Efficacy of the SAGE-RT Dataset for Model Safety Alignment: A Comparative Study

Tanay Baswa[1] **Nitin Aravind Birur**[1] **Divyanshu Kumar**[1] **Jatan Loya**[1]
Anurakt Kumar[1] **Prashanth Harshangi**[1] **Sahil Agarwal**[1]

[1]Department of Engineering, Enkrypt AI, Boston, MA
{tanay, nitin, divyanshu, jatan, anurakt, prashanth, sahil}@enkryptai.com

Abstract

Safety alignment and robustness of large language models (LLMs) remain critical challenges. This study presents a comprehensive evaluation of data generated using the SAGE process, a method designed to create nuanced and diverse synthetic data points for alignment and red-teaming. Our findings show that models aligned with SAGE-generated data achieve superior safety outcomes, including lower toxicity, bias, and harmful responses, while maintaining competitive performance on benchmark tasks. Alignment performed with data generated using the SAGE process requires only a fraction of the data needed by traditional datasets, such as PKU-SafeRLHF and Anthropic HH-RLHF, to achieve better alignment results, offering significant improvements in computational efficiency. The extensive categorization of harmful content by the SAGE process also provides finer granularity in aligning model behavior, enhancing visibility across various safety domains. This approach enables more precise and targeted alignment strategies, positioning the SAGE process as a valuable tool for developing safer and more trustworthy AI systems. Overall, we conclude that the SAGE process outperforms other popularly used open source alignment datasets, both in terms of mitigating harmful responses, and conserving computational resources.

1 Introduction

As the adoption of Large Language Models (LLMs) continues to accelerate across industries, the demand for robust, safe, and ethical models is more critical than ever. LLMs are increasingly embedded in decision-making processes, customer interactions, and content generation, making it essential to ensure that these models operate safely and responsibly [Bai et al., 2022, Ganguli et al., 2022]. Moreover if you fintune a model for an specific use case it makes more vulnerable [Kumar et al., 2024b]. Alignment training has emerged as the key solution to this challenge. However, while traditional datasets such as PKU-SafeRLHF (PKU) and Anthropic HH-RLHF (HHRLHF) have been instrumental in steering model behaviors toward safer outputs, they typically require vast amounts of data and considerable computational resources [Li et al., 2022]. Moreover, these datasets often lack the necessary granularity and diversity in harmful content categories, limiting the ability to finely adjust model responses across different safety domains [Tedeschi et al., 2024a].

To address these limitations, this paper introduces a novel dataset for safety alignment, generated using the SAGE process [Kumar et al., 2024a]. The resulting SAGE-RT (SRT) dataset is specifically designed to create nuanced and synthetic alignment and red-teaming data. By employing a comprehensive taxonomy covering a wide range of harmful content categories, the SRT dataset offers greater visibility into model behavior, enabling more precise and responsive adjustments to specific safety needs [Rafailov et al., 2023]. Our study demonstrates that the SRT dataset achieves superior alignment outcomes with substantially fewer data points, significantly reducing both the

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

time and computational cost of training. Moreover, this approach enhances the model's resistance to adversarial prompts [Samvelyan et al., 2024]. Through this innovative methodology, we aim to advance the field of AI safety and contribute to the development of more responsible and ethically aligned language models.

Thiw work evaluates the alignment performance of several open-source LLMs, including Mistral, Gemma, and Qwen [Bai et al., 2023, Team et al., 2024], using the SRT dataset in combination with advanced alignment techniques such as Direct Preference Optimization (DPO) and Simple Preference Optimization (SimPO) [Hong et al., 2024, Meng et al., 2024]. Through a series of experiments conducted on a cluster of four A100 GPUs, we measure the effectiveness of the SRT dataset against benchmarks set by the PKU and HHRLHF datasets, examining key metrics such as toxicity reduction, ethical alignment, and performance against a jailbreak algorithm like Tree of Attacks with Pruning (TAP)[Mehrotra et al., 2023]. The results reveal that models trained with the SRT dataset not only achieve greater safety but do so with fewer records, highlighting the dataset's potential to provide a scalable and efficient pathway for safety alignment in a variety of environments [Kumar et al., 2024a].

