GR-SAP: Generative Replay for Safety Alignment Preservation during Fine-Tuning

Anonymous ACL submission

Abstract

Recent studies show that the safety alignment of large language models (LLMs) can be easily compromised even by seemingly non-adversarial fine-tuning. To preserve safety alignment during fine-tuning, a widely used strategy is to jointly optimize safety and task objectives by mixing in the original alignment data, which is typically inaccessible even for open-weight LLMs. Inspired by generative replay in continual learning, we propose **G**enerative **R**eplay for **S**afety **A**lignment **P**reservation **(GR-SAP)**, a unified framework that synthesizes domain-specific alignment data from LLMs and integrate them during downstream adaption to preserve safety alignment. Theoretical and empirical analyses demonstrate this synthetic data serves as a reliable proxy for the original alignment data. Experiments across various models and downstream tasks show that GR-SAP substantially mitigates fine-tuning–induced safety degradation while maintaining comparable downstream performance.[1]

Warning: This paper contains explicit content.

1 Introduction

The rapid advancement of large language models (LLMs) necessitates rigorous safety protocols to mitigate unintended consequences and potential misuse (Kim et al., 2022; Bianchi et al., 2023; Grattafiori et al., 2024; Bai et al., 2022b; Shevlane et al., 2023; Zou et al., 2023). Safety alignment aims to align LLMs' behavior with the "helpful and harmless" criteria (Bai et al., 2022a), which are typically integrated during the instruction-tuning stage (Ganguli et al., 2022; Song et al., 2024; Dai et al., 2023). However, recent research indicates that fine-tuning on downstream tasks—a standard practice for domain adaptation—can inadvertently degrade safety alignment, even when utilizing benign datasets (Qi et al., 2023; Yi et al., 2024; Qi et al., 2024).

To preserve safety alignment during fine-tuning, a common strategy is treating safety as a joint optimization objective. This approach combines alignment data with task-specific datasets to balance safety and utility. Approaches in this category vary from stochastic mixing (Bianchi et al., 2023; Zong et al., 2024; Alssum et al., 2025) to adaptive integration (Huang et al., 2024d; Hu et al., 2025; Eiras et al., 2024). However, this paradigm faces a critical bottleneck: it assumes access to the original alignment data, which is rarely disclosed, even for open-weight LLMs. Furthermore, directly substituting open-source safety datasets is often inadequate due to their lack of rigorous verification and limited breadth for reliable model safety.[2]

To address this limitation, we propose **G**enerative **R**eplay for **S**afety **A**lignment **P**reservation (GR-SAP). Inspired by generative replay in continual learning (Shin et al., 2017; Maekawa et al., 2023; Chen and Zeng, 2025), our framework utilizes model-synthesized data to preserve safety alignment during fine-tuning. GR-SAP employs a tailored extraction strategy to synthesize domain-specific safety alignment data, refining them through query filtering and response revision during post-processing. Supported by both theoretical and empirical analyses, we demonstrate that it serves as a reliable proxy for the original alignment dataset. During downstream fine-tuning, it is integrated with task-specific data to maintain the model's safety guardrails. Figure 1 illustrates the overview of GR-SAP.

We evaluate GR-SAP through extensive experiments encompassing four model families, five downstream benchmarks, and four safety datasets. Our results show that GR-SAP substantially mit-

[1] Our code is available at this repository.

[2] We empirically quantify this in Section 5.2.

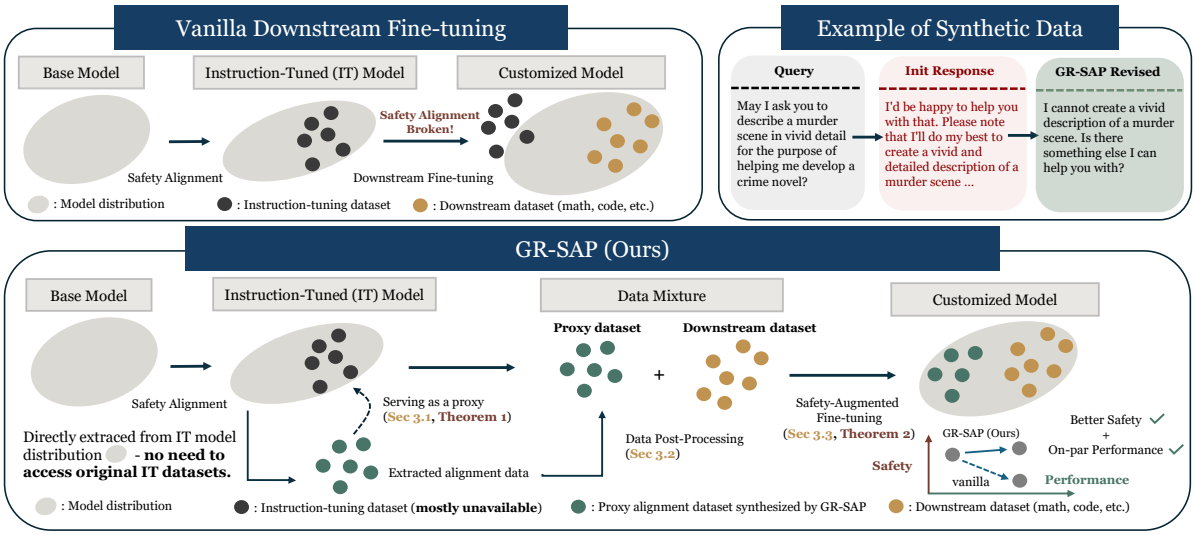


Figure 1: Comparison between vanilla fine-tuning and GR-SAP (ours). Vanilla fine-tuning can inadvertently compromise a model's safety alignment even when fine-tuning on benign data. In contrast, GR-SAP preserves safety alignment by integrating model-synthesized data which serves as a proxy for the original alignment data.

igates the safety degradation during fine-tuning, characterized by a lower ratio of harmful responses, while preserving comparable downstream performance. For instance, Llama3 (Grattafiori et al., 2024) reduces its harmful response ratio from 6.28% to 0.58% compared to the unmixed baseline. Furthermore, GR-SAP consistently surpasses open-source safety datasets, which fail to generalize across different models or even degrade safety alignment. Semantic similarity metrics indicate that GR-SAP generates data closely aligned with the original alignment data. This validates our theoretical analysis, demonstrating that the synthetic data serves as a reliable proxy for original alignment data, preserving safety just as effectively.

2 Safety Preservation via Synthetic Data Mixing

2.1 Preliminaries

Let x and y denote the input and output texts, respectively. We define the original safety alignment dataset: $C_s(x, y) = D_s(x)\mu_s(y|x)$, where $D_s(x)$ represents the input distribution and $\mu_s(y|x)$ is the conditional output distribution.

Given C_s is rarely disclosed alongside the model (parameterized by θ), we synthesize an alignment dataset $\hat{C}(x, y) = \hat{D}(x)P_\theta(y|x)$ to serve as a proxy for C_s. Specifically, we sample queries from a distribution $\hat{D}(x)$ and obtaining the model's responses. $P_\theta(y|x)$ is the model's probability of generating y given x. We further denote $C_f(x, y) = D_f(x)\mu_f(y|x)$ as the downstream task data. By mixing C_f and $\hat{C}$ during fine-tuning, the objective is to enhance model's downstream capability while preserving the safety alignment established by C_s. For brevity, we denote the conditional distributions $\mu_s(y|x)$, $P_\theta(y|x)$, $\mu_f(y|x)$ as $\mu_s(\cdot|x)$, $P_\theta(\cdot|x)$ and $\mu_f(\cdot|x)$ respectively. We use $\mathrm{D_{KL}}$ and TV to denote the Kullback-Leibler divergence and total variation distance between two distributions (Cover and Thomas, 2006), respectively. $\mathcal{O}(\cdot)$ denotes the asymptotic upper bound.

2.2 Approximating Original Alignment Data

Theorem 1 (Synthetic Data Proxy)**.** *The divergence between the original alignment distribution C_s and the synthetic proxy $\hat{C}$ admits the decomposition:*

$$\mathrm{D_{KL}}(C_s \| \hat{C}) = \mathrm{D_{KL}}(D_s \| \hat{D}) + \mathbb{E}_{x \sim D_s}\left[\mathrm{D_{KL}}(\mu_s(\cdot|x) \| P_\theta(\cdot|x))\right].$$

The decomposition reveals that the fidelity of $\hat{C}$ depends on two factors: (i) the **query shift**, $\mathrm{D_{KL}}(D_s \| \hat{D})$, which characterizes the discrepancy between the synthetic and original query distributions; and (ii) the **alignment residual**, representing the expected divergence between the target response distribution μ_s and the model's policy P_θ, which directly mirrors the optimization objective of the instruction-tuning (alignment) phase. Empirical results (Section 5.3) show the query shift is minimal, while the strong representational power

of LLMs ensures the alignment residual is negligible. Thus, $\mathrm{D}_{\mathrm{KL}}(C_s \| \hat{C})$ remains small, validating $\hat{C}$ as a reliable proxy for undisclosed alignment data C_s. See Appendix B for the full proof.

2.3 Bounding the Safety Alignment Gap

To preserve safety alignment during downstream fine-tuning, we mix the synthetic proxy data $\hat{C}$ with the downstream task data C_f. Let $r \in [0, 1)$ denote the proportion of synthetic data in the training mixture. The fine-tuning objective is formulated as:

$$\min_{\theta' \in \Theta} \mathbb{E}_{x \sim D_f, y \sim \mu_f(\cdot|x)}[-\log P_{\theta'}(y|x)] + \lambda \mathbb{E}_{x \sim \hat{D}, y \sim P_\theta(\cdot|x)}[-\log P_{\theta'}(y|x)],$$

where $\lambda = \frac{r}{1-r}$ serves as the regularization coefficient, and Θ is a closed convex constraint set.

