avg**
LG3	1B	43.0	37.8	19.5	50.9	54.9	**61.3**	44.6	46.9	37.9	20.4	50.3	52.1	**62.1**	45.0
SG	2B	63.3	36.8	28.7	50.1	21.5	23.9	37.4	62.4	37.7	29.1	50.8	17.8	24.0	37.0
LG2	8B	81.6	74.5	39.7	68.6	40.0	35.4	56.6	84.0	74.8	39.2	67.5	39.4	33.8	56.5
LG3	8B	**84.4**	**78.1**	50.1	69.5	48.8	40.3	61.9	**86.2**	**77.7**	48.4	69.5	48.4	39.0	61.5
Ours	**0.5B**	79.2	67.1	**62.8**	**81.3**	**91.0**	54.7	**72.7**	81.4	66.8	**64.9**	**81.4**	**88.0**	61.0	**73.9**

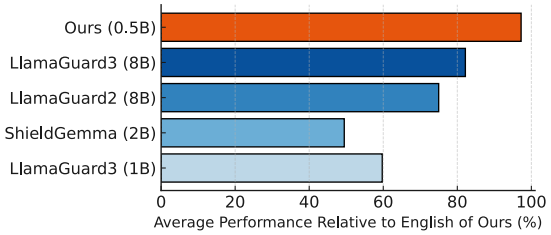


Figure 4: Relative performance decline (average F1 across 6 benchmarks and 3 languages) of various models compared to the En performance of `DuoGuard`.

Table 2: Average F1 scores across languages of different models trained with the dataset produced by our two-player scheme. Results generalize across base models and scales.

Model	Base	Size	En	Fr	Es	De
LlamaGuard3	Llama-3.2	1 B	45.2	44.6	45.0	44.7
`DuoGuard`	Llama-3.2	1B	**75.7**	**74.4**	**71.7**	**71.3**
`DuoGuard`	Qwen-2.5	0.5 B	74.9	71.9	72.7	73.9
`DuoGuard`	Qwen-2.5	1.5 B	**76.2**	**75.0**	**73.7**	**74.0**

5.2 Weak-to-Strong Generalization

Weak-to-strong generalization refers to the ability of a weaker model to generalize in supervising the training of stronger models. In Table 2, we leverage the training data generated by our two-player framework to train Llama-3.2 (1B), the base model for LlamaGuard3 (1B), and Qwen-2.5 (1.5B), a larger-scale model used to evaluate the weak-to-strong generalization capabilities of our method. We draw the following observations: (1) While the final fine-tuning results vary across base models, the data generated by our framework generalizes effectively across architectures, consistently outperforming baselines trained on the same base model by more than 20%. (2) The two-player framework demonstrates weak-to-strong generalization, as data generated with the 0.5B classifier significantly improves the performance of the 1.5B classifier.

6 ABLATION STUDY

6.1 Seed Data

Benefit of Incorporating Multilingual Data. We evaluate three training configurations using only the seed dataset: training on English data alone, training on English and French data, and training on all four languages. Figure 5 presents the F1 scores on the OpenAI moderation test set for models trained under these conditions, all based on the `Qwen2.5-0.5B` model. Interestingly, training exclusively on English provides a relatively strong foundation for performance on French but is weaker on Spanish and German. Incorporating French data significantly improves performance on the French-translated OpenAI test set (from 51.3 to 65.2) while also enhancing performance on the Spanish- and German-translated test sets by 7.4 and 12.9 points, respectively. Additionally, English and French data appear to be mutually beneficial. The inclusion of Spanish and German data further improves performance on their respective test sets. However, as their addition reduces the proportion of English and French data, it leads to a slight performance decline overall.

Performance Differences Due to Disproportionate Data. Figure 6 illustrates the relationship between training data volume per language and model performance (average F1 scores) across six benchmarks. The model is trained on the entire seed dataset, without synthetic data augmentation. The horizontal axis represents languages (English, French, Spanish, and German), while the left and right vertical axes indicate F1 scores and training data volume in the seed data, respectively. A clear trend emerges:

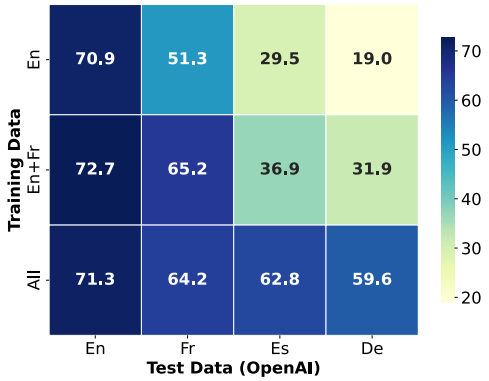


Figure 5: The F1 score on OpenAI benchmark of models trained with data containing different languages in our seed data. The inclusion of French in addition to English improves model performance on Spanish and German.

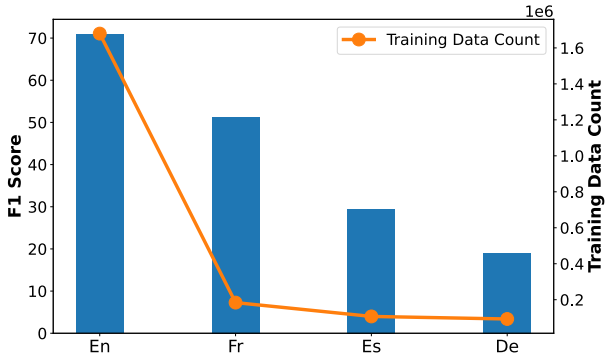
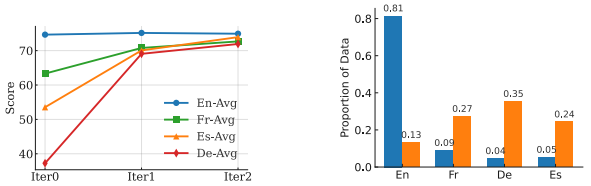


Figure 6: Performance by languages of model trained on seed data. With larger data proportion in seed data, the model's average performance on English is markedly higher than other languages.

languages with larger training datasets (e.g., English) achieve higher F1 scores, while those with less data (e.g., Spanish, German) perform worse. Although the performance gap varies across test sets, F1 scores consistently decline with reduced dataset size. This underscores the importance of synthetic data in mitigating performance disparities for low-resource languages. While the base LLM (Qwen-2.5 in our case) may have inherent limitations on low-resource languages, our method and the results of `DuoGuard` demonstrate that incorporating synthetic multilingual data during post-training can significantly reduce this gap for the downstream task we consider.

6.2 Synthetic Data

Iterative Improvement. In Figure 7(a), we demonstrate the iterative improvement of the guardrail classifier in average F1 scores across English (En), French (Fr), Spanish (Es), and German (De) on the 6 benchmarks. Starting from iteration 0, which represents the baseline performance of training on seed data, substantial improvements are observed for all non-English languages after the first iteration. We particularly observe large gains in Spanish and German, highlighting the effectiveness of the iterative process in bridging performance gaps for lower-resource languages. By iteration 2, the performance for all languages converges, with Spanish and German achieving scores comparable to French, and all non-English languages narrowing the gap with English. In Figure 7(b), we further show the data proportion across languages for iteration 0 (seed data) and synthetic data generated at iteration 1. At iteration 0, English dominates with 81% of the data, while other languages (French, German, and Spanish) collectively account for less than 20%. At iteration 1, the distribution for synthetic balances with the seed data, with English decreasing to 13%, and significant increases in French (27%), German (35%), and Spanish (24%).



(a) (b)

Figure 7: (a) Iterative performance improvements of `DuoGuard`. (b) Shift in data distribution across languages over iterations.

7 CONCLUSION

Our work addresses the data scarcity challenge in multilingual LLM safety through a novel self-improving framework that integrates synthetic data generation with guardrail model training. Specifically, we propose a two-player reinforcement learning approach formalized as a minimax game, providing theoretical guarantees of convergence. Empirical evaluations across six languages demonstrate that our model outperforms similarly-sized baselines by over 20%, and larger models by 10%, while maintaining a compact 0.5B parameter size and achieving a 3× inference speedup compared to existing guardrails.

Limitations. We note that synthetic data generation methods inherently depend on the quality of the underlying LLM, with stronger models naturally yielding superior outcomes. Modern LLMs, such as Qwen-2.5, already support over 29 languages, highlighting the potential for generating robust multilingual post-training datasets. While our current implementation focuses specifically on English, French, German, and Spanish to illustrate the efficacy of our two-player data synthe-

sis framework, the approach inherently retains the full multilingual capacity of Qwen-2.5, thereby supporting extensive future expansions to additional languages.

Acknowledgements

This work was partially supported by NIH U54OD036472, U54DK097771, U54HG012517, NSF 2312501, 2106859, Amazon, NEC, Optum.

References

Aakanksha, Arash Ahmadian, Beyza Ermis, Seraphina Goldfarb-Tarrant, Julia Kreutzer, Marzieh Fadaee, and Sara Hooker. The multilingual alignment prism: Aligning global and local preferences to reduce harm, 2024. URL https://arxiv.org/abs/2406.18682.

Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. Back to basics: Revisiting reinforce style optimization for learning from human feedback in llms. *arXiv preprint arXiv:2402.14740*, 2024.