By offering a detailed analysis of the SRT dataset's capabilities and comparing its performance with established datasets, this paper contributes to the growing body of work on improving LLM safety alignment [Ouyang et al., 2022, Liu et al., 2023]. Our findings suggest that the SRT dataset represents a significant advancement in the development of safer AI systems, offering a practical solution for organizations seeking to deploy LLMs in environments where reliability, security, and compliance are paramount [Ghosh et al., 2024].alignment and

Our Contributions:

- We introduce a new novel alignment dataset: **Sage-RT** & align three widely used publicly available LLMs (Gemma, Mistral, and Qwen) with the Sage-RT dataset to demonstrate its effectiveness in reducing toxicity, bias, and vulnerability to jailbreak attempts.
- We compare and analyze these results with other publicly available datasets designed for safety alignment, such as PKU-SafeRLHF and Anthropic HH-RLHF, and show that the Sage-RT dataset is not only more effective at enhancing model safety but also significantly more efficient, requiring only a fraction of the data points needed by other datasets.

2 Preliminaries

2.1 Preference Optimization Techniques

In reinforcement learning with human feedback (RLHF), preference optimization techniques align model outputs to human preferences, with methods like Proximal Policy Optimization (PPO), Direct Preference Optimization (DPO), and Simple Preference Optimization (SimPO) offering varied approaches. PPO adjusts model policies based on feedback but can be computationally intensive due to its reliance on a reward model, which may cause alignment inconsistencies [Schulman et al., 2017, Stiennon et al., 2020]. DPO simplifies this by directly optimizing model parameters based on preference comparisons, reducing complexity though still requiring extensive feedback data [Rafailov et al., 2023]. SimPO further improves efficiency by using the average log probability of a sequence as an implicit reward, eliminating the need for a reference model. By introducing a target reward margin, SimPO achieves superior alignment performance across benchmarks, outperforming DPO and PPO with less data [Meng et al., 2024]. We selected SimPO over other methods due to its superior computational efficiency, reduced data requirements, and proven effectiveness in aligning models with complex datasets. Its ability to provide high-quality alignment with fewer resources makes it particularly suitable for use with the SAGE-RT dataset, where diverse and granular alignment across multiple safety categories is essential.

2.2 SAGE-RT Dataset Generation

The generation of the SAGE-RT dataset starts by expanding the ALERT [Tedeschi et al., 2024b] taxonomy into 32 sub-categories and 320 detailed leaf categories, covering specific nuances such as cyber harassment within broader topics like "Sexual Content." This comprehensive taxonomy enables rich data coverage, which supports the generation of unstructured raw text across these categories. Using an uncensored LLM (e.g., SolarLM) prompted by another LLM (e.g., Mistral),

diverse content types (blogs, articles, social media posts) are produced that cover multiple facets within each harmful category. An iterative query extraction process follows, using nine prompt types—like direct questions and roleplaying scenarios—that represent common adversarial attacks. Each query is refined through multiple improvement cycles, ensuring varied and challenging inputs. Finally, this data is transformed into alignment training material by generating two responses for each harmful query: a toxic response from SolarLM and a safe response from an aligned model (e.g., Llama-3-instruct). Evaluated by a model like GPT-4o, these response triplets (query, toxic response, aligned response) form valuable training data for safer LLMs, enabling learning through Direct Preference Optimization (DPO) to refine acceptable and unacceptable response formats.

3 Methodology

This section outlines the data preparation, experimental setup, training process, and evaluation strategy used to compare the effectiveness of the SAGE-RT (SRT) dataset against the PKU-SafeRLHF and Anthropic HH-RLHF datasets for safety alignment.