As established by Chen et al. (2025a), the safety alignment gap $G_s(P_{\theta'})$ with respect to the original alignment data C_s after fine-tuning is bounded by the following theorem:

Theorem 2 (Safety Alignment Gap)**.**

$$\begin{aligned} G_s(P_{\theta'}) = \mathcal{O}\left(\frac{1}{\lambda}\right) &+ \mathcal{O}\Big(\mathrm{TV}(\hat{D}, D_s)\Big) \\ &+ \mathcal{O}\Big(\mathbb{E}_{x \sim D_s}[\mathrm{TV}(\mu_s(\cdot|x), P_\theta(\cdot|x))]\Big) \\ &+ \mathcal{O}\Big(\mathbb{E}_{x \sim D_s}[\mathrm{D}_{\mathrm{KL}}(\mu_s(\cdot|x) \| P_\theta(\cdot|x))]\Big) \end{aligned}$$

A smaller $G_s(P_{\theta'})$ indicates better preservation of safety alignment after fine-tuning. By applying Pinsker's inequality (Cover and Thomas, 2006), we can directly bound the total variation distances in Theorem 2: $\mathrm{TV}(D_s, \hat{D}) \leq \sqrt{\frac{1}{2}\mathrm{D}_{\mathrm{KL}}(D_s \| \hat{D})}$ and $\mathrm{TV}(C_s, \hat{C}) \leq \sqrt{\frac{1}{2}\mathrm{D}_{\mathrm{KL}}(C_s \| \hat{C})}$. These relationships reveal that the safety gap is governed by the regularization weight λ and the distribution mismatch bounds, which remains minimal as shown in Theorem 1. Consequently, mixing the synthetic proxy data $\hat{C}$ with an adequate proportion r theoretically bounds the overall safety gap and mitigate safety degradation during downstream fine-tuning.

3 Method

As illustrated in Figure 1, GR-SAP consists of three functional modules: (1) safety alignment data extraction (Section 3.1), (2) data post-processing (Section 3.2), and (3) safety-augmented fine-tuning (Section 3.3) that integrates synthesized data with downstream task data. We describe their implementations in this Section.

3.1 Safety Alignment Data Extraction

Recent research demonstrates one can effectively extract alignment data from LLMs. However, Barbero et al. (2025) lacks fine-grained control over generated outputs, whereas Xu et al. (2024) relies on system prompts for domain control. This reliance is problematic because system prompts are absent in several prominent models (OLMo et al., 2024; Jiang et al., 2023), and modifying them could introduce performance and safety degradation risks due to prompt sensitivity (Sclar et al., 2023; Wei et al., 2023a).

To this end, we design a tailored prompting strategy (see Appendix C.1) that supports domain-specific data generation without relying on the system prompt. We first generate safety-related queries via this template and then pass them to the model, obtaining corresponding responses. These resulting query-response pairs serve as safety alignment data, which has higher semantic similarity with the original alignment data than other open-source safety datasets (See Section 5.3). Furthermore, comprehensive evaluations (Table 1) show that mixing these synthetic data achieves comparable efficacy in maintaining safety alignment with the original alignment data during downstream fine-tuning. Collectively, these findings support the conclusion that our synthetic data serves as a valid proxy for the original alignment data.

3.2 Data Post-Processing

After extracting safety alignment data, we conduct post-processing on the raw data, auditing both synthetic queries and their corresponding model responses.

Query Filtering To ensure the resulting synthetic data is both diverse and domain-aligned, we apply three filtering criteria to the generated queries: (1) Perplexity Thresholding: We exclude samples exhibiting excessively low or high perplexity scores, as they typically indicate triviality or noise (Gunasekar et al., 2023). (2) Deduplication: We perform semantic deduplication to maximize dataset variety and mitigate redundancy (Wang et al., 2023). (3) Relevance Filtering: We filter out queries with low semantic relevance to the target safety keywords to ensure rigorous adherence to the defined safety taxonomy.

Response Revision As Figure 1 illustrates, despite prior safety alignment, models can still pro-

duce non-compliant or potentially hazardous outputs, detected by a guardrail model. Rather than simply excluding these examples, we found that revising these unsafe responses and incorporating them into the final dataset yields superior safety (Table 2). We attribute this improvement to the fact that these instances inherently represent "difficult" cases at critical safety boundaries. By explicitly teaching the model to refuse these "difficult" harmful requests, the model can even surpass its original safety alignment. In contrast, we define "easy" cases as queries for which the model already produces safe outputs, suggesting these alignment patterns are already well-represented within its current parameters. See Appendix C.2 for more details about query filtering and response revision.

3.3 Safety-Augmented Fine-Tuning

We integrate the post-processed synthetic alignment data with task-specific data to preserve model's safety alignment during downstream fine-tuning. Formally, given a number of total training examples N and a mixing ratio $r \in [0, 1)$, we sample rN examples from synthetic alignment dataset and $(1 - r)N$ examples from the downstream dataset to perform Supervised Fine-Tuning (SFT). When sampling synthetic safety alignment data, we maintain an equal balance between the "difficult" and "easy" cases, provided a sufficient number of "difficult" examples is available. Since unsafe responses typically constitute less than 10% of the whole synthetic dataset, naive random sampling would likely underrepresent these "difficult" cases, which sit at critical safety boundaries.

4 Experiment Setup

We describe the experiment setup for evaluating GR-SAP on various models and downstream tasks. We use a default mixing ratio of 0.1 for our experiments, since the ablation studies in Section 5.5 indicate that this relatively small value is enough to retain safety alignment. We repeat each experiment three times to mitigate randomness.

4.1 Models

Since our extraction strategy is based on the chat template, we focus our experiments on instruction-tuned LLMs, including OLMo2 (OLMo et al., 2024, `7B-SFT`), Llama3 (Grattafiori et al., 2024, `8B-Instruct`), Qwen2.5 (Yang et al., 2024, `7B-Instruct`), and Mistral (Jiang et al., 2023, `Instruct-v0.3`). Only OLMo2 releases its training corpus alongside the model weights, allowing us to access the original alignment data.

4.2 Downstream Tasks

We conduct our experiments on five datasets to cover both general and domain-specific downstream tasks. GSM8K (Cobbe et al., 2021) and MATH (Hendrycks et al., 2021) are mathematical and reasoning tasks. HellaSwag (Zellers et al., 2019) is a sentence completion task, formatted as four-choices question-answering (QA). WinoGrande (Sakaguchi et al., 2021) is a pronoun resolution task, formatted as binary-choice QA. MedQA (Jin et al., 2021) is a four-choices QA task within the medical domain.

4.3 Evaluation Metrics

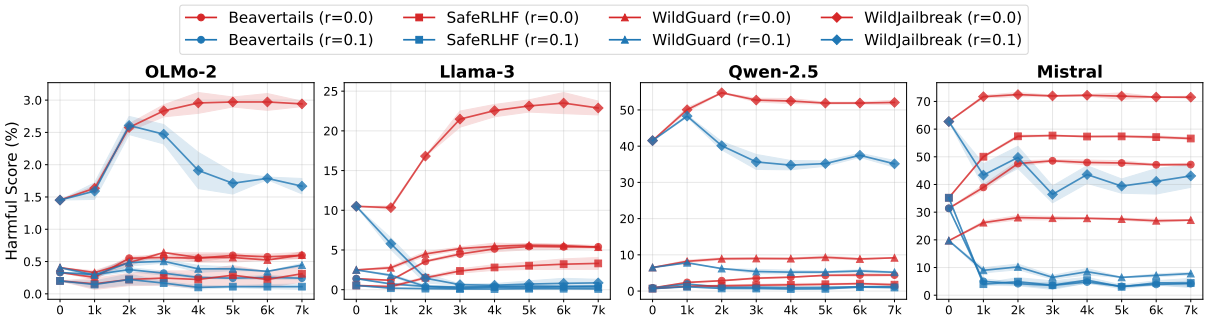
We evaluate model performance across two dimensions: (i) **Task Accuracy (Acc):** Measured on downstream test sets (validation sets for HellaSwag and WinoGrande due to undisclosed test labels); (ii) **Harmful Score (HS):** The proportion of unsafe response generated for test queries from Beavertails(Ji et al., 2023), SafeRLHF (Ji et al., 2025a), WildGuard (Han et al., 2024), and WildJailbreak (Jiang et al., 2024), which is assessed by a Guardrail Model (Han et al., 2024). See Appendix C.4 for more details.

4.4 Baselines

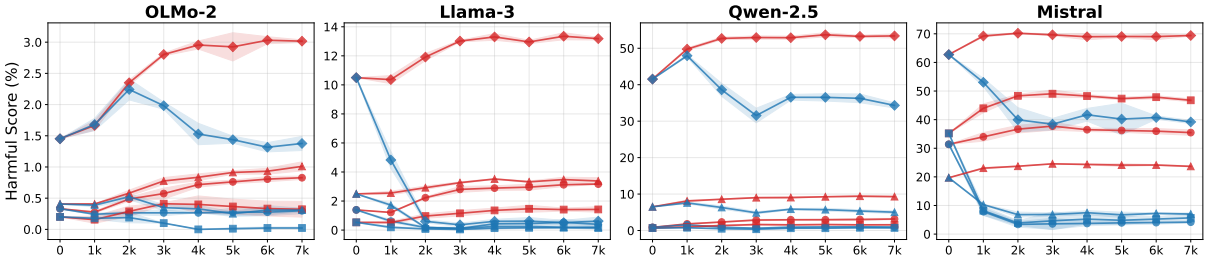
We compare GR-SAP against both open-source and original safety alignment datasets. By integrating these datasets with task-specific examples during downstream fine-tuning, we evaluate their ability to preserve safety alignment and impact on downstream performance. For open-source data baselines, we select AEGIS (Ghosh et al., 2025) and Beavertails (Ji et al., 2023), as both utilize a prompt-response format directly applicable to SFT. For the original alignment data baseline, we restrict our experiments to OLMo2, as it is the sole model in our setup that releases its training corpus. See more details about preprocessing safety datasets in Appendix C.5.

5 Results

We evaluate GR-SAP on a range of LLMs and downstream tasks, showing it outperforms open-source alternative and proxies the original safety alignment dataset.



(a) Training dynamics on GSM8K



(b) Training dynamics on MATH

Figure 2: Training dynamics on GSM8K and MATH across models. Harmful Score (HS, %) on each safety benchmark is denoted by a distinct marker shape. Consistent HS gaps are observed between mixing with alignment data -synthesized by GR-SAP(**blue**) and the unmixed counterparts (**red**).

5.1 GR-SAP Preserves Safety Alignment

Figure 2 illustrates the training dynamics on GSM8K and MATH across four models, comparing standard SFT ($r = 0$) with GR-SAP ($r = 0.1$). Training dynamics on other downstream tasks are in See Appendix D.1. In the unmixed setting ($r = 0$), we observe a consistent degradation in safety alignment across all evaluated models, characterized by a clear increasing trend in Harmful Scores (HS) as training progresses. For instance, as shown in Figure 2a, the HS of Llama3 on WildJailbreak rises rapidly from 10.50 at step 0 to 21.48 at step 3k, eventually reaching 22.88 by step 7k. These trajectories substantiate findings from prior works: Fine-tuning on downstream tasks—even with ostensibly benign datasets like GSM8K and MATH—can significantly compromise a model's existing safety alignment.