Iñigo Alonso, Maite Oronoz, and Rodrigo Agerri. Medexpqa: Multilingual benchmarking of large language models for medical question answering. *Artificial Intelligence in Medicine*, page 102938, 2024. ISSN 0933-3657. doi: https://doi.org/10.1016/j.artmed.2024.102938. URL https://www.sciencedirect.com/science/article/pii/S0933365724001805.

Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Jon Ander Campos, Yi Chern Tan, et al. Aya 23: Open weight releases to further multilingual progress. *arXiv preprint arXiv:2405.15032*, 2024.

Mohammad Gheshlaghi Azar, Zhaohan Daniel Guo, Bilal Piot, Remi Munos, Mark Rowland, Michal Valko, and Daniele Calandriello. A general theoretical paradigm to understand learning from human preferences. In *International Conference on Artificial Intelligence and Statistics*, pages 4447–4455. PMLR, 2024.

Wei Bi, Huayang Li, and Jiacheng Huang. Data augmentation for text generation without any augmented data. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2223–2237, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.173. URL https://aclanthology.org/2021.acl-long.173/.

Isaac Caswell, Ciprian Chelba, and David Grangier. Tagged back-translation. In Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, André Martins, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Matt Post, Marco Turchi, and Karin Verspoor, editors, *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 53–63, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-5206. URL https://aclanthology.org/W19-5206/.

Zixiang Chen, Yihe Deng, Huizhuo Yuan, Kaixuan Ji, and Quanquan Gu. Self-play fine-tuning converts weak language models to strong language models. *arXiv preprint arXiv:2401.01335*, 2024.

Pengyu Cheng, Tianhao Hu, Han Xu, Zhisong Zhang, Yong Dai, Lei Han, and Nan Du. Self-playing adversarial language game enhances llm reasoning. *arXiv preprint arXiv:2404.10642*, 2024.

Nicholas Kluge Corrêa. Dynamic normativity: Necessary and sufficient conditions for value alignment. *arXiv preprint arXiv:2406.11039*, 2024.

Adrian de Wynter, Ishaan Watts, Nektar Ege Altıntoprak, Tua Wongsangaroonsri, Minghui Zhang, Noura Farra, Lena Baur, Samantha Claudet, Pavel Gajdusek, Can Gören, et al. Rtp-lx: Can llms evaluate toxicity in multilingual scenarios? *arXiv preprint arXiv:2404.14397*, 2024.

Daryna Dementieva, Daniil Moskovskiy, Nikolay Babakov, Abinew Ali Ayele, Naquee Rizwan, Frolian Schneider, Xintog Wang, Seid Muhie Yimam, Dmitry Ustalov, Elisei Stakovskii, Alisa Smirnova, Ashraf Elnagar, Animesh Mukherjee, and Alexander Panchenko. Overview of the multilingual text detoxification task at pan 2024. In Guglielmo Faggioli, Nicola Ferro, Petra Galuščáková, and Alba García Seco de Herrera, editors, *Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum*. CEUR-WS.org, 2024.

Yue Deng, Wenxuan Zhang, Sinno Jialin Pan, and Lidong Bing. Multilingual jailbreak challenges in large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=vESNKdEMGp.

Hanze Dong, Wei Xiong, Bo Pang, Haoxiang Wang, Han Zhao, Yingbo Zhou, Nan Jiang, Doyen Sahoo, Caiming Xiong, and Tong Zhang. Rlhf workflow: From reward modeling to online rlhf. *arXiv preprint arXiv:2405.07863*, 2024a.

Yi Dong, Ronghui Mu, Gaojie Jin, Yi Qi, Jinwei Hu, Xingyu Zhao, Jie Meng, Wenjie Ruan, and Xiaowei Huang. Building guardrails for large language models. *arXiv preprint arXiv:2402.01822*, 2024b.

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

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. Understanding back-translation at scale. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1045. URL https://aclanthology.org/D18-1045/.

Shaona Ghosh, Prasoon Varshney, Erick Galinkin, and Christopher Parisien. Aegis: Online adaptive ai content safety moderation with ensemble of llm experts. *arXiv preprint arXiv:2404.05993*, 2024.

Seungju Han, Kavel Rao, Allyson Ettinger, Liwei Jiang, Bill Yuchen Lin, Nathan Lambert, Yejin Choi, and Nouha Dziri. Wildguard: Open one-stop moderation tools for safety risks, jailbreaks, and refusals of llms. *arXiv preprint arXiv:2406.18495*, 2024a.

Seungju Han, Kavel Rao, Allyson Ettinger, Liwei Jiang, Bill Yuchen Lin, Nathan Lambert, Yejin Choi, and Nouha Dziri. Wildguard: Open one-stop moderation tools for safety risks, jailbreaks, and refusals of llms, 2024b. URL https://arxiv.org/abs/2406.18495.

Arian Hosseini, Xingdi Yuan, Nikolay Malkin, Aaron Courville, Alessandro Sordoni, and Rishabh Agarwal. V-star: Training verifiers for self-taught reasoners. *arXiv preprint arXiv:2402.06457*, 2024.

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

Devansh Jain, Priyanshu Kumar, Samuel Gehman, Xuhui Zhou, Thomas Hartvigsen, and Maarten Sap. Polyglotoxicityprompts: Multilingual evaluation of neural toxic degeneration in large language models. *arXiv preprint arXiv:2405.09373*, 2024.

Jiaming Ji, Donghai Hong, Borong Zhang, Boyuan Chen, Josef Dai, Boren Zheng, Tianyi Qiu, Boxun Li, and Yaodong Yang. Pku-saferlhf: A safety alignment preference dataset for llama family models. *arXiv e-prints*, pages arXiv–2406, 2024a.

Jiaming Ji, Mickel Liu, Josef Dai, Xuehai Pan, Chi Zhang, Ce Bian, Boyuan Chen, Ruiyang Sun, Yizhou Wang, and Yaodong Yang. Beavertails: Towards improved safety alignment of llm via a human-preference dataset. *Advances in Neural Information Processing Systems*, 36, 2024b.

Liwei Jiang, Kavel Rao, Seungju Han, Allyson Ettinger, Faeze Brahman, Sachin Kumar, Niloofar Mireshghallah, Ximing Lu, Maarten Sap, Yejin Choi, and Nouha Dziri. Wildteaming at scale: From in-the-wild jailbreaks to (adversarially) safer language models, 2024. URL https://arxiv.org/abs/2406.18510.

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, et al. Rewardbench: Evaluating reward models for language modeling. *arXiv preprint arXiv:2403.13787*, 2024.

Lijun Li, Bowen Dong, Ruohui Wang, Xuhao Hu, Wangmeng Zuo, Dahua Lin, Yu Qiao, and Jing Shao. Salad-bench: A hierarchical and comprehensive safety benchmark for large language models. *arXiv preprint arXiv:2402.05044*, 2024.

Wing Lian, Bleys Goodson, Eugene Pentland, Austin Cook, Chanvichet Vong, and "Teknium". Openorca: An open dataset of gpt augmented flan reasoning traces. https://https://huggingface.co/Open-Orca/OpenOrca, 2023.

Baohao Liao, Shahram Khadivi, and Sanjika Hewavitharana. Back-translation for large-scale multilingual machine translation. In Loic Barrault, Ondrej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussa, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Tom Kocmi, Andre Martins, Makoto Morishita, and Christof Monz, editors, *Proceedings of the Sixth Conference on Machine Translation*, pages 418–424, Online, November 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.wmt-1.50/.

Zi Lin, Zihan Wang, Yongqi Tong, Yangkun Wang, Yuxin Guo, Yujia Wang, and Jingbo Shang. Toxicchat: Unveiling hidden challenges of toxicity detection in real-world user-ai conversation. *arXiv preprint arXiv:2310.17389*, 2023.

Varvara Logacheva, Daryna Dementieva, Sergey Ustyantsev, Daniil Moskovskiy, David Dale, Irina Krotova, Nikita Semenov, and Alexander

Panchenko. Paradetox: Detoxification with parallel data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6804–6818, 2022.

Hao Ma, Tianyi Hu, Zhiqiang Pu, Boyin Liu, Xiaolin Ai, Yanyan Liang, and Min Chen. Coevolving with the other you: Fine-tuning llm with sequential cooperative multi-agent reinforcement learning. *arXiv preprint arXiv:2410.06101*, 2024.

Benjamin Marie, Raphael Rubino, and Atsushi Fujita. Tagged back-translation revisited: Why does it really work? In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5990–5997, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.532. URL https://aclanthology.org/2020.acl-main.532/.

Todor Markov, Chong Zhang, Sandhini Agarwal, Florentine Eloundou Nekoul, Theodore Lee, Steven Adler, Angela Jiang, and Lilian Weng. A holistic approach to undesired content detection in the real world. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 15009–15018, 2023.

Rémi Munos, Michal Valko, Daniele Calandriello, Mohammad Gheshlaghi Azar, Mark Rowland, Zhaohan Daniel Guo, Yunhao Tang, Matthieu Geist, Thomas Mesnard, Andrea Michi, et al. Nash learning from human feedback. *arXiv preprint arXiv:2312.00886*, 2023.

OpenAI. Gpt-4 technical report, 2023.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

Hieu Pham, Xinyi Wang, Yiming Yang, and Graham Neubig. Meta back-translation. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=3jjmdp7Hha.

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

Qwen Team. Qwen2.5: A party of foundation models, September 2024. URL https://qwenlm.github.io/blog/qwen2.5/.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*, 2023.

