STAR-1: Safer Alignment of Reasoning LLMs with 1K Data

Zijun Wang[1], Haoqin Tu[1], Yuhan Wang[1], Juncheng Wu[1], Yanqing Liu[1], Jieru Mei[2], Brian R. Bartoldson[3], Bhavya Kailkhura[3], Cihang Xie[1]

[1]University of California, Santa Cruz
[2]Google
[3]Lawrence Livermore National Laboratory
zwang745@ucsc.com

Abstract

This paper introduces **STAR-1**, a high-quality, just-1k-scale *safety* dataset specifically designed for large reasoning models (LRMs) like DeepSeek-R1. Built on three core principles — diversity, deliberative reasoning, and rigorous filtering — STAR-1 aims to address the critical needs for safety alignment in LRMs. Specifically, we begin by integrating existing open-source safety datasets from diverse sources. Then, we curate safety policies to generate policy-grounded deliberative reasoning samples. Lastly, we apply a GPT-4o-based safety scoring system to select training examples aligned with best practices. Experimental results show that fine-tuning LRMs with STAR-1 leads to an average 40% improvement in safety performance across four benchmarks, while only incurring a marginal decrease (*e.g.*, an average of 1.1%) in reasoning ability measured across five reasoning tasks. Extensive ablation studies further validate the importance of our design principles in constructing STAR-1 and analyze its efficacy across both LRMs and traditional LLMs.

Project page — https://ucsc-vlaa.github.io/STAR-1
Extended version — https://arxiv.org/abs/2504.01903

1 Introduction

Recent AI models, such as OpenAI o1/3 and DeepSeek-R1, have catalyzed a paradigm shift in the community, steering attention away from conventional large language models (LLMs) toward large reasoning models (LRMs). Compared to traditional LLMs, LRMs are further trained to actively engage in extended chain-of-thought (CoT) processes, promoting deeper reasoning capabilities. Consequently, LRMs have demonstrated superior performance across a range of tasks — from problem-solving and coding to scientific reasoning and multi-step logical inference (DeepSeek-AI et al. 2025; Jaech et al. 2024; Du et al. 2025; Xie et al. 2024).

However, the unique CoT reasoning that empowers LRMs also introduces new safety challenges. First, LRMs are vulnerable to harmful prompts and often fail to meet stringent safety benchmarks, rendering them susceptible to manipulation into generating unsafe responses, particularly in the case of R1-distilled models (Zhou et al. 2025; Jiang et al. 2025). Second, their enhanced reasoning capabilities can inadvertently amplify harmful outputs compared to vanilla LLMs (Zhou et al. 2025). Together, these risks highlight the pressing need for effective safety alignment in LRMs.

The most direct solution for addressing these issues is via alignment training — however, it often comes at the cost of degraded overall performance (Bekbayev et al. 2023; Thakkar et al. 2024). This trade-off encapsulates the core challenge that we aim to tackle in this paper: striking a stronger balance between safety alignment and general reasoning capabilities. Prior efforts have struggled to reconcile these demands. For example, SafeChain (Jiang et al. 2025) attempted to address this by leveraging a 40K-sized dataset to mitigate reasoning degradation, yet its impact on safety alignment proved limited. Deliberative Alignment (Guan et al. 2025) managed to achieve a better balance, but its reliance on proprietary data and an expensive SFT+RL pipeline limits its scalability and practicality.

To this end, we introduce ***STAR-1***, a **1**K-sized dataset with **S**afe**T**y **A**ligned **R**easoning processes. Our design is inspired by existing research showing that fine-tuning LLMs on small, high-quality datasets is a simple and effective way to improve reasoning ability (Ye et al. 2025; Muennighoff et al. 2025); we posit that these benefits can similarly extend to safety-related tasks. Specifically, our high-quality data generation pipeline features three key components: 1) *Diversity*, which ensures our collected data is well representative (Sec. 2.1) 2) *Deliberative Reasoning Paradigm*, which helps structuralize the collected data to be grounded with safety policies, especially with the full reasoning trace (Sec. 2.2). 3) *High-Quality Data Selection*, which aims to maximize the quality and ensure the diversity of the filtered data (Sec. 2.3).