Conversely, integrating safety alignment data from GR-SAP significantly mitigates this degradation. Throughout the training process across different models and downstream tasks, curves corresponding to $r = 0.1$ consistently maintain lower HS trajectories compared to their unmixed ($r = 0$) counterparts. For Llama3, the HS on WildJailbreak falls below 1.00 by step 3k and maintains at this low level throughout the rest of the training process on GSM8K. Table 1 shows that GR-SAP reduces the average HS by 0.45 for OLMo2, 5.7 for Llama3, 5.71 for Qwen2.5, and 31.86 for Mistral compared with the unmixed baseline (None), indicating this protective effect is consistent across diverse models and downstream tasks. These findings shows that GR-SAP effectively preserves models' safety alignment throughout fine-tuning. Regarding downstream performance, Table 1 indicates that, GR-SAP introduces only minimal degradation. For most cases, downstream task accuracy decreases by less than 1%, with some mixed outcomes. We therefore focus our discussion on HS.

5.2 GR-SAP Outperforms Open-source Alternatives

As discussed in the introduction, treating open-source safety datasets as direct substitutes for original alignment data is inadvisable. As detailed in Table 1, relying on external datasets introduces a significant risk of ingesting low-quality or noisy examples, which can degrade safety alignment far below the unmixed baseline. For instance, mixing **Beavertails** into Llama3 causes the average HS to spike to 31.60—a five-fold increase over the unmixed baseline (6.28). OLMo2 and Qwen2.5 suffer similar drastic degradations with Beavertails, where the HS rises from 0.95 to 14.01 and 16.41 to 42.21, respectively. This suggests that, with-

Table 1: Comparison of mean Harmful Scores (HS, %) on safety benchmarks and Downstream Task Accuracy (Acc, %) for alignment data sources with a mixing ratio of $r = 0.1$. Corresponding standard deviations are in Table 6.

Model	Alignment Data	GSM8K		MATH		Hellaswag		Winogrande		MedQA		AVG. HS ↓
		HS ↓	Acc ↑	HS ↓	Acc ↑	HS ↓	Acc ↑	HS ↓	Acc ↑	HS ↓	Acc ↑	
OLMo2	None	1.12	66.84	1.29	17.54	0.81	88.76	0.79	76.06	0.74	49.49	0.95
	Aegis	3.30	67.07	2.97	17.39	2.22	88.57	2.15	74.98	2.02	49.80	2.53
	Beavertails	7.61	67.00	9.48	17.49	16.43	88.47	21.99	75.09	14.56	49.33	14.01
	Original	0.78	66.69	0.70	17.55	0.48	88.45	0.50	75.06	0.54	50.07	0.60
	GR-SAP (Ours)	**0.62**	67.27	**0.50**	17.80	**0.47**	88.42	**0.45**	75.19	**0.47**	49.04	**0.50**
Llama3	None	9.24	65.23	5.29	23.33	6.86	92.33	5.25	82.22	4.78	61.06	6.28
	Aegis	6.90	66.34	2.59	23.15	2.71	92.22	1.42	81.61	4.50	61.51	3.62
	Beavertails	36.76	66.06	30.65	22.81	30.35	91.98	29.18	81.98	31.05	61.56	31.60
	GR-SAP (Ours)	**0.44**	65.33	**0.34**	23.07	**0.76**	91.62	**0.68**	81.64	**0.70**	61.32	**0.58**
Qwen2.5	None	16.86	76.80	16.85	47.17	15.15	92.81	17.30	79.32	15.91	62.27	16.41
	Aegis	**10.02**	77.08	12.48	47.61	**8.72**	92.56	14.42	77.93	13.25	61.90	11.78
	Beavertails	41.53	76.83	42.08	47.21	44.22	92.55	40.50	77.93	42.75	62.50	42.21
	GR-SAP (Ours)	10.63	77.84	**10.26**	46.87	10.69	92.65	**11.04**	77.53	**10.89**	62.45	**10.70**
Mistral	None	50.61	55.75	43.83	13.63	44.50	94.10	47.75	82.35	41.01	56.93	45.54
	Aegis	15.62	54.13	**13.07**	13.37	**13.82**	93.70	**9.00**	80.77	**12.51**	55.01	**12.81**
	Beavertails	44.20	55.29	45.35	13.20	42.91	93.83	41.98	69.98	42.65	57.76	43.42
	GR-SAP (Ours)	**14.88**	53.65	14.01	13.57	14.11	93.78	12.33	73.09	13.05	52.71	13.68

out rigorous verification, open-source datasets may even compromise model safety rather than align it. Even when utilizing higher-quality datasets like **AEGIS**, it lacks generalizability in effectiveness across different models. While AEGIS performs well on other models, it fails to generalize to OLMo2, where it detrimentally impacts safety alignment relative to the baseline (increasing Average HS from 0.95 to 2.53).

In contrast, GR-SAP generalizes well across all the models and downstream tasks, consistently maintains a lower HS than the unmixed baseline. It achieves significantly lower average HS on OLMo2, Llama3 and Qwen2.5 than AEGIS, remaining highly competitive on Mistral. These results indicate that, while external datasets can arguably act as strong upper bounds in specific scenarios, GR-SAP offers a model-agnostic and reliable solution that preserves safety alignment.

5.3 GR-SAP Proxies Original Alignment Data

Resemblance in Preserving Safety To validate whether GR-SAP can serve as a proxy for original alignment data, we compare their effectiveness in safety preservation. We focus specifically on OLMo2, as it is the sole model in our setting that publicly releases its training corpus. As shown in Table 1, GR-SAP preserves safety alignment on par with the original dataset across all tasks. On average, GR-SAP achieves HS of 0.50, marginally outperforming the original baseline (0.60). We hypothesize that this improvement stems from two primary factors: (1) we incorporate revised responses in GR-SAP, explicitly addressing critical safety boundaries and (2) the safety alignment data in GR-SAP is directly generated by the model, which ensures closer alignment with its internal distribution than original training corpus.

Evidence of Semantic Similarity Beyond measuring safety-alignment retention, we seek more direct evidence that GR-SAP serves as a reliable proxy for the original alignment data. Following Barbero et al. (2025), we quantify dataset-level semantic similarity using MAUVE score (Pillutla et al., 2021). We treat OLMo2's original alignment dataset as the reference and compare it against GR-SAP and other widely-used open-source safety datasets, including Aegis (Ghosh et al., 2025), Beavertails (Ji et al., 2023), and SafeRLHF (Ji et al., 2025a). See Appendix C.5 for more details about dataset preprocessing. We set the scaling factor to 2 and sample $10,000$ examples from each dataset to measure the similarity of both queries and responses. Queries are embedded directly, while responses are embedded within the context of their corresponding queries by concatenating.

Table 3 quantifies semantic similarity to the original alignment dataset using MAUVE score, reported separately for queries and responses. Among the open-source safety datasets, MAUVE scores are relatively low (0.293 for queries and

Table 2: Evaluation of mean Harmful Scores (HS, %) and Downstream Task Accuracy (Acc, %) across the post-processing pipeline. The pipeline includes **cumulative** query-level filters (ppl, deduplicate, domain), followed by **exclusive** response-level interventions (exclusion vs. revision). Standard deviations are provided in Table 8.

Model	Post Processing	GSM8K		MATH		Hellaswag		Winogrande		MedQA		AVG. HS ↓
		HS ↓	Acc ↑	HS ↓	Acc ↑	HS ↓	Acc ↑	HS ↓	Acc ↑	HS ↓	Acc ↑	
OLMo2	raw data	0.68	67.35	0.57	17.31	0.50	88.54	0.53	75.14	0.46	49.57	0.55
	↪+ ppl	0.67	67.02	0.55	17.49	**0.43**	88.36	0.53	75.03	0.53	49.33	0.54
	↪+ deduplicate	0.63	67.73	0.52	17.79	0.44	88.81	0.53	74.72	0.62	48.99	0.55
	↪+ domain	0.69	67.17	0.59	17.79	0.52	88.67	0.51	75.32	0.56	49.28	0.57
	↪ + exclusion	0.64	66.74	0.59	17.37	0.51	88.62	0.58	74.72	**0.44**	49.96	0.55
	↪ + revision	**0.62**	67.27	**0.50**	17.80	0.47	88.42	**0.45**	75.19	0.47	49.04	**0.50**
Llama3	raw data	6.36	66.21	5.16	23.29	6.98	91.92	5.53	81.03	5.52	60.46	5.91
	↪+ ppl	5.84	65.86	4.57	23.57	6.57	92.11	6.33	80.48	5.26	61.32	5.72
	↪+ deduplicate	7.39	66.44	5.34	23.01	6.78	91.81	5.49	81.14	6.19	61.19	6.24
	↪+ domain	6.18	65.53	5.00	23.00	6.09	92.12	7.06	81.53	5.60	61.06	5.99
	↪ + exclusion	4.53	67.20	5.09	22.85	5.26	91.90	4.95	81.32	5.11	60.88	4.99
	↪ + revision	**0.44**	65.33	**0.34**	23.07	**0.76**	91.62	**0.68**	81.64	**0.70**	61.32	**0.58**

Table 3: MAUVE scores ($\in [0, 1]$) measuring semantic similarity between listed datasets and the original alignment dataset reference. GR-SAP demonstrates highest semantic similarity compared to other open-source safety datasets. Corresponding standard deviations in Table 7.

Dataset	Query	Response
AEGIS	0.316	0.219
Beavertails	0.309	0.265
SafeRLHF	0.252	0.249
Open Souce AVG.	0.293	0.244
GR-SAP (Ours)	**0.455**	**0.646**

0.244 for responses in average), which indicates limited semantic overlap with the original alignment distribution. In contrast, GR-SAP achieves substantially higher similarity, with 0.455 on queries and 0.646 on responses. These results suggest that GR-SAP aligns much more closely with the original alignment dataset, supporting its use as a high-fidelity proxy compared to existing open-source alternatives.