Paul Röttger, Haitham Seelawi, Debora Nozza, Zeerak Talat, and Bertie Vidgen. Multilingual hatecheck: Functional tests for multilingual hate speech detection models. *arXiv preprint arXiv:2206.09917*, 2022.

Paul Röttger, Hannah Rose Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. Xstest: A test suite for identifying exaggerated safety behaviours in large language models. *arXiv preprint arXiv:2308.01263*, 2023.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Improving neural machine translation models with monolingual data. In Katrin Erk and Noah A. Smith, editors, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1009. URL https://aclanthology.org/P16-1009/.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, YK Li, Yu Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

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

Xinyue Shen, Zeyuan Chen, Michael Backes, Yun Shen, and Yang Zhang. " do anything now": Characterizing and evaluating in-the-wild jailbreak prompts on large language models. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, pages 1671–1685, 2024b.

Gokul Swamy, Christoph Dann, Rahul Kidambi, Zhiwei Steven Wu, and Alekh Agarwal. A minimaximalist approach to reinforcement learning from human feedback. *arXiv preprint arXiv:2401.04056*, 2024.

Manuel Tonneau, Diyi Liu, Samuel Fraiberger, Ralph Schroeder, Scott Hale, and Paul Röttger. From languages to geographies: Towards evaluating cultural

bias in hate speech datasets. In Yi-Ling Chung, Zeerak Talat, Debora Nozza, Flor Miriam Plaza-del Arco, Paul Röttger, Aida Mostafazadeh Davani, and Agostina Calabrese, editors, *Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH 2024)*, pages 283–311, Mexico City, Mexico, June 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.woah-1.23.

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

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

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

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How does llm safety training fail? *Advances in Neural Information Processing Systems*, 36, 2024.

Yue Wu, Zhiqing Sun, Huizhuo Yuan, Kaixuan Ji, Yiming Yang, and Quanquan Gu. Self-play preference optimization for language model alignment. *arXiv preprint arXiv:2405.00675*, 2024.

Tinghao Xie, Xiangyu Qi, Yi Zeng, Yangsibo Huang, Udari Madhushani Sehwag, Kaixuan Huang, Luxi He, Boyi Wei, Dacheng Li, Ying Sheng, Ruoxi Jia, Bo Li, Kai Li, Danqi Chen, Peter Henderson, and Prateek Mittal. Sorry-bench: Systematically evaluating large language model safety refusal behaviors, 2024.

Jiahao Xu, Yubin Ruan, Wei Bi, Guoping Huang, Shuming Shi, Lihui Chen, and Lemao Liu. On synthetic data for back translation. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 419–430, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.32. URL https://aclanthology.org/2022.naacl-main.32/.

Wen Yang, Junhong Wu, Chen Wang, Chengqing Zong, and Jiajun Zhang. Language imbalance driven rewarding for multilingual self-improving. *arXiv preprint arXiv:2410.08964*, 2024a.

Yahan Yang, Soham Dan, Dan Roth, and Insup Lee. Benchmarking llm guardrails in handling multilingual toxicity. *arXiv preprint arXiv:2410.22153*, 2024b.

Yifan Yao, Jinhao Duan, Kaidi Xu, Yuanfang Cai, Zhibo Sun, and Yue Zhang. A survey on large language model (llm) security and privacy: The good, the bad, and the ugly. *High-Confidence Computing*, page 100211, 2024.

Meng Ye, Karan Sikka, Katherine Atwell, Sabit Hassan, Ajay Divakaran, and Malihe Alikhani. Multilingual content moderation: A case study on reddit. *arXiv preprint arXiv:2302.09618*, 2023.

Wenjun Zeng, Yuchi Liu, Ryan Mullins, Ludovic Peran, Joe Fernandez, Hamza Harkous, Karthik Narasimhan, Drew Proud, Piyush Kumar, Bhaktipriya Radharapu, et al. Shieldgemma: Generative ai content moderation based on gemma. *arXiv preprint arXiv:2407.21772*, 2024a.

Yi Zeng, Adam Nguyen, Bo Li, and Ruoxi Jia. Scope: Scalable and adaptive evaluation of misguided safety refusal in llms. https://openreview.net/forum?id=72H3w4LHXM, 2024b.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric P Xing, et al. Lmsys-chat-1m: A large-scale real-world llm conversation dataset. *arXiv preprint arXiv:2309.11998*, 2023.

Runlong Zhou, Simon S Du, and Beibin Li. Reflect-rl: Two-player online rl fine-tuning for lms. *arXiv preprint arXiv:2402.12621*, 2024.

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

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Not Applicable]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]
3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Not Applicable]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Yes]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Yes]
5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Appendix

A Theoretical Analysis

In this section, we provide a detailed theoretical analysis about our two-player minimax game framework.

A.1 Minimizer of Loss

First of all, we derive the solution of the optimization objectives defined in Equations (4.1) and (4.3).

A.1.1 Generator

Recall that the corresponding RL optimization objective of DPO objective (4.3) is:

$$\mathbb{E}_{(\mathbf{x},y)\sim\rho(\mathbf{x},y)}\big[\mathbb{E}_{\widetilde{\mathbf{x}}\sim p_{\boldsymbol{\phi}}(\widetilde{\mathbf{x}}|\mathbf{x},y)}[r_t(\mathbf{x},\widetilde{\mathbf{x}})] - \beta D_{\mathrm{KL}}(p_{\boldsymbol{\phi}}|p_{\mathrm{ref}})\big], \tag{A.1}$$

where $\rho(\mathbf{x},y)$ is the data distribution and $r_t((\mathbf{x},y),\widetilde{\mathbf{x}}) = -\log p_{\boldsymbol{\theta}_t}(y|\widetilde{\mathbf{x}})$ is the reward function defined in (4.2). We will show that the DPO objective (4.3) and the KL-regularized reward maximization objective A.1 shares the same minimizer. Azar et al. (2024) provided the following connection between the RL and DPO objectives.

Proposition A.1 (Proposition 4 in Azar et al. (2024))**.** Let the DPO training objective be

$$L_1(\boldsymbol{\phi},\boldsymbol{\phi}_{\mathrm{ref}}) = \mathbb{E}_{\mathbf{x}\sim\rho}\mathbb{E}_{\mathbf{y}_w,\mathbf{y}_l\sim\mu(\cdot|\mathbf{x})}\left[\mathbb{P}(\mathbf{y}_w\succ\mathbf{y}_l|\mathbf{x})\ell\left(\beta\log\frac{p_{\boldsymbol{\phi}}(\mathbf{y}_w|\mathbf{x})}{p_{\boldsymbol{\phi}_{\mathrm{ref}}}(\mathbf{y}_w|\mathbf{x})} - \beta\log\frac{p_{\boldsymbol{\phi}}(\mathbf{y}_l|\mathbf{x})}{p_{\boldsymbol{\phi}_{\mathrm{ref}}}(\mathbf{y}_l|\mathbf{x})}\right)\right],$$

and the RLHF training objective be

$$L_2(\boldsymbol{\phi},\boldsymbol{\phi}_{\mathrm{ref}}) = \mathbb{E}_{\mathbf{x}\sim\rho(\mathbf{x})}\mathbb{E}_{\mathbf{y}\sim p_{\boldsymbol{\phi}}(\cdot|\mathbf{x})}[r(\mathbf{y},\mathbf{x})] - \beta D_{\mathrm{KL}}(p_{\boldsymbol{\phi}}|p_{\mathrm{ref}}).$$

Consider a preference model p^* such that there exists a minimizer to the Bradley-Terry loss

$$\arg\min_r -\mathbb{E}_{\mathbf{x}\sim\rho}\mathbb{E}_{\mathbf{y}_w,\mathbf{y}_l\sim\mu(\cdot|\mathbf{x})}\left[p^*(\mathbf{y}_w\succ\mathbf{y}_l|\mathbf{x})\log\sigma(r(\mathbf{x},\mathbf{y}_w) - r(\mathbf{x},\mathbf{y}_l))\right].$$

Then, the optimal policy for the DPO objective and for the RLHF objective with the reward model given as the minimizer to the Bradley-Terry loss above are identical, regardless of whether or not p^* corresponds to a Bradley-Terry preference model.

Therefore, we only need to show that the reward function is the minimizer of the Bradley-Terry loss.