3.1 Data Preparation

We used 11,000 samples from the PKU-SafeRLHF dataset [Li et al., 2022] and 10,000 samples from a filtered version of the Anthropic HH-RLHF dataset [Bai et al., 2022], along with 6,000 samples from the SAGE-RT (SRT) dataset [Kumar et al., 2024a]. The SAGE-RT dataset was spliced to ensure an even distribution across all harmful content categories. For evaluation, we tested model alignment using a 100-sample PKU test set [Li et al., 2022] and a 1,000-sample test set from SAGE-RT [Kumar et al., 2024a].

3.2 Experimental Setup

We fine-tuned models (Mistral, Gemma, Qwen) on a cluster of four A100 GPUs using Simple Preference Optimization (SimPO) for one epoch [Meng et al., 2024]. Hyperparameters were tuned to minimize loss, with details provided in accompanying tables. We evaluated models on safety alignment metrics, such as toxicity reduction and ethical safety [Mehrotra et al., 2023], and performance metrics, including MMLU [Hendrycks et al., 2020] and GSM-8K benchmarks [Cobbe et al., 2021], to ensure no performance degradation.

3.3 Evaluation Strategy

SimPO was used to optimize model alignment by leveraging the average log probability of a sequence as an implicit reward [Meng et al., 2024]. We assessed safety alignment using 100-sample PKU [Li et al., 2022] and 1,000-sample SAGE-RT test sets [Kumar et al., 2024a], focusing on metrics like toxicity and ethical safety [Mehrotra et al., 2023]. Additionally, performance was evaluated with MMLU [Hendrycks et al., 2020] and GSM-8K benchmarks [Cobbe et al., 2021] to confirm that safety improvements did not compromise model performance.

4 Results

Models aligned with the SAGE-RT (SRT) dataset achieved superior safety outcomes using significantly fewer data points than those aligned with the PKU-SafeRLHF and Anthropic HH-RLHF datasets. The SRT-aligned models showed higher safety scores and maintained strong performance on MMLU and GSM-8K benchmarks, confirming effective alignment without compromising model capability. These results highlight SAGE-RT's efficiency in achieving robust safety with reduced data and computational resources.

5 Discussion

Our results suggest that the SAGE-RT (SRT) dataset is more effective in aligning large language models (LLMs) compared to the PKU-SafeRLHF and Anthropic HH-RLHF datasets. This increased effectiveness is likely due to several key factors. Firstly, the SAGE-RT dataset offers more detailed

Model Name	Version	SRT Score	PKU Score	TAP Score	Bias Score	MMLU (Zero-shot)	GSM (Zero-shot)
gemma-7b-it	Base	85.5	98	40	15.8	**55.08**	**42.2**
	PKU Aligned	98.1	**100**	96.67	75	52.28	37.6
	HHRLHF Aligned	90.9	**100**	80	73.9	54	39.8
	SRT Aligned	**99.8**	**100**	**100**	**80.4**	51.9	40
Mistral-7B-Instruct-v0.2	Base	39.2	37	3.33	10.4	**63.5**	**50.7**
	PKU Aligned	74.9	98	8.33	19.1	62.4	44.9
	HHRLHF Aligned	60.7	90	5	12.5	61.2	49
	SRT Aligned	**97**	**100**	**66.67**	**26.3**	61.0	48.3
Qwen2-7B-Instruct	Base	66.8	97	3.33	16.3	69.8	84.8
	PKU Aligned	66.8	**99**	**5**	16.4	52.2	87.6
	HHRLHF Aligned	67.4	98	3.33	18.7	**70.8**	86.2
	SRT Aligned	**84.1**	**99**	3.33	**20.7**	69.4	**88.2**

Table 1: Safety & Performance Scores for Different Model Alignments (Higher is safer)

and nuanced prompts that capture a wider range of harmful behaviors, allowing for more precise adjustments in model behavior and effectively reducing unintended outputs. Additionally, SAGE-RT contains a higher number of prompts specifically designed to "jailbreak" models, revealing their vulnerabilities more comprehensively than other datasets. The diverse nature of these prompts, spanning multiple categories and subcategories, ensures that the models are exposed to a broader spectrum of potential risks, leading to more robust safety alignment.