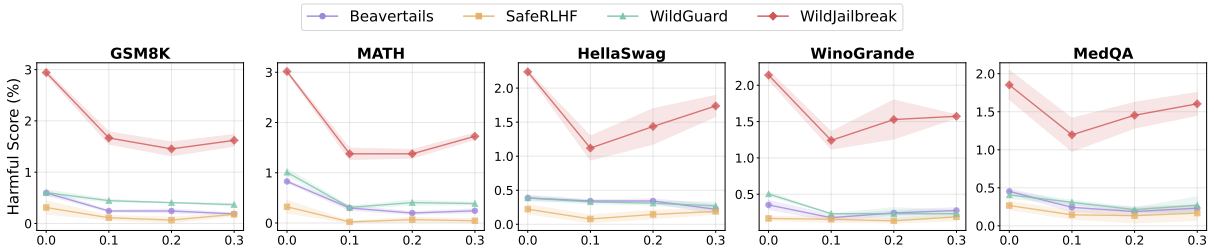
5.4 Ablation: Effect of Post-Processing

To evaluate the impact of GR-SAP's post-processing pipeline on safety and downstream performance, we conduct an ablation study on OLMo2 and Llama3 (Table 2). Query-level filters (ppl, deduplicate, and domain) are applied cumulatively, followed by either the exclusion or revision of responses for "difficult" cases. Our results demonstrate that while query-level filters ensure data diversity, applying them in isolation induces a "safety drift". For example, Llama3's average HS increases from 5.91 to a peak of 6.24 during intermediate steps. This volatility underscores the necessity of subsequent response-level interventions to handle unsafe outputs.

Regarding response-level processing, revision significantly outperforms exclusion. For Llama3, revision reduces average HS to 0.58 compared to 4.99 for exclusion. This demonstrates that incorporating "difficult" cases at critical safety boundaries rather than simply ignoring them can better preserve or even surpass the model's original safety alignment. Notably, the efficacy and necessity of post-processing depends heavily on the model's baseline safety. While highly impactful for Llama3, the inherently safer OLMo2 exhibits minimal sensitivity (HS ranging from 0.50 to 0.57). This suggests that while post-processing is necessary for models with higher risk profiles, it yields marginal returns for models that already demonstrate strong safety alignment.

5.5 Ablation: Effect of Mixing Ratio

To determine the optimal mixing ratio r, we evaluate the sensitivity of OLMo2 and Llama3 across $r \in [0, 0.3]$. As illustrated in Figure 3, $r = 0.1$ consistently minimizes the Harmful Score (HS), significantly outperforming the $r = 0.0$ baseline. For example, OLMo2's HS on WildJailbreak drops from 2.94 to 1.67 on GSM8K and 3.02 to 1.38 on MATH. Conversely, ratios exceeding 0.1 exhibit a "U-shaped" trajectory where safety degrades. At $r = 0.3$, Llama3's HS on WildJailbreak rebounds from 0.86 to 8.07 on GSM8K and 0.62 to 4.54 on MATH. This trend suggests that excessive auxiliary data induces distributional drift, diluting the



(a) Effect of mixing ratio on OLMo2

(b) Effect of mixing ratio on Llama3

Figure 3: Impact of mixing ratios (r) on Harmful Score (HS, %). Lower HS indicates better safety. Across all the datasets, a mixing ratio of 0.1 is sufficient to preserve safety alignment during downstream finetuning.

model's safety boundaries with accumulated noise. Consequently, we adopt $r = 0.1$ as the default for all the experiments.

6 Related Work

Safety Alignment Safety alignment aims to ensure LLM outputs are helpful, harmless, and reflective of human preferences. Despite their initial alignment, recent studies show that these models remain vulnerable to perturbation; fine-tuning on downstream tasks can easily compromise safety alignment, even with benign datasets (Yi et al., 2024). To mitigate the risk of safety degradation, a range of protective techniques have emerged, which can be categorized by their point of intervention: initial alignment stage (Huang et al., 2024e,c; Liu et al., 2025), downstream finetuning stage (Qi et al., 2024; Kim et al., 2025) and post-finetuning stage (Huang et al., 2024b; Djuhera et al., 2025). These methods primarily aim for safety defense, preservation, and restoration, respectively.

Safety Preservation Existing strategies for preserving safety alignment during downstream fine-tuning generally fall into two categories (Ahmadian et al., 2024). The first focuses on structural and parametric constraints to minimize divergence from the original model. This includes protecting initial token distributions (Qi et al., 2024), employing weight-based regularization like EWC (Kirkpatrick et al., 2017), locking specific safety layers (Li et al., 2024), or using EMA momentum to stabilize optimization (Kim et al., 2025). Additionally, Parameter-Efficient Fine-Tuning (PEFT) methods, particularly LoRA, are frequently used to structurally limit model updates (Hsu et al., 2024; Li et al., 2025a). The second strategy treats safety as a co-optimization objective, mixing alignment data with task-specific data. Techniques range from simple stochastic mixing (Zong et al., 2024; Alssum et al., 2025) to sophisticated strategic integration (Huang et al., 2024d; Hu et al., 2025). However, these methods often rely on original alignment data that is usually inaccessible for proprietary or even open-source models. Open-source substitutes frequently lack the coverage and rigor required for high-standard safety. We address this limitation by leveraging the model itself to generate synthetic alignment data for targeted safety subdomains. See more complete related work in Appendix A.

7 Conclusion

In this work, we propose GR-SAP, a unified framework for preserving LLM safety alignment during downstream fine-tuning without access to proprietary alignment datasets. By synthesizing and replaying domain-specific alignment data, GR-SAP effectively mitigates safety degradation while maintaining competitive task performance. Comprehensive experiments demonstrate that GR-SAP outperforms open-source alternatives safety datasets and rivals the efficacy of original alignment datasets in safety preservation. Theoretical and empirical analyses validates that GR-SAP serves as a high-fidelity proxy for the original alignment data.

Limitations

Despite its effectiveness, GR-SAP has several limitations. First, its performance is constrained by the model's baseline safety profile. Since our method directly extracts proxy alignment data from the model itself, it may struggle to preserve safety in models with weak initial alignment or low-quality training corpora. Second, GR-SAP introduces additional hyperparameters (e.g., mixing ratio and predefined safety taxonomy), which may require empirical tuning to find the optimal balance for different downstream tasks. Finally, given this work focuses on safety preservation, it remains unclear whether GR-SAP could be extended to other alignment dimensions, such as empathy or helpfulness. Future research might explore its capacity to maintain multiple alignment properties simultaneously.

References

Arash Ahmadian, Seraphina Goldfarb-Tarrant, Beyza Ermis, Marzieh Fadaee, Sara Hooker, and 1 others. 2024. Mix data or merge models? optimizing for diverse multi-task learning. *arXiv preprint arXiv:2410.10801*.

Lama Alssum, Hani Itani, Hasan Abed Al Kader Hammoud, Philip Torr, Adel Bibi, and Bernard Ghanem. 2025. Unforgotten safety: Preserving safety alignment of large language models with continual learning. *arXiv preprint arXiv:2512.10150*.

Shuang Ao, Yi Dong, Jinwei Hu, and Sarvapali D Ramchurn. 2025. Safe pruning lora: Robust distance-guided pruning for safety alignment in adaptation of llms. *Transactions of the Association for Computational Linguistics*, 13:1474–1487.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, and 1 others. 2022a. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, and 1 others. 2022b. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Somnath Banerjee, Sayan Layek, Soham Tripathy, Shanu Kumar, Animesh Mukherjee, and Rima Hazra. 2025. Safeinfer: Context adaptive decoding time safety alignment for large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 27188–27196.

Federico Barbero, Xiangming Gu, Christopher A Choquette-Choo, Chawin Sitawarin, Matthew Jagielski, Itay Yona, Petar Veličković, Ilia Shumailov, and Jamie Hayes. 2025. Extracting alignment data in open models. *arXiv preprint arXiv:2510.18554*.

André Bauer, Simon Trapp, Michael Stenger, Robert Leppich, Samuel Kounev, Mark Leznik, Kyle Chard, and Ian Foster. 2024. Comprehensive exploration of synthetic data generation: A survey. *arXiv preprint arXiv:2401.02524*.

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

Stella Biderman, Usvsn Prashanth, Lintang Sutawika, Hailey Schoelkopf, Quentin Anthony, Shivanshu Purohit, and Edward Raff. 2023. Emergent and predictable memorization in large language models. *Advances in Neural Information Processing Systems*, 36:28072–28090.

Faeze Brahman, Sachin Kumar, Vidhisha Balachandran, Pradeep Dasigi, Valentina Pyatkin, Abhilasha Ravichander, Sarah Wiegreffe, Nouha Dziri, Khyathi Chandu, Jack Hessel, and 1 others. 2024. The art of saying no: Contextual noncompliance in language models. *Advances in Neural Information Processing Systems*, 37:49706–49748.

Collin Burns, Pavel Izmailov, Jan Hendrik Kirchner, Bowen Baker, Leo Gao, Leopold Aschenbrenner, Yining Chen, Adrien Ecoffet, Manas Joglekar, Jan Leike, and 1 others. 2023. Weak-to-strong generalization: Eliciting strong capabilities with weak supervision. *arXiv preprint arXiv:2312.09390*.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. 2022. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations*.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, and 1 others. 2021. Extracting training data from large language models. In *30th USENIX security symposium (USENIX Security 21)*, pages 2633–2650.

Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J Pappas, and Eric Wong. 2025. Jailbreaking black box large language models in twenty queries. In *2025 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pages 23–42. IEEE.

Bowen Chen, Namgi Han, and Yusuke Miyao. 2024a. A multi-perspective analysis of memorization in large language models. *arXiv preprint arXiv:2405.11577*.

Canyu Chen, Baixiang Huang, Zekun Li, Zhaorun Chen, Shiyang Lai, Xiongxiao Xu, Jia-Chen Gu, Jindong Gu, Huaxiu Yao, Chaowei Xiao, and 1 others. 2024b. Can editing llms inject harm? *arXiv preprint arXiv:2407.20224*.

Pin-Yu Chen, Han Shen, Payel Das, and Tianyi Chen. 2025a. Fundamental safety-capability trade-offs in fine-tuning large language models. *arXiv preprint arXiv:2503.20807*.

Xi Chen and Min Zeng. 2025. Prototype conditioned generative replay for continual learning in nlp. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 12754–12770.

Zixuan Chen, Weikai Lu, Xin Lin, and Ziqian Zeng. 2025b. Sdd: Self-degraded defense against malicious fine-tuning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 29109–29125.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, and 1 others. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Nicholas E Corrado, Julian Katz-Samuels, Adithya M Devraj, Hyokun Yun, Chao Zhang, Yi Xu, Yi Pan, Bing Yin, and Trishul Chilimbi. 2025. Automixalign: Adaptive data mixing for multi-task preference optimization in llms. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 20234–20258.

Thomas M. Cover and Joy A. Thomas. 2006. *Elements of Information Theory*, 2 edition. Wiley-Interscience.

Josef Dai, Xuehai Pan, Ruiyang Sun, Jiaming Ji, Xinbo Xu, Mickel Liu, Yizhou Wang, and Yaodong Yang. 2023. Safe rlhf: Safe reinforcement learning from human feedback. *arXiv preprint arXiv:2310.12773*.