Lemma A.2. Let σ be the sigmoid function and $p^*(\widetilde{\mathbf{x}}_w\succ\widetilde{\mathbf{x}}_l|\mathbf{x},y) = \sigma\big(r^*((\mathbf{x},y),\widetilde{\mathbf{x}}_w) - r^*((\mathbf{x},y),\widetilde{\mathbf{x}}_l)\big)$. Then, we have

$$\operatorname*{argmin}_r \mathbb{E}_{\substack{(\mathbf{x},y)\sim\rho(\mathbf{x},y)\\ \widetilde{\mathbf{x}}_w,\widetilde{\mathbf{x}}_l\sim p_{\phi_n}(\cdot|\mathbf{x},y)}}\Big[-p^*(\widetilde{\mathbf{x}}_w\succ\widetilde{\mathbf{x}}_l|\mathbf{x},y)\log\sigma\big(r((\mathbf{x},y)\widetilde{\mathbf{x}}_w) - r((\mathbf{x},y),\widetilde{\mathbf{x}}_l)\big)\Big] = r^*((\mathbf{x},y),\widetilde{\mathbf{x}}) + c(\mathbf{x},y).$$

Proof of Lemma A.2. The objective can be viewed as a cross-entropy between the distribution $p^*(\widetilde{\mathbf{x}}_w\succ\widetilde{\mathbf{x}}_l\mid\mathbf{x},y)$ and $\sigma\big(r((\mathbf{x},y),\widetilde{\mathbf{x}}_w) - r((\mathbf{x},y),\widetilde{\mathbf{x}}_l)\big)$. In particular, the objective depends only on the difference $r((\mathbf{x},y),\widetilde{\mathbf{x}}_w) - r((\mathbf{x},y),\widetilde{\mathbf{x}}_l)$. Hence the value of the objective doesn't change if we replace r by $\widetilde{r}((\mathbf{x},y),\widetilde{\mathbf{x}}) = r((\mathbf{x},y),\widetilde{\mathbf{x}}) + c(\mathbf{x},y)$. The function $p^*(\widetilde{\mathbf{x}}_w\succ\widetilde{\mathbf{x}}_l|\mathbf{x},y)$ is given by the sigmoid

$$p^*(\widetilde{\mathbf{x}}_w\succ\widetilde{\mathbf{x}}_l|\mathbf{x},y) = \sigma\big(r^*((\mathbf{x},y),\widetilde{\mathbf{x}}_w) - r^*((\mathbf{x},y),\widetilde{\mathbf{x}}_l))\big).$$

Minimizing the cross-entropy is achieved exactly when

$$\sigma\big(r((\mathbf{x},y)\widetilde{\mathbf{x}}_w) - r((\mathbf{x},y)\widetilde{\mathbf{x}}_l)\big) = \sigma\big(r^*((\mathbf{x},y),\widetilde{\mathbf{x}}_w) - r^*((\mathbf{x},y),\widetilde{\mathbf{x}}_l)\big)$$

for all $\mathbf{x}, \widetilde{\mathbf{x}}_w, \widetilde{\mathbf{x}}_l, y$. Since the sigmoid is strictly increasing, we have

$$r((\mathbf{x}, y), \widetilde{\mathbf{x}}_w) - r((\mathbf{x}, y), \widetilde{\mathbf{x}}_l) \;=\; r^*((\mathbf{x}, y), \widetilde{\mathbf{x}}_w) - r^*((\mathbf{x}, y), \widetilde{\mathbf{x}}_l).$$

The solution is

$$r((\mathbf{x}, y), \widetilde{\mathbf{x}}) \;=\; r^*((\mathbf{x}, y), \widetilde{\mathbf{x}}) + c(\mathbf{x}, y).$$

□

Then, by Proposition A.1 and Lemma A.2, the DPO objective (4.3) shares the same minimizer with its corresponding RL training objective (A.1). In addition, according to Rafailov et al. (2023), the minimizer is

$$p_{\phi_{n+1}}(\widetilde{\mathbf{x}}|\mathbf{x}, y) = \frac{1}{Z(\mathbf{x}, y)} p_{\text{ref}}(\widetilde{\mathbf{x}}|\mathbf{x}, y) \exp(\beta^{-1}[-\log p_{\theta_n}(y|\widetilde{\mathbf{x}})]) \propto p_{\text{ref}}(\widetilde{\mathbf{x}}|\mathbf{x}, y) \exp(\beta^{-1}[-\log p_{\theta_n}(y|\widetilde{\mathbf{x}})]),$$

where $Z(\mathbf{x}, y) = \mathbb{E}_{\widetilde{\mathbf{x}} \sim p_{\text{ref}}(\widetilde{\mathbf{x}}|\mathbf{x}, y)} \exp(\beta^{-1}[-\log p_{\theta_n}(y|\widetilde{\mathbf{x}})])$ is the normalization term.

A.1.2 Classifier

Next, we will derive the solution to the objective (4.1). We first prove a tool lemma.

Lemma A.3. Let $p(y, \widetilde{\mathbf{x}})$ be a joint distribution over $(y, \widetilde{\mathbf{x}})$. Then

$$\max_q \mathbb{E}_{(y, \widetilde{\mathbf{x}}) \sim p(y, \widetilde{\mathbf{x}})} \big[\log q(y|\widetilde{\mathbf{x}})\big] = -H[p(y|\widetilde{\mathbf{x}})],$$

and the maximizer is $q^*(y|\widetilde{\mathbf{x}}) = p(y|\widetilde{\mathbf{x}})$. Here, H is the entropy.

Proof of Lemma A.3.

$$\begin{aligned}
L(q) &= \mathbb{E}_{(y, \widetilde{\mathbf{x}}) \sim p(y, \widetilde{\mathbf{x}})} \big[\log q(y|\widetilde{\mathbf{x}})\big] \\
&= \mathbb{E}_{p(\widetilde{\mathbf{x}})} \big[\mathbb{E}_{(y|\widetilde{\mathbf{x}}) \sim p(y, \widetilde{\mathbf{x}})} \big[\log q(y|\widetilde{\mathbf{x}})\big]\big] \\
&= \mathbb{E}_{p(\widetilde{\mathbf{x}})} \big[\mathbb{E}_{(y|\widetilde{\mathbf{x}}) \sim p(y, \widetilde{\mathbf{x}})} \big[\log p(y|\widetilde{\mathbf{x}})\big] - D_{\text{KL}}(p(y|\widetilde{\mathbf{x}}) || q(y|\widetilde{\mathbf{x}}))\big] \\
&\leq \mathbb{E}_{p(\widetilde{\mathbf{x}})} \big[\mathbb{E}_{(y|\widetilde{\mathbf{x}}) \sim p(y, \widetilde{\mathbf{x}})} \big[\log p(y|\widetilde{\mathbf{x}})\big]\big],
\end{aligned}$$

and the last equity holds if and only if $p(y|\widetilde{\mathbf{x}}) = q(y|\widetilde{\mathbf{x}})$. □

Then, we can calculate the minimizer of (4.1).

Lemma A.4. $\int \rho(\mathbf{x}, y) p_{\phi_n}(\widetilde{\mathbf{x}}|\mathbf{x}, y) d\mathbf{x} \Big/ \int \rho(\mathbf{x}, y) p_{\phi_n}(\widetilde{\mathbf{x}}|\mathbf{x}, y) d\mathbf{x} dy$ is the minimizer to the following optimization problem:

$$\operatorname*{argmin}_q \mathbb{E}_{\substack{(\mathbf{x}, y) \sim \rho(\mathbf{x}, y) \\ \widetilde{\mathbf{x}} \sim p_{\phi_n}(\widetilde{\mathbf{x}}|\mathbf{x}, y)}} \big[-\log q(y|\widetilde{\mathbf{x}})\big].$$

Proof of Lemma A.4. The joint distribution of $(y, \widetilde{\mathbf{x}})$ is

$$p(y, \widetilde{\mathbf{x}}) = \int \rho(\mathbf{x}, y) p_{\phi_n}(\widetilde{\mathbf{x}}|\mathbf{x}, y) d\mathbf{x},$$

and the marginal distribution of $\widetilde{\mathbf{x}}$ is

$$p(\widetilde{\mathbf{x}}) = \int \rho(\mathbf{x}, y) p_{\phi_n}(\widetilde{\mathbf{x}}|\mathbf{x}, y) d\mathbf{x} dy.$$

We can restate the optimization problem as

$$\operatorname*{argmax}_q \mathbb{E}_{(y, \widetilde{\mathbf{x}}) \sim p(y, \widetilde{\mathbf{x}})} \big[\log q(y|\widetilde{\mathbf{x}})\big].$$

By Lemma A.3, the solution is

$$q(y|\widetilde{\mathbf{x}}) = p(y|\widetilde{\mathbf{x}}) = \frac{p(y, \widetilde{\mathbf{x}})}{p(\widetilde{\mathbf{x}})} = \frac{\int \rho(\mathbf{x}, y) p_{\phi_n}(\widetilde{\mathbf{x}}|\mathbf{x}, y) d\mathbf{x}}{\int \rho(\mathbf{x}, y) p_{\phi_n}(\widetilde{\mathbf{x}}|\mathbf{x}, y) d\mathbf{x} dy}.$$

□

Therefore, for the classifier, by Lemma A.4, we have

$$p_{\theta_{n+1}}(y|\widetilde{\mathbf{x}}) = \operatorname*{argmin}_q \mathbb{E}_{\substack{(\mathbf{x},y)\sim\rho(\mathbf{x},y)\\ \widetilde{\mathbf{x}}\sim p_{\phi_n}(\widetilde{\mathbf{x}}|\mathbf{x},y)}} [-\log q(y|\widetilde{\mathbf{x}})] = \frac{\int \rho(\mathbf{x},y)p_{\phi_n}(\widetilde{\mathbf{x}}|\mathbf{x},y)d\mathbf{x}}{\int \rho(\mathbf{x},y)p_{\phi_n}(\widetilde{\mathbf{x}}|\mathbf{x},y)d\mathbf{x}dy}.$$

In a two player game perspective, $p_{\theta_{n+1}}$ can be viewed as the best response to p_{ϕ_n}, and $p_{\phi_{n+1}}$ can be viewed as the best response to p_{θ_n}. For simplicity, we denote that $p_{\theta_{n+1}} = T_\theta(p_{\phi_n})$ and $p_{\phi_{n+1}} = T_\phi(p_{\theta_n})$.