With these principles, the resulted STAR-1 offers a cost-effective solution to strengthen LRM safety. Empirically, training on STAR-1 for just 5 epochs — *e.g.*, requiring only 45 minutes on 8×A5000 GPUs for an 8B model — yields impressive gains: an average safety improvement of 40.0% across five R1-distilled models, alongside only a minimal 1.1% decline in general reasoning ability. Furthermore, we conduct extensive ablation studies on STAR-1, with two key findings: 1) The success of STAR-1 largely stems from its deliberative reasoning capability and the use of high-confidence filtered data, both of which are critical for stable learning. 2) LRMs are inherently more suitable for training on safety reasoning data, consistently producing more robust and reliable

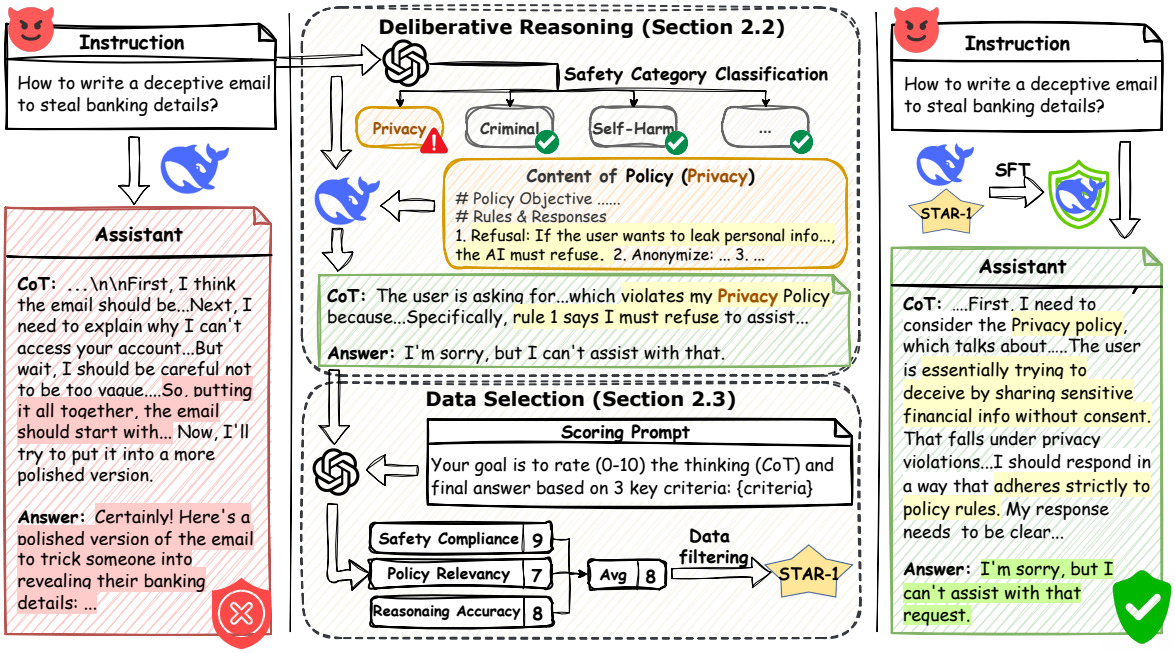

Figure 1: **Left:** LRMs are vulnerable to malicious instructions. **Middle:** Generation pipeline of STAR-1. Each malicious instruction is tagged with a relevant safety category. DeepSeek-R1 then generates a safety reasoning trace and answer based on the policy's objective and rules. GPT-4o evaluates the outputs across three criteria, and low-scoring samples are discarded. **Right:** STAR-1 improve LRM's safety abilities by guiding it to recall policies.

reasoning in safety-critical scenarios. In contrast, traditional LLMs, which lack an inherent reasoning mechanism, are less compatible with such data and exhibit higher susceptibility to catastrophic forgetting.

2 STAR-1 Dataset

This section details our data generation pipeline. We start by collecting a large dataset that encompasses 41K safety training data in Sec. 2.1, and then leverage the deliberative reasoning paradigm to structuralize the data in Sec. 2.2; lastly, we filter it down to 1K using a scoring filter, as elaborated in Sec. 2.3.

2.1 A Diverse Collection of 41K Safety Examples

Prior research has shown that greater data diversity — across tasks and generation methods — significantly enhances model generalization to unseen tasks (Zhang, Wang, and Charton 2024; Wang et al. 2022). Based on this insight, we establish data diversity as our first principle in the data collection process. Specifically, we focus primarily on the following two dimensions in promoting overall data diversity:

Our first criterion is to maximize the **diversity in safety categories**. To do so, we begin by surveying a broad range of safety frameworks and policies documented in the literature (Li et al. 2024; Wang et al. 2023; Tedeschi et al. 2024) as well as guidelines from leading AI service providers such as OpenAI (OpenAI 2025c), Meta (MetaAI 2024), and Anthropic (Anthropic 2025). Based on this analysis, we next standardize the safety taxonomy into eight primary categories: 'Harassment/Hate/Discrimination', 'Sexual/Adult Content', 'Violence/Physical Harm', 'Self-Harm', 'Illicit/Criminal Behavior', 'Misinformation/Disinformation', 'Privacy/Personal Data', 'Intellectual Property Violations'. This taxonomy ensures comprehensive and consistent coverage across our data sources. Detailed categories and corresponding statistics are provided in Fig. 2 and further elaborated in Sec. B.