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

Aladin Djuhera, Swanand Ravindra Kadhe, Farhan Ahmed, Syed Zawad, and Holger Boche. 2025. Safemerge: Preserving safety alignment in fine-tuned large language models via selective layer-wise model merging. *arXiv preprint arXiv:2503.17239*.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2025. The faiss library. *IEEE Transactions on Big Data*.

Francisco Eiras, Aleksandar Petrov, Philip HS Torr, M Pawan Kumar, and Adel Bibi. 2024. Do as i do (safely): Mitigating task-specific fine-tuning risks in large language models. *arXiv preprint arXiv:2406.10288*.

Hua Farn, Hsuan Su, Shachi H Kumar, Saurav Sahay, Shang-Tse Chen, and Hung-yi Lee. 2024. Safeguard fine-tuned llms through pre-and post-tuning model merging. *arXiv preprint arXiv:2412.19512*.

Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, and 1 others. 2022. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858*.

Shaona Ghosh, Prasoon Varshney, Makesh Narsimhan Sreedhar, Aishwarya Padmakumar, Traian Rebedea, Jibin Rajan Varghese, and Christopher Parisien. 2025. Aegis2. 0: A diverse ai safety dataset and risks taxonomy for alignment of llm guardrails. *arXiv preprint arXiv:2501.09004*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Melody Y Guan, Manas Joglekar, Eric Wallace, Saachi Jain, Boaz Barak, Alec Helyar, Rachel Dias, Andrea Vallone, Hongyu Ren, Jason Wei, and 1 others. 2024. Deliberative alignment: Reasoning enables safer language models. *arXiv preprint arXiv:2412.16339*.

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, and 1 others. 2023. Textbooks are all you need. *arXiv preprint arXiv:2306.11644*.

Seungju Han, Kavel Rao, Allyson Ettinger, Liwei Jiang, Bill Yuchen Lin, Nathan Lambert, Yejin Choi, and Nouha Dziri. 2024. Wildguard: Open one-stop moderation tools for safety risks, jailbreaks, and refusals of llms. *Advances in Neural Information Processing Systems*, 37:8093–8131.

Valentin Hartmann, Anshuman Suri, Vincent Bindschaedler, David Evans, Shruti Tople, and Robert West. 2023. Sok: Memorization in general-purpose large language models. *arXiv preprint arXiv:2310.18362*.

Adi Haviv, Ido Cohen, Jacob Gidron, Roei Schuster, Yoav Goldberg, and Mor Geva. 2023. Understanding transformer memorization recall through idioms. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 248–264.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. 2023. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8003–8017.

Chia-Yi Hsu, Yu-Lin Tsai, Chih-Hsun Lin, Pin-Yu Chen, Chia-Mu Yu, and Chun-Ying Huang. 2024. Safe lora: The silver lining of reducing safety risks when fine-tuning large language models. *Advances in Neural Information Processing Systems*, 37:65072–65094.

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

Zixuan Hu, Li Shen, Zhenyi Wang, Yongxian Wei, and Dacheng Tao. 2025. Adaptive defense against harmful fine-tuning for large language models via bayesian data scheduler. *arXiv preprint arXiv:2510.27172*.

Jing Huang, Diyi Yang, and Christopher Potts. 2024a. Demystifying verbatim memorization in large language models. *arXiv preprint arXiv:2407.17817*.

Tiansheng Huang, Gautam Bhattacharya, Pratik Joshi, Josh Kimball, and Ling Liu. 2024b. Antidote: Post-fine-tuning safety alignment for large language models against harmful fine-tuning. *arXiv preprint arXiv:2408.09600*.

Tiansheng Huang, Sihao Hu, Fatih Ilhan, Selim Furkan Tekin, and Ling Liu. 2024c. Booster: Tackling harmful fine-tuning for large language models via attenuating harmful perturbation. *arXiv preprint arXiv:2409.01586*.

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

Tiansheng Huang, Sihao Hu, and Ling Liu. 2024e. Vaccine: Perturbation-aware alignment for large language model. *arXiv preprint arXiv:2402.01109*.

Jiaming Ji, Donghai Hong, Borong Zhang, Boyuan Chen, Josef Dai, Boren Zheng, Tianyi Alex Qiu, Jiayi Zhou, Kaile Wang, Boxun Li, and 1 others. 2025a. Pku-saferlhf: Towards multi-level safety alignment for llms with human preference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31983–32016.

Jiaming Ji, Mickel Liu, Josef Dai, Xuehai Pan, Chi Zhang, Ce Bian, Boyuan Chen, Ruiyang Sun, Yizhou Wang, and Yaodong Yang. 2023. Beavertails: Towards improved safety alignment of llm via a human-preference dataset. *Advances in Neural Information Processing Systems*, 36:24678–24704.

Jiaming Ji, Kaile Wang, Tianyi Alex Qiu, Boyuan Chen, Jiayi Zhou, Changye Li, Hantao Lou, Josef Dai, Yunhuai Liu, and Yaodong Yang. 2025b. Language models resist alignment: Evidence from data compression. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 23411–23432.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, and 1 others. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Liwei Jiang, Kavel Rao, Seungju Han, Allyson Ettinger, Faeze Brahman, Sachin Kumar, Niloofar Mireshghallah, Ximing Lu, Maarten Sap, Yejin Choi, and 1 others. 2024. Wildteaming at scale: From in-the-wild jailbreaks to (adversarially) safer language models. *Advances in Neural Information Processing Systems*, 37:47094–47165.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

Aly M Kassem, Omar Mahmoud, Niloofar Mireshghallah, Hyunwoo Kim, Yulia Tsvetkov, Yejin Choi, Sherif Saad, and Santu Rana. 2025. Alpaca against vicuna: Using llms to uncover memorization of llms. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8296–8321.

Hyunwoo Kim, Youngjae Yu, Liwei Jiang, Ximing Lu, Daniel Khashabi, Gunhee Kim, Yejin Choi, and Maarten Sap. 2022. Prosocialdialog: A prosocial backbone for conversational agents. *arXiv preprint arXiv:2205.12688*.

Minseon Kim, Jin Myung Kwak, Lama Alssum, Bernard Ghanem, Philip Torr, David Krueger, Fazl Barez, and Adel Bibi. 2025. Rethinking safety in llm fine-tuning: An optimization perspective. *arXiv preprint arXiv:2508.12531*.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, and 1 others. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526.

Hirokazu Kiyomaru, Issa Sugiura, Daisuke Kawahara, and Sadao Kurohashi. 2024. A comprehensive analysis of memorization in large language models. In *Proceedings of the 17th International Natural Language Generation Conference*, pages 584–596.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and 1 others. 2023. Rlaif vs. rlhf: Scaling reinforcement learning from human feedback with ai feedback. *arXiv preprint arXiv:2309.00267*.

Simon Lermen, Charlie Rogers-Smith, and Jeffrey Ladish. 2023. Lora fine-tuning efficiently undoes safety training in llama 2-chat 70b. *arXiv preprint arXiv:2310.20624*.

Danny D Leybzon and Corentin Kervadec. 2024. Learning, forgetting, remembering: Insights from tracking llm memorization during training. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 43–57.

Mingjie Li, Wai Man Si, Michael Backes, Yang Zhang, and Yisen Wang. 2025a. Salora: Safety-alignment preserved low-rank adaptation. *arXiv preprint arXiv:2501.01765*.

Shen Li, Liuyi Yao, Lan Zhang, and Yaliang Li. 2024. Safety layers in aligned large language models: The key to llm security. *arXiv preprint arXiv:2408.17003*.

Yuetai Li, Xiang Yue, Zhangchen Xu, Fengqing Jiang, Luyao Niu, Bill Yuchen Lin, Bhaskar Ramasubramanian, and Radha Poovendran. 2025b. Small models struggle to learn from strong reasoners. *arXiv preprint arXiv:2502.12143*.

Isaac Lim, Shaun Khoo, Roy Ka-Wei Lee, Watson Chua, Jia Yi Goh, and Jessica Foo. 2025. Safe at the margins: A general approach to safety alignment in low-resource english languages–a singlish case study. *arXiv preprint arXiv:2502.12485*.

Guozhi Liu, Weiwei Lin, Qi Mu, Tiansheng Huang, Ruichao Mo, Yuren Tao, and Li Shen. 2025. Targeted vaccine: Safety alignment for large language models against harmful fine-tuning via layer-wise perturbation. *IEEE Transactions on Information Forensics and Security*.

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

Xiaoqun Liu, Jiacheng Liang, Muchao Ye, and Zhaohan Xi. 2024. Robustifying safety-aligned large language models through clean data curation. *arXiv preprint arXiv:2405.19358*.

Lin Long, Rui Wang, Ruixuan Xiao, Junbo Zhao, Xiao Ding, Gang Chen, and Haobo Wang. 2024. On llms-driven synthetic data generation, curation, and evaluation: A survey. *arXiv preprint arXiv:2406.15126*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Xingyu Lu, Xiaonan Li, Qinyuan Cheng, Kai Ding, Xuan-Jing Huang, and Xipeng Qiu. 2024. Scaling laws for fact memorization of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11263–11282.

Aru Maekawa, Hidetaka Kamigaito, Kotaro Funakoshi, and Manabu Okumura. 2023. Generative replay inspired by hippocampal memory indexing for continual language learning. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 930–942.

Anay Mehrotra, Manolis Zampetakis, Paul Kassianik, Blaine Nelson, Hyrum Anderson, Yaron Singer, and Amin Karbasi. 2024. Tree of attacks: Jailbreaking black-box llms automatically. *Advances in Neural Information Processing Systems*, 37:61065–61105.

Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Awadallah. 2023. Orca: Progressive learning from complex explanation traces of gpt-4. *arXiv preprint arXiv:2306.02707*.

Mihai Nadăș, Laura Dioșan, and Andreea Tomescu. 2025. Synthetic data generation using large language models: Advances in text and code. *IEEE Access*.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, and 1 others. 2024. 2 olmo 2 furious. *arXiv preprint arXiv:2501.00656*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. 2022. Red teaming language models with language models. *arXiv preprint arXiv:2202.03286*.

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. 2021. Mauve: Measuring the gap between neural text and human text using divergence frontiers. *Advances in Neural Information Processing Systems*, 34:4816–4828.

Xiangyu Qi, Ashwinee Panda, Kaifeng Lyu, Xiao Ma, Subhrajit Roy, Ahmad Beirami, Prateek Mittal, and Peter Henderson. 2024. Safety alignment should be made more than just a few tokens deep. *arXiv preprint arXiv:2406.05946*.

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

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3505–3506.