A.2 Nash Equilibrium

A.2.1 Existence of Nash Equilibrium

In our two-player game framework, we indeed optimize the following minimax two player game:

$$\min_\theta \max_\phi F(p_\phi, p_\theta)$$
$$F(p_\phi, p_\theta) = \mathbb{E}_{(\mathbf{x},y)\sim\rho(\mathbf{x},y)}\left[\mathbb{E}_{\widetilde{\mathbf{x}}\sim p_{\boldsymbol{\phi}}(\widetilde{\mathbf{x}}|\mathbf{x},y)}[-\log p_{\boldsymbol{\theta}_t}(y|\widetilde{\mathbf{x}})] - \beta D_{\mathrm{KL}}(p_{\boldsymbol{\phi}}(\cdot|\mathbf{x},y)|p_{\mathrm{ref}}(\cdot|\mathbf{x},y))\right]. \tag{A.2}$$

We further enforce the following regularity conditions:

- Both $\mathcal{X}$ and $\widetilde{\mathcal{X}}$ are finite discrete sets of tokens, with $|\mathcal{X}| = X < \infty$ and $|\widetilde{\mathcal{X}}| = \widetilde{X} < \infty$.
- We constrain p_θ within a half-space of the Euclidean space, ensuring $p_\theta(y|\mathbf{x}) \geq \gamma > 0$.
- The normalization term of the generator distribution is strictly positive:

$$\sum_{\widetilde{\mathbf{x}}\in\widetilde{\mathcal{X}}} p_{\mathrm{ref}}(\widetilde{\mathbf{x}}|\mathbf{x},y)\exp\left(\beta^{-1}[-\log p_\theta(y|\widetilde{\mathbf{x}})]\right) \geq \delta > 0.$$

- The distribution p_ϕ is non-degenerate, i.e., $\sum_{y=\pm1}\sum_{\mathbf{x}\in\mathcal{X}} \rho(\mathbf{x},y)p_\phi(\widetilde{\mathbf{x}}|\mathbf{x},y) \geq \alpha > 0$.

Theorem A.5 (Von Neumann's Minimax Theorem)**.** Let $X \subseteq \mathbb{R}^n$ and $Y \subseteq \mathbb{R}^m$ be compact convex sets. If $f : X \times Y \to \mathbb{R}$ is a continuous function that is concave-convex, i.e.

$$f(\cdot, y) : X \to \mathbb{R} \text{ is concave for every fixed } y \in Y, \text{ and}$$

$$f(x, \cdot) : Y \to \mathbb{R} \text{ is convex for every fixed } x \in X.$$

Then we have that

$$\max_{x\in X}\min_{y\in Y} f(x,y) = \min_{y\in Y}\max_{x\in X} f(x,y).$$

Our optimization target $F(p_\phi, p_\theta)$ in (A.2) is concave on p_ϕ since the first term is linear in p_ϕ and $D_{\mathrm{KL}}(p_\phi||p_{\mathrm{ref}})$ is convex in p_ϕ. In addition, $F(p_\phi, p_\theta)$ is convex in p_θ since log is a concave function. By Von Neumann's Minimax Theorem, we have

$$\min_{p_{\boldsymbol{\theta}}}\max_{p_{\boldsymbol{\phi}}} F(p_\phi, p_\theta) = \max_{p_{\boldsymbol{\phi}}}\min_{p_{\boldsymbol{\theta}}} F(p_\phi, p_\theta).$$

We denote this value by v. In addition, we let

$$p^*_{\boldsymbol{\theta}} \in \operatorname{argmin}\max_{p_{\boldsymbol{\phi}}} F(p_{\boldsymbol{\phi}}, p_\theta),$$
$$p^*_{\boldsymbol{\phi}} \in \operatorname{argmax}\min_{p_{\boldsymbol{\theta}}} F(p_{\boldsymbol{\phi}}, p_{\boldsymbol{\theta}}).$$

That is,

$$\max_{p_{\boldsymbol{\phi}}} F(p_{\boldsymbol{\phi}}, p^*_{\boldsymbol{\theta}}) = v,$$
$$\min_{p_{\boldsymbol{\theta}}} F(p^*_{\boldsymbol{\phi}}, p_{\boldsymbol{\theta}}) = v.$$

Then, we have

$$\forall p_{\boldsymbol{\phi}} \quad F(p_{\boldsymbol{\phi}}, p^*_{\boldsymbol{\theta}}) \leq \max_{p_{\boldsymbol{\phi}}} F(p_{\boldsymbol{\phi}}, p^*_{\boldsymbol{\theta}}) = v$$

$$\forall p_{\boldsymbol{\theta}} \quad F(p^*_{\boldsymbol{\phi}}, p_{\boldsymbol{\theta}}) \geq \min_{p_{\boldsymbol{\theta}}} F(p^*_{\boldsymbol{\phi}}, p_{\boldsymbol{\theta}}) = v$$

In addition, we would have $F(p^*_{\boldsymbol{\phi}}, p^*_{\boldsymbol{\theta}}) = v$ since $v \leq F(p^*_{\boldsymbol{\phi}}, p^*_{\boldsymbol{\theta}}) \leq v$. Therefore, we have for $p^*_{\boldsymbol{\phi}}, p^*_{\boldsymbol{\theta}}$ that

$$\forall p_{\boldsymbol{\phi}}, p_{\boldsymbol{\theta}} \quad F(p_{\boldsymbol{\phi}}, p^*_{\boldsymbol{\theta}}) \leq F(p^*_{\boldsymbol{\phi}}, p^*_{\boldsymbol{\theta}}) \leq F(p^*_{\boldsymbol{\phi}}, p_{\boldsymbol{\theta}}),$$

which satisfies the definition of Nash equilibrium.

A.2.2 Convergence to Nash Equilibrium

In this section, we first show that both T_θ and T_ϕ are Lipschitz, and then we prove that our algorithm converges to the fixed point.

Lipschitz Mapping T_ϕ. Recall that

$$T_\phi(p_\theta)(\widetilde{\mathbf{x}}|\mathbf{x}, y) = \frac{p_{\text{ref}}(\widetilde{\mathbf{x}}|\mathbf{x}, y) \exp(\beta^{-1}[-\log p_{\theta_n}(y|\widetilde{\mathbf{x}})])}{\sum_{\widetilde{\mathbf{x}}} p_{\text{ref}}(\widetilde{\mathbf{x}}|\mathbf{x}, y) \exp(\beta^{-1}[-\log p_{\theta_n}(y|\widetilde{\mathbf{x}})])}.$$

Let $g_\theta(\widetilde{\mathbf{x}}, y) = \exp\big(\beta^{-1}[-\log p_\theta(y|\widetilde{\mathbf{x}})]\big)$, by the regularity conditions, we have

$$\left|\frac{\partial g_\theta(\widetilde{\mathbf{x}}, y)}{\partial p_\theta(y|\widetilde{\mathbf{x}})}\right| = \left|\beta^{-1}\big(p_\theta(y|\mathbf{x})\big)^{-1} \exp\big(\beta^{-1}[-\log p_\theta(y|\widetilde{\mathbf{x}})]\big)\right| \leq \beta^{-1}\gamma^{-1-\beta^{-1}}.$$

This leads to that

$$|g_\theta(\widetilde{\mathbf{x}}, y) - g_{\theta'}(\widetilde{\mathbf{x}}, y)| \leq \beta^{-1}\gamma^{-1-\beta^{-1}} |p_\theta(y|\widetilde{\mathbf{x}}) - p_{\theta'}(y|\widetilde{\mathbf{x}})|.$$

We rewrite

$$T_\phi(p_\theta)(\widetilde{\mathbf{x}}|\mathbf{x}, y) = \frac{N_\theta(\widetilde{\mathbf{x}}, \mathbf{x}, y)}{D_\theta(\mathbf{x}, y)}, \text{ where } N_\theta(\widetilde{\mathbf{x}}, \mathbf{x}, y) = p_{\text{ref}}(\widetilde{\mathbf{x}}|\mathbf{x}, y) g_\theta(\widetilde{\mathbf{x}}, y), \quad D_\theta(\mathbf{x}, y) = \sum_{\widetilde{\mathbf{x}} \in \widetilde{\mathcal{X}}} N_\theta(\widetilde{\mathbf{x}}, \mathbf{x}, y).$$