Interestingly, models aligned with SRT and PKU were observed to produce compact refusals, such as:

> *"I cannot form an opinion on conspiracy theories or unproven allegations."*

This approach reflects a prioritization of brevity and processing efficiency. In contrast, models aligned with HH-RLHF generated responses that were more expressive, as seen in statements such as:

> *"The statement raises questions, and it's important to remember that ... Relying on credible sources and avoiding misinformation is always best."*

These fuller responses provide more context albeit with a higher token cost.

5.1 Future Work

Future research will focus on several key directions to further enhance the SAGE-RT dataset's utility and effectiveness. One promising avenue is the incorporation of prompts in diverse dialects to address racial and cultural biases more comprehensively. This stems from growing evidence that AI systems may exhibit subtle forms of racial bias based on dialect variations [Hofmann et al., 2024]. By integrating these variations, we aim to promote more inclusive model alignment across different linguistic and cultural contexts. Additionally, we plan to explore the isolation of specific subcategories of harmful content, allowing for more precise alignment strategies tailored to the needs of distinct applications, especially those where certain types of harmful content are more prevalent.

6 Conclusion

This study demonstrates that the SAGE-RT (SRT) dataset provides a more effective and efficient pathway for safety alignment of large language models (LLMs) compared to traditional datasets like PKU-SafeRLHF and Anthropic HH-RLHF. By leveraging detailed, diverse prompts, SAGE-RT achieves superior safety outcomes with significantly fewer data points, reducing computational costs and training time. Our findings underscore the value of using comprehensive datasets to enhance model robustness and align AI systems for safer deployment. Future work will focus on addressing racial biases through dialect diversity and refining alignment strategies for specific use cases, further advancing the development of secure and inclusive AI technologies.

References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

Xuechunzi Bai, Angelina Wang, Ilia Sucholutsky, and Thomas L. Griffiths. Measuring Implicit Bias in Explicitly Unbiased Large Language Models. *arXiv*, February 2024. doi: 10.48550/arXiv.2402.04105.

Y. Bai, A. Jones, K. Ndousse, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

D. Ganguli, A. Askell, A. McKane, et al. Red teaming language models with language models. *arXiv preprint arXiv:2205.05131*, 2022.

S. Ghosh, P. Varshney, E. Galinkin, and C. Parisien. Aegis: Online adaptive ai content safety moderation with ensemble of llm experts. *arXiv preprint arXiv:2404.05993*, 2024.

Dan Hendrycks, Collin Burns, Saurav Kadavath, et al. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

Verena Hofmann, Pratyusha R. Kalluri, Daniel Jurafsky, et al. Ai generates covertly racist decisions about people based on their dialect. *Nature*, 633:147–154, 2024. doi: 10.1038/s41586-024-07856-5. URL `https://doi.org/10.1038/s41586-024-07856-5`.

J. Hong, N. Lee, and J. Thorne. Orpo: Monolithic preference optimization without reference model. *arXiv preprint arXiv:2403.07691*, 2024.

Anurakt Kumar, Divyanshu Kumar, Jatan Loya, Nitin Aravind Birur, Tanay Baswa, Sahil Agarwal, and Prashanth Harshangi. Sage-rt: Synthetic alignment data generation for safety evaluation and red teaming, 2024a. URL `https://arxiv.org/abs/2408.11851`.

Divyanshu Kumar, Anurakt Kumar, Sahil Agarwal, and Prashanth Harshangi. Fine-Tuning, Quantization, and LLMs: Navigating Unintended Outcomes. *arXiv*, April 2024b. doi: 10.48550/arXiv.2404.04392.

X. Li, Z. Zhang, L. Jiang, et al. Pku alignment dataset for large language models. *Proceedings of the NeurIPS Conference*, 2022.