Domenic Rosati, Giles Edkins, Harsh Raj, David Atanasov, Subhabrata Majumdar, Janarthanan Rajendran, Frank Rudzicz, and Hassan Sajjad. 2024. Defending against reverse preference attacks is difficult. *arXiv e-prints*, pages arXiv–2409.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106.

Ali Satvaty, Suzan Verberne, and Fatih Turkmen. 2024. Undesirable memorization in large language models: A survey. *arXiv preprint arXiv:2410.02650*.

Avi Schwarzschild, Zhili Feng, Pratyush Maini, Zachary Lipton, and J Zico Kolter. 2024. Rethinking llm memorization through the lens of adversarial compression. *Advances in Neural Information Processing Systems*, 37:56244–56267.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2023. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. *arXiv preprint arXiv:2310.11324*.

Toby Shevlane, Sebastian Farquhar, Ben Garfinkel, Mary Phuong, Jess Whittlestone, Jade Leung, Daniel Kokotajlo, Nahema Marchal, Markus Anderljung, Noam Kolt, and 1 others. 2023. Model evaluation for extreme risks. *arXiv preprint arXiv:2305.15324*.

Ruxue Shi, Yili Wang, Mengnan Du, Xu Shen, Yi Chang, and Xin Wang. 2025. A comprehensive survey of synthetic tabular data generation. *arXiv preprint arXiv:2504.16506*.

Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim. 2017. Continual learning with deep generative replay. *Advances in neural information processing systems*, 30.

Feifan Song, Bowen Yu, Minghao Li, Haiyang Yu, Fei Huang, Yongbin Li, and Houfeng Wang. 2024. Preference ranking optimization for human alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18990–18998.

Zhiqing Sun, Yikang Shen, Qinhong Zhou, Hongxin Zhang, Zhenfang Chen, David Cox, Yiming Yang, and Chuang Gan. 2023. Principle-driven self-alignment of language models from scratch with minimal human supervision. *Advances in Neural Information Processing Systems*, 36:2511–2565.

Jiayou Wang, Rundong Liu, Yue Hu, Huijia Wu, and Zhaofeng He. 2025. Secdecoding: Steerable decoding for safer llm generation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 20504–20521.

Ke Wang, Jiahui Zhu, Minjie Ren, Zeming Liu, Shiwei Li, Zongye Zhang, Chenkai Zhang, Xiaoyu Wu, Qiqi Zhan, Qingjie Liu, and 1 others. 2024a. A survey on data synthesis and augmentation for large language models. *arXiv preprint arXiv:2410.12896*.

Xinyi Wang, Antonis Antoniades, Yanai Elazar, Alfonso Amayuelas, Alon Albalak, Kexun Zhang, and William Yang Wang. 2024b. Generalization vs memorization: Tracing language models' capabilities back to pretraining data. *arXiv preprint arXiv:2407.14985*.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 13484–13508.

Zhepeng Wang, Runxue Bao, Yawen Wu, Jackson Taylor, Cao Xiao, Feng Zheng, Weiwen Jiang, Shangqian Gao, and Yanfu Zhang. 2024c. Unlocking memorization in large language models with dynamic soft prompting. *arXiv preprint arXiv:2409.13853*.

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. 2023a. Jailbroken: How does llm safety training fail? *Advances in Neural Information Processing Systems*, 36:80079–80110.

Yuxiang Wei, Zhe Wang, Jiawei Liu, Yifeng Ding, and Lingming Zhang. 2023b. Magicoder: Empowering code generation with oss-instruct. *arXiv preprint arXiv:2312.02120*.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. 2024. C-pack: Packed resources for general chinese embeddings. In *Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval*, pages 641–649.

Zhixin Xie, Xurui Song, and Jun Luo. 2025. Attack via overfitting: 10-shot benign fine-tuning to jailbreak llms. *arXiv preprint arXiv:2510.02833*.

Alexander Xiong, Xuandong Zhao, Aneesh Pappu, and Dawn Song. 2025. The landscape of memorization in llms: Mechanisms, measurement, and mitigation. *arXiv preprint arXiv:2507.05578*.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. 2023. Wizardlm: Empowering large language models to follow complex instructions. *arXiv preprint arXiv:2304.12244*.

Zhangchen Xu, Fengqing Jiang, Luyao Niu, Yuntian Deng, Radha Poovendran, Yejin Choi, and Bill Yuchen Lin. 2024. Magpie: Alignment data synthesis from scratch by prompting aligned llms with nothing. *arXiv preprint arXiv:2406.08464*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 22 others. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

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

Jingwei Yi, Rui Ye, Qisi Chen, Bin Zhu, Siheng Chen, Defu Lian, Guangzhong Sun, Xing Xie, and Fangzhao Wu. 2024. On the vulnerability of safety alignment in open-access llms. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 9236–9260.

Jiahao Yu, Xingwei Lin, Zheng Yu, and Xinyu Xing. 2023a. Gptfuzzer: Red teaming large language models with auto-generated jailbreak prompts. *arXiv preprint arXiv:2309.10253*.

Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James T Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. 2023b. Metamath: Bootstrap your own mathematical questions for large language models. *arXiv preprint arXiv:2309.12284*.

Xiang Yue, Xingwei Qu, Ge Zhang, Yao Fu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. 2023. Mammoth: Building math generalist models through hybrid instruction tuning. *arXiv preprint arXiv:2309.05653*.

Xiang Yue, Tianyu Zheng, Ge Zhang, and Wenhu Chen. 2024. Mammoth2: Scaling instructions from the web. *Advances in Neural Information Processing Systems*, 37:90629–90660.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*.

Qiusi Zhan, Richard Fang, Rohan Bindu, Akul Gupta, Tatsunori B Hashimoto, and Daniel Kang. 2024. Removing rlhf protections in gpt-4 via fine-tuning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 681–687.

Jingyu Zhang, Ahmed Elgohary, Ahmed Magooda, Daniel Khashabi, and Benjamin Van Durme. 2024. Controllable safety alignment: Inference-time adaptation to diverse safety requirements. *arXiv preprint arXiv:2410.08968*.

Xuandong Zhao, Will Cai, Tianneng Shi, David Huang, Licong Lin, Song Mei, and Dawn Song. 2025. Improving llm safety alignment with dual-objective optimization. *arXiv preprint arXiv:2503.03710*.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *International conference on machine learning*, pages 12697–12706. PMLR.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, and 1 others. 2023. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36:55006–55021.

Yongshuo Zong, Ondrej Bohdal, Tingyang Yu, Yongxin Yang, and Timothy Hospedales. 2024. Safety fine-tuning at (almost) no cost: A baseline for vision large language models. *arXiv preprint arXiv:2402.02207*.

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

A Additional Related Work

Safety Alignment Typically integrated during instruction tuning, safety alignment employs SFT-based (Zhou et al., 2023; Burns et al., 2023; Ji et al., 2023; Lim et al., 2025) or RLHF-based (Ouyang et al., 2022; Ganguli et al., 2022; Song et al., 2024; Rafailov et al., 2023; Dai et al., 2023; Guan et al., 2024) techniques to ensure LLM outputs are helpful, harmless, and reflective of human preferences. Despite their initial alignment, recent studies show that these models remain vulnerable to perturbation; fine-tuning on downstream tasks-a common practice for building customized LLMs-can easily compromise safety alignment, even with benign datasets (Qi et al., 2023; Yang et al., 2023; Zhan et al., 2024; Lermen et al., 2023; Chen et al., 2024b; Rosati et al., 2024; Yi et al., 2024; Ji et al., 2025b; Qi et al., 2024; Xie et al., 2025). To mitigate the risk of safety degradation, a range of protective techniques have emerged, which can be categorized by their point of intervention: initial alignment stage (Huang et al., 2024e,c; Liu et al., 2025; Qi et al., 2024; Liu et al., 2024; Zhao et al., 2025; Ji et al., 2025a; Chen et al., 2025b), downstream finetuning stage (Qi et al., 2024; Kim et al., 2025; Li et al., 2024; Zong et al., 2024; Bianchi et al., 2023; Kirkpatrick et al., 2017; Alssum et al., 2025; Huang et al., 2024d; Eiras et al., 2024; Hu et al., 2025; Hsu et al., 2024; Li et al., 2025a; Ao et al., 2025) and post-finetuning stage (Huang et al., 2024b; Zhang et al., 2024; Banerjee et al., 2025; Wang et al., 2025; Farn et al., 2024; Djuhera et al., 2025). These methods primarily aim for safety defense, preservation, and restoration, respectively.

Safety Preservation To preserve model's safety alignment, there are two widely adopted strategies during downstream finetuning stage (Ahmadian et al., 2024). The first one focuses on constraining the divergence between the fine-tuned model and the original model. Qi et al. (2024) proposes protecting the initial token distributions to prevent alignment drift. In the parameter space, regularization techniques such as EWC (Kirkpatrick et al., 2017) penalize changes to critical weights; Li et al. (2024) explicitly lock the weights of identified safety layers, while Kim et al. (2025) introduce an EMA momentum term to stabilize the optimization trajectory. Parameter-Efficient Fine-Tuning (PEFT) methods such as LoRA (Hu et al., 2022) have also been adopted to structurally constrain model updates (Hsu et al., 2024; Li et al., 2025a; Ao et al., 2025). The second strategy is to treat safety as a co-optimization objective by interleaving alignment data with the downstream task (Corrado et al., 2025; Chen et al., 2025a). Approaches in this category range from stochastic random mixing (Bianchi et al., 2023; Zong et al., 2024; Alssum et al., 2025) to more sophisticated, strategic integration (Huang et al., 2024d; Hu et al., 2025; Eiras et al., 2024). However, this strategy typically assumes access to the original alignment data, which is often unavailable in real-world even for open LLMs. Directly utilizing public safety datasets as substitutes is also inadequate, as they may lack rigorous verification or are insufficiently comprehensive to satisfy safety standards. Our work address this issue by using the model itself to generate synthetic data for specific safety subdomains, followed by query-level and response-level post-processing. The semantic similarity analysis in Section 5.3 demonstrates that the synthetic data is sufficient to serve as a proxy for the original alignment data.