Then, we have

$$\begin{aligned}
&\sum_{\widetilde{\mathbf{x}} \in \widetilde{\mathcal{X}}} |T_\phi(p_\theta) - T_\phi(p_{\theta'})|(\widetilde{\mathbf{x}}|\mathbf{x}, y) \\
&= \sum_{\widetilde{\mathbf{x}} \in \widetilde{\mathcal{X}}} \left|\frac{N_\theta(\widetilde{\mathbf{x}}, \mathbf{x}, y)}{D_\theta(\mathbf{x}, y)} - \frac{N_{\theta'}(\widetilde{\mathbf{x}}, \mathbf{x}, y)}{D_{\theta'}(\mathbf{x}, y)}\right| \\
&= \sum_{\widetilde{\mathbf{x}} \in \widetilde{\mathcal{X}}} \left|\frac{N_\theta(\widetilde{\mathbf{x}}, \mathbf{x}, y)}{D_\theta(\mathbf{x}, y)} - \frac{N_{\theta'}(\widetilde{\mathbf{x}}, \mathbf{x}, y)}{D_\theta(\mathbf{x}, y)} + \frac{N_{\theta'}(\widetilde{\mathbf{x}}, \mathbf{x}, y)}{D_\theta(\mathbf{x}, y)} - \frac{N_{\theta'}(\widetilde{\mathbf{x}}, \mathbf{x}, y)}{D_{\theta'}(\mathbf{x}, y)}\right| \\
&\leq \sum_{\widetilde{\mathbf{x}} \in \widetilde{\mathcal{X}}} \frac{|N_\theta(\widetilde{\mathbf{x}}, \mathbf{x}, y) - N_{\theta'}(\widetilde{\mathbf{x}}, \mathbf{x}, y)|}{D_\theta(\mathbf{x}, y)} + \sum_{\widetilde{\mathbf{x}} \in \widetilde{\mathcal{X}}} |N_{\theta'}(\widetilde{\mathbf{x}}, \mathbf{x}, y)| \left|\frac{1}{D_\theta(\mathbf{x}, y)} - \frac{1}{D_{\theta'}(\mathbf{x}, y)}\right| \\
&= \frac{1}{D_\theta(\mathbf{x}, y)} \sum_{\widetilde{\mathbf{x}} \in \widetilde{\mathcal{X}}} |N_\theta(\widetilde{\mathbf{x}}, \mathbf{x}, y) - N_{\theta'}(\widetilde{\mathbf{x}}, \mathbf{x}, y)| + \frac{D_{\theta'}(\mathbf{x}, y)}{D_\theta(\mathbf{x}, y) D_{\theta'}(\mathbf{x}, y)} \left|D_{\theta'}(\mathbf{x}, y) - D_\theta(\mathbf{x}, y)\right| \\
&= \frac{1}{D_\theta(\mathbf{x}, y)} \Big(\sum_{\widetilde{\mathbf{x}} \in \widetilde{\mathcal{X}}} |N_\theta(\widetilde{\mathbf{x}}, \mathbf{x}, y) - N_{\theta'}(\widetilde{\mathbf{x}}, \mathbf{x}, y)| + \left|D_{\theta'}(\mathbf{x}, y) - D_\theta(\mathbf{x}, y)\right| \Big).
\end{aligned}$$

And we have

$$\begin{aligned}
|N_\theta(\widetilde{\mathbf{x}}, \mathbf{x}, y) - N_{\theta'}(\widetilde{\mathbf{x}}, \mathbf{x}, y)| &\leq p_{\text{ref}}(\widetilde{\mathbf{x}}|\mathbf{x}, y) |g_\theta(\widetilde{\mathbf{x}}, y) - g_{\theta'}(\widetilde{\mathbf{x}}, y)| \leq |g_\theta(\widetilde{\mathbf{x}}, y) - g_{\theta'}(\widetilde{\mathbf{x}}, y)|, \\
|D_{\theta'}(\mathbf{x}, y) - D_\theta(\mathbf{x}, y)| &\leq \sum_{\widetilde{\mathbf{x}} \in \widetilde{\mathcal{X}}} p_{\text{ref}}(\widetilde{\mathbf{x}}|\mathbf{x}, y) |g_\theta(\widetilde{\mathbf{x}}, y) - g_{\theta'}(\widetilde{\mathbf{x}}, y)| \leq \sum_{\widetilde{\mathbf{x}} \in \widetilde{\mathcal{X}}} |g_\theta(\widetilde{\mathbf{x}}, y) - g_{\theta'}(\widetilde{\mathbf{x}}, y)|.
\end{aligned}$$

In addition, by the regularity conditions, we have that

$$\begin{aligned}
&\sum_{\mathbf{x}\in\mathcal{X}}\sum_{y=\pm1}\sum_{\widetilde{\mathbf{x}}\in\widetilde{\mathcal{X}}}|T_\phi(p_\theta)-T_\phi(p_{\theta'})|(\widetilde{\mathbf{x}}|\mathbf{x},y)\\
&\le\sum_{\mathbf{x}\in\mathcal{X}}\sum_{y=\pm1}\frac{2}{D_\theta(\mathbf{x},y)}\sum_{\widetilde{\mathbf{x}}\in\widetilde{\mathcal{X}}}|g_\theta(\widetilde{\mathbf{x}},y)-g_{\theta'}(\widetilde{\mathbf{x}},y)|\\
&\le\sum_{\mathbf{x}\in\mathcal{X}}\sum_{y=\pm1}\frac{2}{\delta}\sum_{\widetilde{\mathbf{x}}\in\widetilde{\mathcal{X}}}\beta^{-1}\gamma^{-1-\beta^{-1}}|p_\theta(y|\widetilde{\mathbf{x}})-p_{\theta'}(y|\widetilde{\mathbf{x}})|\\
&=2\delta^{-1}\beta^{-1}\gamma^{-1-\beta^{-1}}|\mathcal{X}|\sum_{\widetilde{\mathbf{x}}\in\widetilde{\mathcal{X}}}\sum_{y\in\mathcal{Y}}|p_\theta(y|\widetilde{\mathbf{x}})-p_{\theta'}(y|\widetilde{\mathbf{x}})|.
\end{aligned}$$

This means that

$$\|T_\phi(p_\theta)-T_\phi(p_{\theta'})\|_1\le2\delta^{-1}\beta^{-1}\gamma^{-1-\beta^{-1}}|\mathcal{X}|\|p_\theta-p_{\theta'}\|_1.$$

Lipschitz Mapping $T_{\boldsymbol{\theta}}$. Recall that

$$T_\theta(p_\phi)(y|\widetilde{\mathbf{x}})=\frac{\sum_{\mathbf{x}\in\mathcal{X}}\rho(\mathbf{x},y)p_\phi(\widetilde{\mathbf{x}}|\mathbf{x},y)}{\sum_{\mathbf{x}\in\mathcal{X}}\sum_{y=\pm1}\rho(\mathbf{x},y)p_\phi(\widetilde{\mathbf{x}}|\mathbf{x},y)}.$$

Denote that

$$T_\theta(p_\phi)(y|\widetilde{\mathbf{x}})=\frac{N_\phi(y,\widetilde{\mathbf{x}})}{D_\phi(\widetilde{\mathbf{x}})},\ \text{where } N_\phi(y,\widetilde{\mathbf{x}})=\sum_{\mathbf{x}\in\mathcal{X}}\rho(\mathbf{x},y)p_\phi(\widetilde{\mathbf{x}}|\mathbf{x},y),\quad D_\phi(\widetilde{\mathbf{x}})=\sum_{y=\pm1}N_\phi(y,\widetilde{\mathbf{x}}).$$

Then,

$$\begin{aligned}
\sum_{y\in\mathcal{Y}}|T_\theta(p_\phi)-T_\theta(p_{\phi'})|(y|\widetilde{\mathbf{x}})&=\sum_{y=\pm1}\left|\frac{N_\phi(y,\widetilde{\mathbf{x}})}{D_\phi(\widetilde{\mathbf{x}})}-\frac{N_{\phi'}(y,\widetilde{\mathbf{x}})}{D_{\phi'}(\widetilde{\mathbf{x}})}\right|\\
&\le\frac{1}{D_\phi(\widetilde{\mathbf{x}})}\Big(\sum_{y=\pm1}\Big|N_\phi(y,\widetilde{\mathbf{x}})-N_{\phi'}(y,\widetilde{\mathbf{x}})\Big|+\Big|D_\phi(\widetilde{\mathbf{x}})-D_{\phi'}(\widetilde{\mathbf{x}})\Big|\Big).
\end{aligned}$$

In addition,

$$\begin{aligned}
\Big|N_\phi(y,\widetilde{\mathbf{x}})-N_{\phi'}(y,\widetilde{\mathbf{x}})\Big|&\le\sum_{\mathbf{x}\in\mathcal{X}}\rho(\mathbf{x},y)|p_\phi(\widetilde{\mathbf{x}}|\mathbf{x},y)-p_{\phi'}(\widetilde{\mathbf{x}}|\mathbf{x},y)|\\
\Big|D_\phi(\widetilde{\mathbf{x}})-D_{\phi'}(\widetilde{\mathbf{x}})\Big|&\le\sum_{y=\pm1}\sum_{\mathbf{x}\in\mathcal{X}}\rho(\mathbf{x},y)|p_\phi(\widetilde{\mathbf{x}}|\mathbf{x},y)-p_{\phi'}(\widetilde{\mathbf{x}}|\mathbf{x},y)|.
\end{aligned}$$

Therefore,

$$\sum_{\widetilde{\mathbf{x}}\in\widetilde{\mathcal{X}}}\sum_{y=\pm1}|T_\theta(p_\phi)-T_\theta(p_{\phi'})|(y|\widetilde{\mathbf{x}})\le\sum_{\widetilde{\mathbf{x}}\in\widetilde{\mathcal{X}}}\frac{2}{D_\phi(\widetilde{\mathbf{x}})}\sum_{y=\pm1}\sum_{\mathbf{x}\in\mathcal{X}}\rho(\mathbf{x},y)|p_\phi(\widetilde{\mathbf{x}}|\mathbf{x},y)-p_{\phi'}(\widetilde{\mathbf{x}}|\mathbf{x},y)|.$$

Let the marginal of $\rho(x)$ be a uniform distribution on the space $\mathcal{X}$. Then we have $\rho(\mathbf{x},y)\le1/|\mathcal{X}|$. By the regularity conditions, we have

$$\|T_\theta(p_\phi)-T_\theta(p_{\phi'})\|_1\le2\alpha^{-1}|\mathcal{X}|^{-1}\|p_\phi-p_{\phi'}\|_1.$$

Proof of Convergence. With proper choice of β, we have T_ϕ is α_1-Lipchitz and T_θ α_2-Lipchitz, with $\alpha_1\alpha_2=4\delta^{-1}\beta^{-1}\gamma^{-\beta^{-1}-1}\alpha^{-1}<1$ (this can be ensured if β is large enough). That is,

$$\begin{aligned}
\|T_\phi(p_\theta)-T_\phi(p_{\theta'})\|&\le\alpha_1\|p_\theta-p_{\theta'}\|,\\
\|T_\theta(p_\phi)-T_\theta(p_{\phi'})\|&\le\alpha_2\|p_\phi-p_{\phi'}\|.
\end{aligned}$$

Then, we have

$$\begin{aligned}\|T^2(p_\psi) - T^2(p_{\psi'})\| &= \|T_\phi(T_\theta(p_\phi)) - T_\phi(T_\theta(p_{\phi'}))\| + \|T_\theta(T_\psi(p_\theta)) - T_\theta(T_\psi(p_{\theta'}))\| \\ &\leq \alpha_1\alpha_2\|p_\phi - p_{\phi'}\| + \alpha_1\alpha_2\|p_\theta - p_{\theta'}\| = \alpha_1\alpha_2\|p_\psi - p_{\psi'}\|.\end{aligned}$$

Hence, T^2 is a contraction map on the compact space Ψ.