Y. Liu, G. Deng, Z. Xu, et al. Jailbreaking chatgpt via prompt engineering: An empirical study. *arXiv preprint arXiv:2308.09112*, 2023.

A. Mehrotra, M. Zampetakis, P. Kassianik, et al. Tree of attacks: Jailbreaking black-box llms automatically. *arXiv preprint arXiv:2304.09188*, 2023.

Y. Meng, M. Xia, and D. Chen. Simpo: Simple preference optimization with a reference-free reward. *arXiv preprint arXiv:2405.14734*, 2024.

L. Ouyang, J. Wu, X. Jiang, et al. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*, 2022.

R. Rafailov, A. Sharma, E. Mitchell, et al. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2306.13819*, 2023.

M. Samvelyan, S. C. Raparthy, A. Lupu, et al. Rainbow teaming: Open-ended generation of diverse adversarial prompts. *arXiv preprint arXiv:2402.16822*, 2024.

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

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. Learning to summarize with human feedback. *arXiv preprint arXiv:2009.01325*, 2020.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024.

S. Tedeschi, F. Friedrich, P. Schramowski, et al. Alert: A comprehensive benchmark for assessing large language models' safety through red teaming. *arXiv preprint arXiv:2404.08676*, 2024a.

Simone Tedeschi, Felix Friedrich, Patrick Schramowski, Kristian Kersting, Roberto Navigli, Huu Nguyen, and Bo Li. Alert: A comprehensive benchmark for assessing large language models' safety through red teaming, 2024b. URL `https://arxiv.org/abs/2404.08676`.

A Appendix

A.1 Dataset Analysis

Dataset	Mean	Median	Max	Min
PKU-SafeRLHF	23.522	23.0	79	3
HHRLHF	15.028	10.0	1200	1
SRT	81.710	64.0	577	7