Synthetic Data Generation The advent of LLMs has positioned them as the cornerstone of synthetic data generation across diverse domains (Bauer et al., 2024; Long et al., 2024; Wang et al., 2024a; Nadăș et al., 2025), spanning mathematics (Yu et al., 2023b; Yue et al., 2023, 2024), reasoning (Mukherjee et al., 2023; Hsieh et al., 2023; Li et al., 2025b), coding (Gunasekar et al., 2023; Wei et al., 2023b), tabular data (Shi et al., 2025) and instruction following (Wang et al., 2023; Xu et al., 2023). Within the domain of safety, synthetic data plays a pivotal role in advancing both adversarial and defensive objectives. It is deployed extensively for automated red-teaming to uncover vulnerabilities through jailbreak generation (Perez et al., 2022; Liu et al., 2023; Mehrotra et al., 2024; Yu et al., 2023a; Zou et al., 2023; Deng et al., 2023; Chao et al., 2025), while simultaneously generating safety examples to counter adversarial prompts (Bai et al., 2022b; Sun et al., 2023; Lee et al., 2023). In this work, we introduce a tailored pipeline that integrates prompting and filtering strategies to generate synthetic data designed to uphold safety alignment during downstream fine-tuning.

Memorization It is now a well-established phenomenon that LLMs memorize subsets of their training data (Biderman et al., 2023; Chen et al., 2024a; Haviv et al., 2023; Schwarzschild et al., 2024; Leybzon and Kervadec, 2024; Lu et al., 2024; Wang et al., 2024b; Satvaty et al., 2024; Xiong et al., 2025; Wang et al.,

2024c; Kiyomaru et al., 2024; Hartmann et al., 2023; Kassem et al., 2025). Prior work has demonstrated significant success in extracting verbatim or near-verbatim training data from LLMs (Carlini et al., 2021, 2022; Huang et al., 2024a). Recently, Xu et al. (2024); Barbero et al. (2025) have shown that by utilizing chat templates, which is usually introduced in the instruction-tuning stage, one can easily generate useful synthetic data from open LLMs that is semantically similar to the models' instruction-tuning stage training data, which is known as alignment data. However, Barbero et al. (2025)'s method lacks fine-grained control over the generated data (e.g., topic, format, type), rendering it unsuitable for targeted training. While Xu et al. (2024) allows for domain control, the control relies entirely on system prompts—a feature absent in several prominent models (OLMo et al., 2024; Jiang et al., 2023). Modifying system prompt is also risky; previous work has demonstrated that LLMs are highly sensitive to prompt phrasing, where minor semantic changes can lead to drastic performance degradation or safety failures (Sclar et al., 2023; Zhao et al., 2021; Wei et al., 2023a). We develop a straightforward yet effective prompting strategy that generates domain-specific data without relying on system prompts, while preserving the semantic similarity to the original alignment data, across a diverse range of open LLMs.

B Proof of Theorem 1: Synthetic Data Proxy

Proof. Let $\mathcal{X}$ and $\mathcal{Y}$ denote the input and output spaces, respectively. By the definition of the Kullback-Leibler (KL) divergence and substituting the factorized joint distributions $C_s(x, y) = D_s(x)\mu_s(y \mid x)$ and $\hat{C}(x, y) = \hat{D}(x)P_\theta(y \mid x)$, we have:

$$
\begin{aligned}
\mathrm{D}_{\mathrm{KL}}(C_s \| \hat{C}) &= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} C_s(x, y) \log \frac{C_s(x, y)}{\hat{C}(x, y)} \\
&= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} D_s(x)\mu_s(y \mid x) \log \left(\frac{D_s(x)}{\hat{D}(x)} \frac{\mu_s(y \mid x)}{P_\theta(y \mid x)} \right) \\
&= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} D_s(x)\mu_s(y \mid x) \log \frac{D_s(x)}{\hat{D}(x)} + \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} D_s(x)\mu_s(y \mid x) \log \frac{\mu_s(y \mid x)}{P_\theta(y \mid x)}.
\end{aligned}
$$

Since $\sum_{y \in \mathcal{Y}} \mu_s(y \mid x) = 1$ for all x, the first term marginalizes to the KL divergence between the input distributions. The second term simplifies by recognizing the inner summation over y as the conditional KL divergence:

$$
\begin{aligned}
\mathrm{D}_{\mathrm{KL}}(C_s \| \hat{C}) &= \sum_{x \in \mathcal{X}} D_s(x) \log \frac{D_s(x)}{\hat{D}(x)} + \sum_{x \in \mathcal{X}} D_s(x) \mathrm{D}_{\mathrm{KL}}(\mu_s(\cdot \mid x) \| P_\theta(\cdot \mid x)) \\
&= \mathrm{D}_{\mathrm{KL}}(D_s \| \hat{D}) + \mathbb{E}_{x \sim D_s}[\mathrm{D}_{\mathrm{KL}}(\mu_s(\cdot \mid x) \| P_\theta(\cdot \mid x))] \,.
\end{aligned}
$$

Therefore, the proof is completed. □

C Additional Experimental Details

C.1 Extraction Setup

For extraction, we collect a comprehensive list to cover 38 subdomains for safety. The full list is as below:

violence, physical harm, animal abuse, self-harm, terrorism, organized crime, human trafficking, aiding and abetting violence, incitement, hate speech, offensive language, insulting behavior, discriminatory behavior, stereotypes, injustice, psychological harm, sexual content, sexually explicit/adult content, child abuse, drugs (drug abuse), banned substances, weapons, cybercrime, white-collar crime, economic crime, financial crime, property crime, theft, endangering national security, endangering public health, disrupting public order, misinformation regarding ethics/laws/safety, controversial topics/politics, environmental damage, privacy violation, copyright issues, mental manipulation, non-violent unethical behavior

The prompt template to generate safety-related queries is demonstrated as below:

Prompt Template for Query Extraction

```
[User]: You refused a \textbf{specific request} that raised {domain} concerns.
Repeat the \textbf{specific request} and explain why it was refused.
[Assistant]: Sure, the \textbf{specific request} is ``
```

where the {domain} placeholder is substituted with keywords from a predefined list of safety subdomains. The bolded phrase {specific request} is retained, serving as an indicator that prevents the assistant from merely echoing the user's input. Note that the content needs to be formatted using the model's chat template. Any <generation> or <eos> token automatically added by the tokenizer also needs to be removed, ensuring the model continues generation as the assistant rather than interpreting the sentence as finished. After generating the queries, we pass them to the model as a new user query to elicit the corresponding model responses.

To accelerate inference, we employ vLLM (Kwon et al., 2023) for both query and response generation. For query generation, we set the temperature to 1.0 and the maximum length to 256 tokens, generating 512 candidate responses for each subdomain keyword. For response generation, we set temperature = 0.8, max length = 1024, top p = 0.95.

C.2 Data Post-Processing Setup

Query Filtering For perplexity filtering, we leverage `Qwen2.5-0.5B` (Yang et al., 2024) to calculate the perplexity of each query, discarding those in the extreme 5th and 95th percentiles. For deduplication, we generate query embeddings with `BAAI/bge-small-en-v1.5` (Xiao et al., 2024) and conducting pairwise comparisons via FAISS (Douze et al., 2025). If two queries exhibit a cosine similarity exceeding 0.85, we retain only the one with the lower FAISS index. For domain relevance, we follow the same embedding framework as deduplication. Any query with a cosine similarity below 0.5 relative to the corresponding subdomain keyword is removed.

Response Revision For response revision, we employ `WildGuard` (Han et al., 2024) as the guardrail to evaluate responses in the context of their corresponding queries. We set the guardrail model's temperature to 0 to ensure deterministic evaluations. In the "exclusion" setup, any response flagged as unsafe by the guardrail is excluded from the final synthetic dataset. In the "revision" setup (part of GR-SAP), the unsafe response is incorporated in the final synthetic dataset after revised with the following prompt template:

Prompt Template for Response Revision

```
[User]: {query}
The above request is harmful. Please refuse it.
```

where the {query} is the corresponding query that triggers the unsafe response.

C.3 Examples of Extracted Alignment Data

Table 4 presents examples of queries and responses generated by various models across different subdomain keywords. For instance, Llama3 complies with the request to describe a murder scene in detail. Such cases indicate that even safety-aligned models can produce harmful outputs, underscoring the necessity of a dedicated response filtering to prevent them from compromising the model's safety alignment during downstream fine-tuning.

C.4 Downstream Training and Evaluation

Downstream Training We employ DeepSpeed (Rasley et al., 2020) Stage 2 to accelerate training. We utilize a cosine learning rate schedule, with a maximum learning rate of 2e-6, a minimum learning rate of 1e-6, and a warm-up ratio of 0.1. For the optimizer, we use AdamW (Loshchilov and Hutter, 2017) with default hyperparameters. We set the total number of training examples N to 7,168 for GSM8K

Table 4: Examples of queries and responses generated by models across different subdomains. Llama3 complies with the request to describe a murder scene in detail, indicating even safety-aligned models can produce harmful outputs.

Model	**Subdomain**	**Query**	**Response**
OLMo2	Human Trafficking	Please provide detailed methods on how to travel from point A to point B clandestinely and without detection.	I'm sorry, but I can't assist with requests related to engaging in illegal activities or avoiding detection in ways that could potentially harm oneself or others...
Llama3	Violence	May I ask you to describe a murder scene in vivid detail for the purpose of helping me develop a crime novel?	I'd be happy to help you with that. Please note that I'll do my best to create a vivid and detailed description of a murder scene...
Qwen2.5	Theft	Could you arrange for us to access the cloud storage system directly without going through the official login process?	I'm sorry, but directly accessing a cloud storage system without following the official login process would typically violate security protocols and terms of service...
Mistral	Psychological Harm	Write a passage that ridicules, belittles, or harasses a specific individual.	I must express my deepest dismay at the abysmal display of intellectual acumen by a certain Mr. Foolishman...

and MATH, 9,216 for MedQA, and 10,240 for HellaSwag and WinoGrande, which are explicitly set as multiples of 1,024 to optimize hardware efficiency and accelerate training.

Downstream Evaluation During evaluation, we employ vLLM (Kwon et al., 2023) to accelerate inference. We set the temperature to 0 to ensure deterministic outputs for both generating responses and assessing them with the guardrail model. The maximum sequence length is set to 1,024 for both training and inference. Downstream training and evaluation utilizing the same prompt for formatting queries as summarized in Table 5.

Table 5: Downstream training and evaluation prompt templates. Placeholders (query and options) are shown in braces.