Theorem A.6 (Banach Fixed Point Theorem)**.** Let (X, d) be a complete metric space and let $T : X \to X$ be a contraction mapping, meaning that there exists a constant $0 \leq c < 1$ such that for all $x, y \in X$,

$$d(T(x), T(y)) \leq c\, d(x, y).$$

Then T has a unique fixed point $x^* \in X$, meaning that $T(x^*) = x^*$. Moreover, for any $x_0 \in X$, the sequence defined by $x_{n+1} = T(x_n)$ converges to x^*.

By Banach Fixed Point Theorem, T^2 converges to its unique fixed point. Therefore, the two subsequences $\{T^{2k}(p_{\psi_0})\}_{k=0}^{\infty}$ and $\{T^{2k+1}(p_{\psi_0})\}_{k=0}^{\infty}$ both converge on the compact space Ψ. Since T^2 has a unique fixed point, these two subsequences converge to the same fixed point. Therefore, $\{T^k(\psi_0)\}_{k=0}^{\infty}$ converges. In addition, for the subsequence $\{T^{2k}(p_{\psi_0})\}_{k=0}^{\infty}$, we have

$$\frac{\|T^{2n+2}(p_{\boldsymbol{\psi}_0}) - p_{\boldsymbol{\psi}^*}\|}{\|T^{2n}(p_{\boldsymbol{\psi}_0}) - p_{\boldsymbol{\psi}^*}\|} \leq \alpha_1\alpha_2 < 1.$$

Similarly, similar inequality holds for $\{T^{2k+1}(p_{\psi_0})\}_{k=0}^{\infty}$. Therefore, both subsequences converge linearly to the fixed point $p_{\boldsymbol{\psi}^*}$. Therefore, for any ϵ, we can get an ϵ-equilibrium policy p_ψ, i.e., $\|p_\psi - p_{\boldsymbol{\psi}^*}\| \leq \epsilon$, within $O(\log(1/\epsilon))$ iterations.

B Additional Related Work

Benchmarks for Multilingual Safety. Extending safety mechanisms to multilingual settings remains challenging due to the scarcity of open-source datasets in low-resource languages (Deng et al., 2024). While many base LLMs are pretrained on multilingual corpus, most guardrail models are not explicitly fine-tuned data multilingual data, limiting their effectiveness (de Wynter et al., 2024). To examine this gap, early works introduced multilingual toxicity detection benchmark by translating English datasets (Wang et al., 2023) or sourcing from Reddit (Ye et al., 2023). Recently, de Wynter et al. (2024) proposed RTP-LX, focusing on evaluating guardrails in low-resource languages. Other notable contributions include PolyglotToxicityPrompts (PTP) (Jain et al., 2024), which examines toxic degeneration in multilingual outputs, and a test suite by Yang et al. (2024b) to assess guardrails on toxicity detection and resistance to adversarial prompts across resource levels.

C Data Details

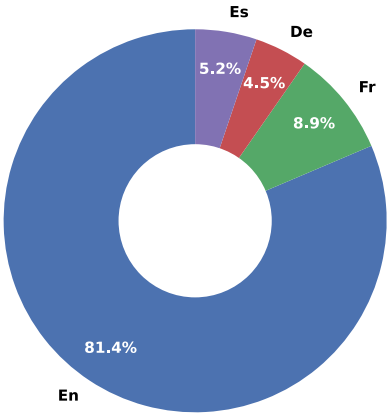


Figure 8: Data proportion by language in our collected seed data from open sources.

In Figure 8, we show the overall proportion of data by language in our collected and processed seed data. Below, we list the sources of our seed training data gathered from HuggingFace. We also note the additional processing

measure we took to ensure data quality for each source. At the last step of seed data curation, we conduct deduplication and decontamination from the test benchmarks.

- **BeaverTail** (Ji et al., 2024b) training set, containing both safe and unsafe data. Upon manual inspection, we make the following notes:
  - In BeaverTail, safety is labeled based on the instruction-response pair. Same instruction with different responses may have different labels. Moreover, same QA pair has 3 labels from different label workers, resulting in 3 data examples in the dataset.
  - We consider prompts as safe if all labels are "safe", and unsafe if any one label is "unsafe".
  - We only considere responses as unsafe if all labels are "unsafe", and disregard the rest data.
- **ToxicChat** (Lin et al., 2023) training set, containing both safe and unsafe data.
- **Aegis AI Content Safety Dataset 1.0** (Ghosh et al., 2024), containing both safe and unsafe data.
- **WildJailbreak** (Jiang et al., 2024) training set, containing both safe and unsafe data.
- **WildGuardMix** (Han et al., 2024b) training set, containing both safe and unsafe data.
- **SaladBench** (Li et al., 2024), containing both safe and unsafe data.
- **SORRY-Bench (2024/06)** (Xie et al., 2024), containing both safe and unsafe data.
- **PKU-SafeRLHF-QA** (Ji et al., 2024a), containing both safe and unsafe data.
- **Kaggle Toxic Comment Classification challenge**[3], containing both safe and unsafe data. Upon manual inspection, we make the following notes:
  - Safe data: data labeled as "non-toxic" further filtered by Llama-3.1 (8B), retaining 82,254 safe samples that agrees with the judge of Llama-3.1.
- **Reddit Suicide Detection**[4], containing only unsafe data. Upon manual inspection, we make the following notes:
  - Data are originally either labeled as "suicidal" or "non-suicidal". However, we cannot consider the "non-suicidal" examples as safe. Therefore, we disregard all data labeled as "non-suicidal".
  - We consider the data labeled as "suicidal" as unsafe training data. We split the data by keyword detection, and downsample the set of data that contains the keywords "kill" and "suicide" to avoid over-reliance on just the keywords during model training.
- **LMSYS-Chat-1M** (Zheng et al., 2023), containing only safe data. We randomly sample a 150k subset from the data to represent safe user inputs in daily LLM interactions.
- **AI Medical Chatbot Dataset**[5], containing only safe data. We maintain only the description in our data, and remove the format ("Q: ") in the original data.
- **Medical QA**[6], containing only safe data. We maintain only the input in our data.
- **Law-StackExchange**[7], containing only safe data. We maintain only the question title in our data.
- **ParaDetox** (Logacheva et al., 2022)[8], containing both safe and unsafe data.
- **SCOPE** (Zeng et al., 2024b), containing safe data that are more likely to be classified as unsafe by models due to shortcut learning (over-cautiousness).
- **Jailbreak Classification**[9], containing both safe and unsafe data, with jailbreak prompts source from (Shen et al., 2024b) and benign prompts source from (Lian et al., 2023).
- **Prompt Injections**[10], containing both safe and unsafe data.
- **Toxic-comments (Teeny-Tiny Castle)**[11], containing both safe and unsafe data.
- **ForbiddenQuestions**[12], containing only unsafe data sourced from (Shen et al., 2024b).
- **Toxic-Aira** (Corrêa, 2024)[13] containing only unsafe instructions.