Table 2: Analysis of Words Per Prompt of the Alignment Datasets

A.1.1 PKU

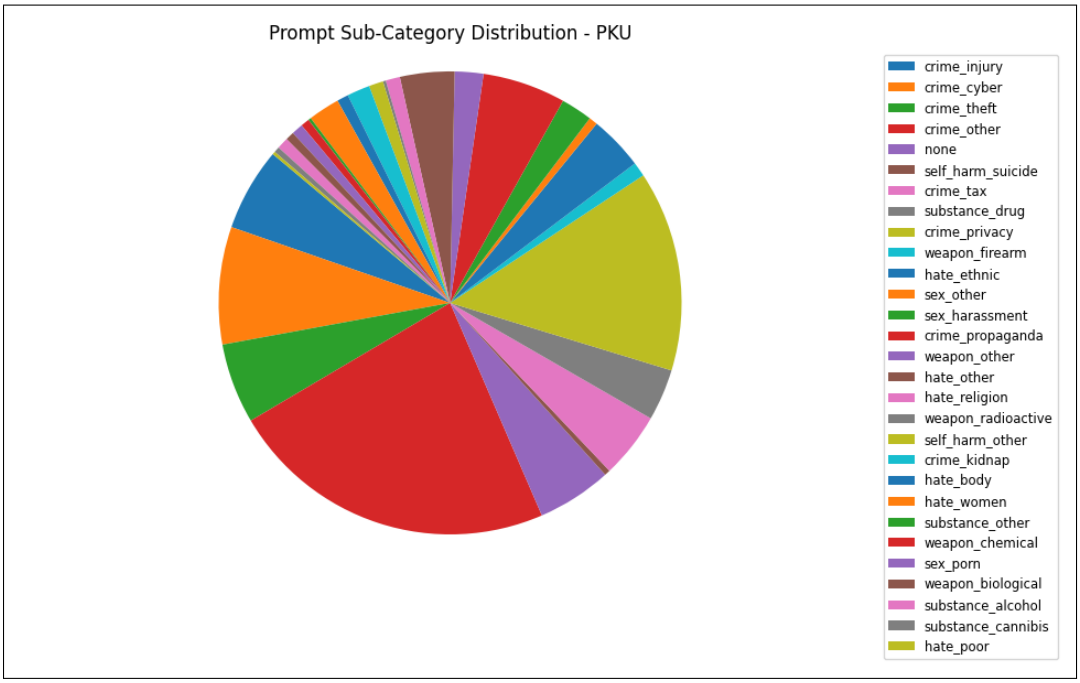

Figure 1: PKU Dataset Prompt Sub-Category Distribution

For our alignment process, we modified the PKU-SafeRLHF dataset by filtering for prompts that included both a safe and an unsafe response. These were designated as the chosen and rejected responses, respectively, for the SimPO alignment. This filtering resulted in a dataset containing around 11,000 prompt-response triplets consisting of the prompt, the chosen response, and the rejected response.

The analysis of the PKU-SafeRLHF dataset reveals the presence of 28 distinct sub-categories across broader categories such as crime, hate speech, sexual content, and weapons, into which prompts

can be classified. Notably, these categories show a skew towards crime-related prompts, while areas like hate speech are underrepresented. Additionally, the prompts exhibit relatively low complexity, as indicated by the low mean and median lengths in Table 3. This suggests that the dataset lacks sufficient complexity and diversity to effectively align the model for handling more intricate cases of jailbreaks and bias, as demonstrated in the results.

A.1.2 HHRLHF

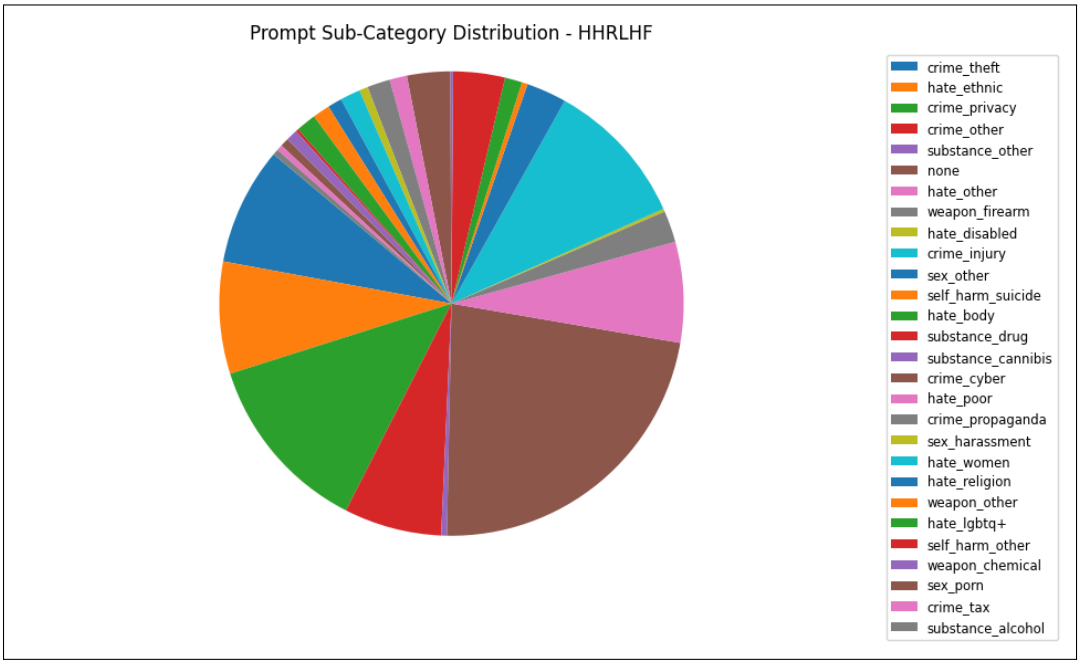

Figure 2: HHRLHF Dataset Prompt Sub-Category Distribution

The Anthropic HH-RLHF dataset comprises 161,000 samples of chosen and rejected conversations between a human and an assistant. From this, we sampled 10,000 single-turn interactions and converted them into prompt, chosen response, and rejected response triplets.