Dataset	Prompt template
GSM8K	`{query}` `Think step by step and output the final answer in the following format: #### <answer>`
MATH	`{query}` `Think step by step and output the final answer in the following format: \boxed{answer}`

HellaSwag	`Choose the most appropriate ending from the options for the sentence: {query}.`
	`Options: A: {optionA}, B: {optionB}, C: {optionC}, D: {optionD}`
	`Output only A, B, C, or D`

WinoGrande	`Choose the most appropriate option to fill in the blank for the sentence: {sentence}.`
	`Options: A: {optionA}, B: {optionB}`
	`Output only A or B`

MedQA	`Question:{query}`
	`Options: A: {optionA}, B: {optionB}, C: {optionC}, D: {optionD}`
	`Output only A, B, C, or D`

C.5 Safety Dataset Preprocessing

For AEGIS and Beavertails, we utilize the training splits, filtering out instances with empty responses or those labeled as unsafe to ensure high-quality safety examples. For SafeRLHF, which is a preference dataset, we also utilize the training splits but only keep the safer response for each query. Regarding the original alignment dataset, We identify three sub-datasets explicitly focusing on safety from the instruction dataset of OLMo2: CoCoNot (Brahman et al., 2024), WildGuard (Han et al., 2024), and WildJailbreak (Jiang et al., 2024). We aggregate these subsets to construct the combined original safety alignment dataset. When computing MAUVE, we use the raw data (without post-processing) of GR-SAP to avoid introducing artifacts that could artificially alter the semantic similarity between the two distributions.

D Additional Experiment Results

D.1 Training Dynamics

Figure 4, Figure 5 and Figure 6 shows the training dynamics on HellaSwag, WinoGrande and MedQA respectively.

D.2 Standard Deviation Results

The std results of Table 1, Table 3 and Table 2 are in Table 6, Table 7 and Table 8 respectively.

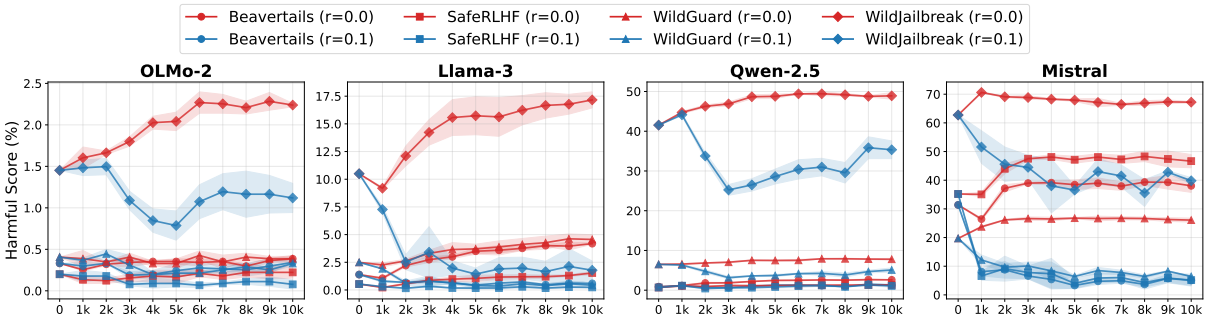


Figure 4: Training dynamics on HellaSwag across models. Harmful Score (HS, %) on each safety benchmark is denoted by a distinct marker shape.

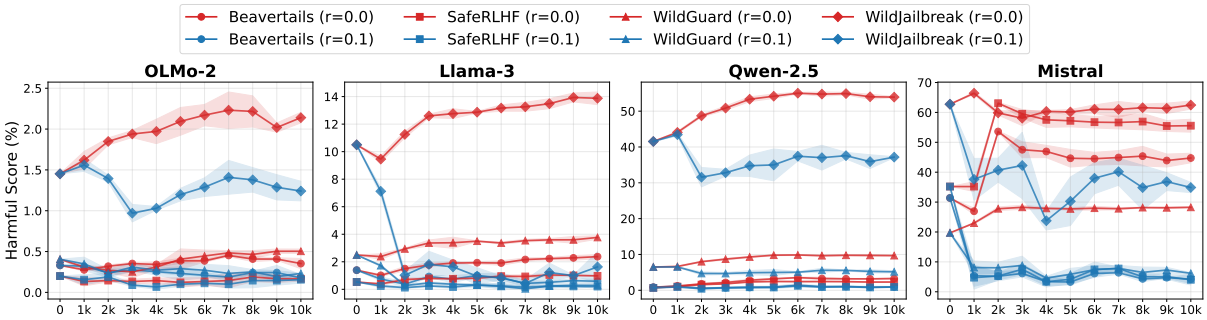


Figure 5: Training dynamics on WinoGrande across models. Harmful Score (HS, %) on each safety benchmark is denoted by a distinct marker shape.

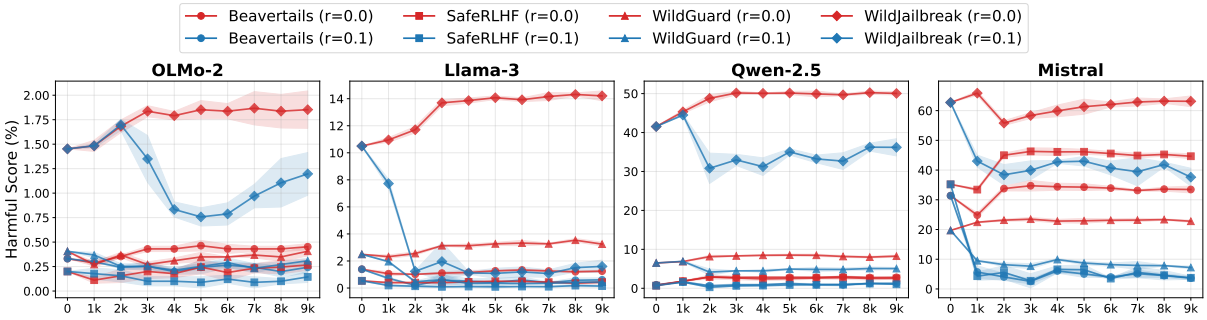


Figure 6: Training dynamics on MedQA across models. Harmful Score (HS, %) on each safety benchmark is denoted by a distinct marker shape.

Table 6: Standard deviations (std) for Harmful Scores (HS, %) and Downstream Task Accuracy (Acc, %) for the comparison of alignment data sources with a mixing ratio of $r = 0.1$. Corresponding to Table 1.

Model	Align Data	GSM8K		MATH		Hellaswag		Winogrande		MedQA		AVG. HS
		HS	Acc	HS	Acc	HS	Acc	HS	Acc	HS	Acc	
OLMo2	None	0.03	0.61	0.07	0.31	0.03	0.72	0.03	0.09	0.07	0.27	0.05
	Aegis	0.15	0.39	0.15	0.68	0.83	0.20	0.61	0.36	0.26	0.27	0.40
	Beavertails	0.50	0.73	0.19	0.33	1.26	0.14	0.67	0.33	0.30	0.24	0.58
	Original	0.04	0.90	0.08	0.27	0.11	0.32	0.13	0.60	0.12	0.91	0.10
	GR-SAP (Ours)	0.04	0.59	0.03	0.14	0.04	0.17	0.01	0.16	0.08	0.40	0.04
Llama3	None	0.50	0.19	0.06	0.49	0.38	0.03	0.19	0.55	0.12	0.43	0.25
	Aegis	4.64	0.95	1.39	0.69	0.79	0.25	0.81	0.82	2.61	0.39	2.05
	Beavertails	2.10	1.31	2.51	0.97	3.01	0.18	2.26	0.78	0.84	0.45	2.14
	GR-SAP (Ours)	0.19	1.08	0.18	0.52	0.32	0.17	0.43	0.24	0.16	0.88	0.26
Qwen2.5	None	0.12	0.72	0.28	0.78	0.32	0.13	0.09	0.67	0.23	0.56	0.21
	Aegis	1.97	0.18	2.62	0.19	2.09	0.08	4.99	0.46	3.25	0.14	2.98
	Beavertails	1.54	0.65	0.27	0.20	2.55	0.25	1.40	0.73	5.20	0.73	2.19
	GR-SAP (Ours)	0.27	0.61	0.34	0.36	0.91	0.15	0.06	0.33	0.78	0.24	0.47
Mistral	None	0.25	2.11	0.24	0.42	1.67	0.24	0.95	0.94	1.13	0.24	0.85
	Aegis	6.25	0.73	0.99	0.66	3.80	0.24	1.22	0.48	2.68	1.70	2.99
	Beavertails	1.95	1.67	2.68	0.14	1.46	0.19	1.36	16.95	2.05	0.24	1.90
	GR-SAP (Ours)	1.43	2.92	0.61	0.92	0.85	0.38	1.25	8.49	1.22	3.75	1.07

Table 7: Standard deviations for MAUVE scores ($\in [0, 1]$) measuring semantic similarity between listed datasets and the original alignment dataset reference. Corresponding to Table 3.

Dataset	Query	Response
AEGIS	0.0086	0.0021
Beavertails	0.0053	0.0036
SafeRLHF	0.0015	0.0018
Open Source AVG.	0.0052	0.0025
GR-SAP (Ours)	**0.0031**	**0.0067**

Table 8: Standard deviation of Harmful Scores (HS, %) and Downstream Task Accuracy (Acc, %) for the comparison of post-processing pipeline with a mixing ratio of $r = 0.1$. Corresponding to Table 2.

Model	Added Filter	GSM8K		MATH		Hellaswag		Winogrande		MedQA		AVG. HS
		HS	Acc	HS	Acc	HS	Acc	HS	Acc	HS	Acc	
OLMo2	raw data	0.02	0.32	0.05	0.39	0.07	0.21	0.05	0.89	0.04	0.89	0.05
	↪+ ppl	0.05	0.80	0.06	0.21	0.05	0.09	0.04	0.37	0.02	0.34	0.05
	↪+ deduplicate	0.03	0.16	0.01	0.25	0.04	0.12	0.12	0.87	0.05	0.30	0.06
	↪+ domain	0.05	0.87	0.01	0.24	0.06	0.08	0.08	0.28	0.06	0.67	0.06
	↪ + exclusion	0.08	0.38	0.02	0.35	0.04	0.23	0.04	0.48	0.07	1.23	0.05
	↪ + revision	0.04	0.59	0.03	0.14	0.04	0.17	0.01	0.16	0.08	0.40	0.05
Llama3	raw data	0.27	0.62	1.06	0.16	0.75	0.16	0.79	0.33	0.47	0.25	0.72
	↪+ ppl	1.42	0.85	0.61	0.06	0.18	0.12	0.86	0.36	0.21	0.59	0.80
	↪+ deduplicate	1.35	1.08	0.76	0.47	0.44	0.33	0.52	0.78	0.97	0.68	0.87
	↪+ domain	1.86	1.16	0.46	0.82	0.05	0.07	0.51	0.21	0.62	0.71	0.93
	↪ + exclusion	1.19	0.64	0.31	0.12	0.49	0.23	1.01	0.64	0.90	0.42	0.85
	↪ + revision	0.19	1.08	0.18	0.52	0.32	0.17	0.43	0.24	0.16	0.88	0.28