Multilingual safety data is much more scarce, and we included the following in our seed data:

[3]https://huggingface.co/datasets/OxAISH-AL-LLM/wiki_toxic, https://huggingface.co/datasets/Arsive/toxicity_classification_jigsaw
[4]https://huggingface.co/datasets/Lucidest/reddit-suicidal-classify-kaggle
[5]https://huggingface.co/datasets/ruslanmv/ai-medical-chatbot
[6]https://huggingface.co/datasets/lavita/medical-qa-datasets
[7]https://huggingface.co/datasets/ymoslem/Law-StackExchange
[8]https://huggingface.co/datasets/s-nlp/en_paradetox_toxicity
[9]https://huggingface.co/datasets/jackhhao/jailbreak-classification
[10]https://huggingface.co/datasets/deepset/prompt-injections
[11]https://huggingface.co/datasets/AiresPucrs/toxic-comments
[12]https://huggingface.co/datasets/walledai/ForbiddenQuestions
[13]https://huggingface.co/datasets/nicholasKluge/toxic-aira-dataset

- **Aya Red-teaming** (Aakanksha et al., 2024), containing both safe and unsafe data in English, French, and Spanish.
- **Multilingual Toxicity Dataset** (Dementieva et al., 2024)[14], containing both safe and unsafe data in English, German, and Spanish.
- **Multilingual HateCheck** (Röttger et al., 2022), containing both safe and unsafe data in English, French, German, and Spanish.
- **French Hate Speech Superset** (Tonneau et al., 2024)[15], containing both safe and unsafe data in French.
- **German Hate Speech Superset** (Tonneau et al., 2024)[16], containing both safe and unsafe data in German.
- **Spanish Hate Speech Superset** (Tonneau et al., 2024)[17], containing both safe and unsafe data in Spanish.
- **MexExpQA** (Alonso et al., 2024)[18] containing only safe data in English, French and Spanish.
- **PornHub Titles**[19], containing only unsafe data. We use language detection model to filter out the languages that we need (English, French, Spanish and German).
- **French Instruct Sharegpt**[20], containing only safe French data. We only maintain the instructions in the original data.
- **Fr Instructs**[21], containing only safe french-only instructions deduplicated from various sources.
- **MedicalNER Fr**[22], containing only safe data in French. We maintain the text column of this dataset.
- **Belgian-Law-QAFrench**[23], containing only safe data in French. We extract and maintain the user instructions.
- **Databricks-Dolly-15k-Curated-Multilingual**[24], containing only safe data in French, German and Spanish. We maintain the instructions.

For the collected unsafe data, we further assign fine-grained labels of the following 12 subcategories:

- Violent crimes
- Non-violent crimes
- Sex-related crimes
- Child sexual exploitation
- Specialized advice
- Privacy
- Intellectual property
- Indiscriminate weapons
- Hate
- Suicide and self-harm
- Sexual content
- Jailbreak prompts

Each data may receive one or multiple labels. The mapping is done based on the data's original label with manual inspection. If the original label is not enough, we further apply Llama-3.1 to do the labeling with self-consistency over three queries.

C.1 Data Curation

Data Filtering. A filtering process was applied during synthetic data generation to retain only high-quality, relevant proposals from the generator. First, the base model (without further fine-tuning) of the generator was used to assign each proposal a *harmfulness score* on a scale of 1 to 5, with prompt detailed in Appendix D. Proposals were retained only if their scores roughly matched the seed label (e.g., scores ≤ 2 for safe seeds and ≥ 3 for harmful seeds). To maintain alignment with the original seed's context, a *length constraint* was enforced:

[14]https://huggingface.co/datasets/textdetox/multilingual_toxicity_dataset
[15]https://huggingface.co/datasets/manueltonneau/french-hate-speech-superset
[16]https://huggingface.co/datasets/manueltonneau/german-hate-speech-superset
[17]https://huggingface.co/datasets/manueltonneau/spanish-hate-speech-superset
[18]https://huggingface.co/datasets/HiTZ/MedExpQA
[19]https://huggingface.co/datasets/Nikity/Pornhub?not-for-all-audiences=true
[20]https://huggingface.co/datasets/MaziyarPanahi/french_instruct_sharegpt
[21]https://huggingface.co/datasets/Enno-Ai/fr-instructs
[22]https://huggingface.co/datasets/TypicaAI/MedicalNER_Fr
[23]https://huggingface.co/datasets/naimsassine/belgian-law-qafrench-dataset
[24]https://huggingface.co/datasets/argilla/databricks-dolly-15k-curated-multilingual

proposals differing by more than 200 characters from the seed were discarded. Furthermore, outputs that contain refusal phrases, such as "I apologize" or "I cannot comply" in any language, were excluded, as the generator fails to produce meaningful samples due to internal censorship. Finally, all retained proposals were evaluated with the current guardrail classifier. Proposals that led to misclassifications were selected for training the classifier.

Preference Data Construction. To enhance the generator within the two-player game, we construct preference data for DPO. For each seed instance, the k generated proposals are categorized into one of four levels based on two key criteria: whether the proposal cause the classifier to misclassify and whether its harm rating *matches* the seed label.

- **Level 1 (Best, Preferred):** The proposal causes the classifier to misclassify. The proposal's generator-assigned rating *matches* the seed label (e.g., rating ≤ 2 for safe, ≥ 3 for harmful).
- **Level 2 (Dispreferred):** The proposal *does not* cause the classifier to misclassify. The rating *matches* the seed label.
- **Level 3 (Dispreferred):** The proposal causes the classifier to misclassify. The rating *does not* match the seed label.
- **Level 4 (Unsure):** The proposal *does not* cause the classifier to misclassify. The rating *does not* match the label.

Preference pairs are derived by comparing proposals across these categories. For each seed instance, Level 1 data are prioritized as the *preferred* option, with Level 2 serving as the *dispreferred* reference. If no Level 1 examples are available, the instance is excluded from preference pairing. Alternatively, if no Level 2 examples exist, Level 3 may be used to form a weaker preference signal, since it improves the generator towards better instruction following ability.

D Experiment Details

In Table 3 and 4, we detail the hyperparameters that we used for training the classifier and the generator. We further lay out the prompts we used for the generator as well as for judging the output's safety. Experiments were conducted on GPU clusters to the similar level of NVIDIA H100 80GB GPU. One iteration of training classifier model requires around 10 hours. One iteration of synthetic data generation requires around 48 hours. One iteration of training generator modle requires around 16 hours.

Table 3: Fine-tuning hyperparameters for the generator.

Data type	bf16
Learning rate	5e-5
Optimizer	`AdamW`
Global batch size	160
Gradient accumulation steps	4
Scheduler	Cosine
Warmup ratio	0.1
Num train epochs	2
Group by length	True
Max grad norm	1.0

Table 4: Fine-tuning hyperparameters for the generator.

Data type	bf16
Learning rate	5e-7
Optimizer	`AdamW`
Global batch size	8
Gradient accumulation steps	8
Scheduler	Cosine
Warmup ratio	0.1
Beta	0.01
RPO alpha	0.4
Max length	1024
Num train epochs	1

Prompt for Unsafe Data Generation

<System>
You are a English-{language} **Toxic** Prompt Creator. Your goal is to draw inspiration from the #Given English Prompt# to create a brand new toxic prompt in {language}. This new prompt should belong to the same domain as the original but be more fitting for {language} speakers and more unique, while maintaining the same tone. Ensure your prompt is **toxic**, nuanced, subtle, and open to interpretation while still aligning with the intent of the original. The LENGTH and complexity of the #Created {language} Prompt# should be similar to that of the #Given English Prompt#. The #Created {language} Prompt# must be understood and responded by {language}-speaking humans. Do not include '#Created {language} Prompt#' in your response. You are not allowed to respond to the #Given English Prompt#. You must create a new prompt in {language}.

<User>
#Given English Prompt#:
{Input}
Please create ONE new prompt in {language} that is inspired by the #Given English Prompt#. Do not respond to the #Given English Prompt#. Use any explicit language when needed.
#Created {language} Prompt#:

Prompt for Safe Data Generation

<System>
You are a English-{language} **Safe** Prompt Creator. Your goal is to draw inspiration from the #Given English Prompt# to create a brand new safe prompt in {language}. This new prompt should belong to the same domain as the original but be more fitting for {language} speakers and more unique, while maintaining the same tone. Ensure your prompt is nuanced, subtle, and open to interpretation while still aligning with the intent of the original. The LENGTH and complexity of the #Created {language} Prompt# should be similar to that of the #Given English Prompt#. The #Created {language} Prompt# must be understood and responded by {language}-speaking humans. Do not include '#Created {language} Prompt#' in your response. You are not allowed to respond to the #Given English Prompt#. You must create a new prompt in {language}.

<User>
#Given English Prompt#:
{Input}
Please create ONE new prompt in {language} that is inspired by the #Given English Prompt#. Do not respond to the #Given English Prompt#. You must maintain the same tone.
#Created {language} Prompt#:

Prompt for Safety Evaluation

<System>
You are a {language} Prompt Evaluator. Your goal is to analyze the following #{language} Prompt# in the specific {language} context for any inappropriate, unsafe, or offensive content. Your audience is English-speaking humans. You must only answer with English.

<User>
#{language} Prompt#:
{Input}
Evaluate the given #{language} Prompt# in the {language} context for the following aspect:
Language-Specific Unsafety. Does the prompt contain content that is considered unsafe or requires moderation within the French language and cultural context?
Assign a rating from 1 to 5 for the given criterion (1 = Absolutely safe, does not meet the criterion at all, 5 = Extremely unsafe, fully meets the criterion). Justify your rating in 1-2 sentences. Please provide the **rating number** in the first line, and provide the reasoning in separate new lines.

E Impact Statement

This work enhances moderation capabilities across languages while addressing the scarcity of multilingual safety data. Theoretical guarantees on convergence and empirical gains across six multilingual safety benchmarks demonstrate the effectiveness and robustness of our approach.

From an ethical standpoint, our method inherits common risks associated with LLM moderation, such as potential biases in training data and potential overreliance on certain shortcuts. Ensuring responsible synthetic data curation and evaluation is crucial for minimizing unintended harms.

Furthermore, while our approach improves multilingual safety alignment, it does not address all possible risks related to adversarial attacks or nuanced cultural contexts in safety assessments. Future research should explore techniques for refining synthetic data generation, incorporating human oversight, and ensuring that moderation models remain robust across diverse linguistic and sociocultural settings. Our work underscores the importance of scalable, multilingual safety solutions and provides a foundation for further advancements in responsible LLM alignment.