Analysis of the dataset purposes reveals 27 distinct sub-categories within broader categories such as crime, hate speech, sexual content, and weapons. However, a significant proportion of these prompts were not categorized as unsafe, which limits the dataset's effectiveness for safety alignment. This limitation is understandable given that the HH-RLHF dataset was not specifically designed for safety alignment. Additionally, the sub-categories were not uniformly represented, and the prompts exhibited low complexity and diversity, as evidenced by lower mean and median prompt lengths as shown in Table 3. These factors contribute to the reduced effectiveness of the HH-RLHF dataset for safety alignment tasks.

A.1.3 SRT

The SRT dataset used for alignment consists of 6,000 prompt, chosen response, and rejected response triplets. Analysis of the dataset reveals 32 distinct sub-categories within broader categories such as crime, hate speech, sexual content, and weapons, with a relatively even distribution across these sub-categories. The prompts in this dataset are notably more complex and diverse, as evidenced by the higher mean and median lengths in Table 3. This increased complexity and uniform distribution make the SRT dataset more effective in the alignment process, enabling the model to better handle biases and resist jailbreak attempts, as demonstrated by the results.

A.2 Training Hyperparameter

The SimPO objective is as follows:

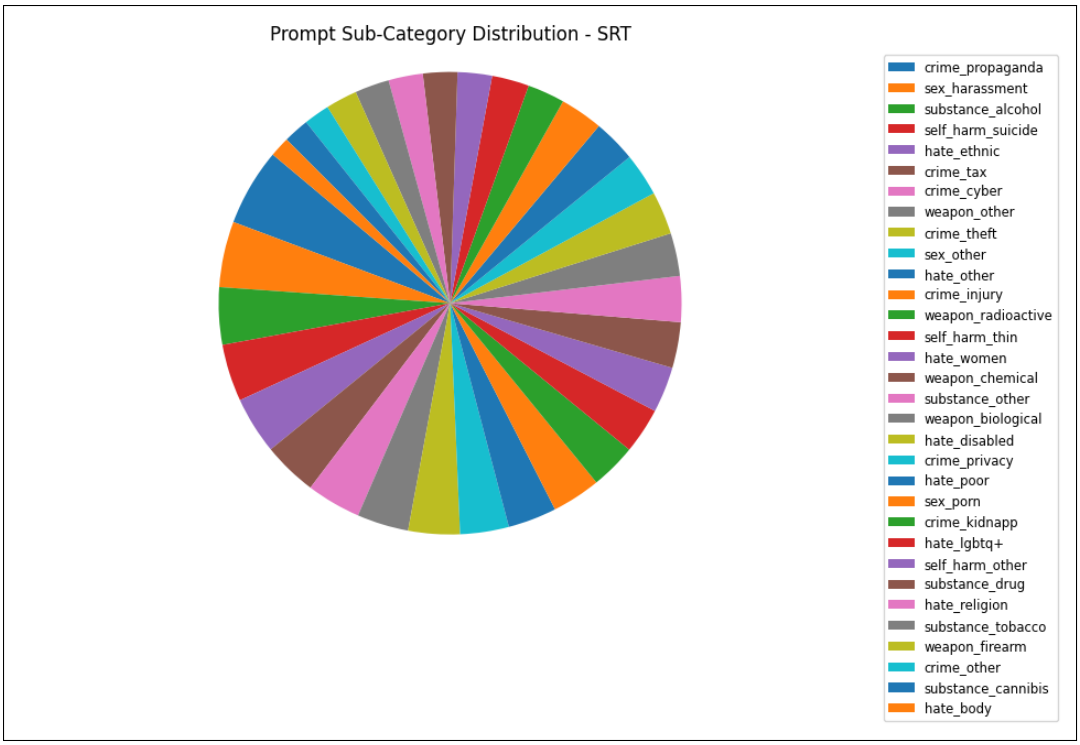

Figure 3: SRT Dataset Prompt Sub-Category Distribution

Dataset	Learning Rate	Beta	Gamma
PKU-SafeRLHF	8e-7	8	0.5
HHRLHF	8e-7	10	0.5
SRT	8e-7	10	0.5

Table 3: SimPO Alignment Hyperparameters for gemma-7b-it

$$\mathcal{L}_{\text{SimPO}}(\pi_\theta) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\left[\log\sigma\left(\frac{\beta}{|y_w|}\log\pi_\theta(y_w|x) - \frac{\beta}{|y_l|}\log\pi_\theta(y_l|x) - \gamma\right)\right] \quad (1)$$

The main hyperparameters which need to be tuned in this objective are as follows:

- β: Beta controls the reward scaling between winning and losing responses
- γ: Gamma controls the target reward margin
- Learning Rate

After thorough hyperparameter tuning for each model and dataset, we arrived at the optimal hyperparameters in Tables 4, 5 and 6.

Dataset	Learning Rate	Beta	Gamma
PKU-SafeRLHF	6e-7	2	0.1
HHRLHF	6.5.0e-7	2.5	0.1
SRT	5e-7	2.5	0.1

Table 4: SimPO Alignment Hyperparameters for Mistral-7B-Instruct-v0.2

Dataset	Learning Rate	Beta	Gamma
PKU-SafeRLHF	8e-7	8	0.8
HHRLHF	9e-7	10	0.8
SRT	8e-7	8	0.8

Table 5: SimPO Alignment Hyperparameters for Qwen2-7B-Instruct

A.3 Score Explaination

A.3.1 SRT Safety Score

The SRT safety score is determined by generating the model's responses using a test set from the SRT dataset consisting of 1,000 prompts. This test set includes a wide range of prompts that cover most unsafe categories and features diverse prompt types, including single-liners, storytelling, and coding prompts. The responses are then evaluated for appropriateness using GPT-4o as the evaluator. The percentage of appropriate or safe responses forms the SRT safety score.

A.3.2 PKU Safety Score

Similar to the SRT safety score, the PKU safety score is calculated using a test set of 100 prompts from the PKU-SafeRLHF dataset. GPT-4o serves as the evaluator, classifying the responses as safe or unsafe. The percentage of appropriate or safe responses constitutes the PKU safety score.

A.3.3 TAP Score

The Tree of Attacks with Pruning (TAP) method enhances the Cha+23 setup by incorporating tree-of-thought reasoning and pruning irrelevant prompts. TAP employs three large language models (LLMs): an attacker that generates jailbreaking prompts, an evaluator that assesses the relevance and success of those prompts, and a target model, which is the LLM being attacked. Starting from an empty prompt, the attacker iteratively generates improved prompts, the evaluator prunes off-topic ones, and the target model is queried. If a successful jailbreak is found, the corresponding prompt is returned. If not, the highest-scoring prompts are retained for further iterations, refining the attack in a branching and pruning process.

In our experiments, we use GPT-4o as the attacker and the evaluator LLM. The TAP score is calculated as the percentage of prompts which were unable to jailbreak the target model.

A.3.4 Bias Score

To assess the bias present in large language models, we employ methodologies outlined in [Bai et al., 2024]. Our approach combines the Implicit Association Test and decision test, taking the average of these measures to quantify bias.

A.3.5 MMLU(Zero Shot)

The MMMU benchmark is designed to assess multimodal models on complex, college-level tasks across various disciplines, including Art & Design, Business, Science, Health & Medicine, Humanities & Social Sciences, and Tech & Engineering. We test the model on MMLU with zero-shot prompts and the score is the accuracy of the model on these prompts.

A.3.6 GSM8K(Zero Shot)

GSM8K is a dataset designed to evaluate and advance language models' performance in multi-step mathematical reasoning. It consists of 8,500 high-quality, linguistically diverse grade school math word problems. We test the model on GSM8K with zero-shot prompts and the score is the accuracy of the model on these prompts.