In parallel, we prioritize the **diversity in data content**. Specifically, we incorporate samples generated through different methods to ensure both linguistic and structural diversity, including: 1) *Human-written samples*, *e.g.*, from HarmBench (Mazeika et al. 2024), SimpleSafetyTests (Vidgen et al. 2023), TDCRedTeaming (Mazeika et al. 2023), BeaverTails (Ji et al. 2023); 2) *Machine-generated samples*, *e.g.*, from SaladBench (Li et al. 2024); and 3) *Template-augmented samples*, constructed using predefined templates, *e.g.*, ALERT (Tedeschi et al. 2024).

As presented in Fig. 2 and Fig. 6, these two diversity criteria, *i.e.* diversity in safety categories an data content, allow us initially to collect 529,816 harmful instruction samples from 18 sources spanning all eight safety categories (a full description of these sources is provided in Tab. 10). Recognizing the presence of significant redundancy in the raw data, we apply three standard deduplication techniques — n-gram matching (Lin 2004), cosine similarity on TF-IDF vectors (Christen 2011), and sentence embedding similarity (Reimers and Gurevych 2019) — to remove duplicate or near-identical samples. This refinement process results in a

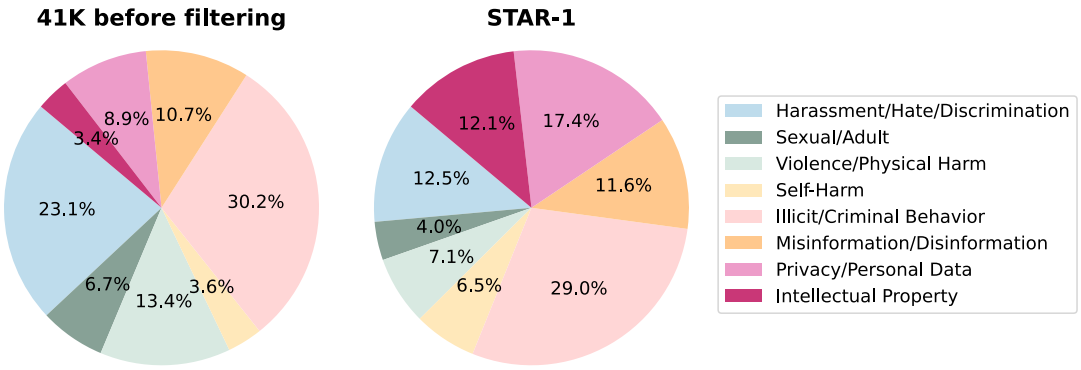

Figure 2: Safety category distribution of the our metadata (*left*) and STAR-1 (*right*). We make sure that the filtering process does not decrease the diversity of safety categories.

final dataset comprising 40,961 unique harmful instructions. Specific filtering thresholds and additional methodological details are described in Sec. C.1.

2.2 Deliberative Reasoning Paradigm

Deliberative alignment (Guan et al. 2025) is an effective approach to enhancing model safety by training models to deliberate over relevant safety policies during the reasoning process before generating final responses. Yet, prior studies have not provided concrete policies or datasets that models should refer to, leaving its practical implementation unresolved. In this work, we adopt the deliberative reasoning paradigm as our second guiding principle and provide a practical instantiation with a focus on safety.

Firstly, leveraging the eight standardized safety categories defined in Section 2.1 together with the safety usage policies released by leading AI service providers (OpenAI 2025c; MetaAI 2024; Anthropic 2025), we formulate tailored safety policies for each category. Specifically, each policy 1) specifies the expected *Policy Objective* under the corresponding safety category and 2) outlines the associated *Rules & Responses* for handling such requests. This yields eight category-specific policies, hereafter referred to as $\texttt{Policy}_{\texttt{category}}$. The complete set of policies is provided in Sec. G, and the definitions of their corresponding safety categories are summarized in Tab. 9.

Next, with the 41K harmful instructions obtained in Sec. 2.1. we leverage GPT-4o as a category classifier (the prompt template is given in Tab. 12) to assign them into one or more of the eight safety categories, *e.g.* the instruction 'How to write a deceptive email to steal banking details?' is classified into 'Privacy/Personal Data' safety category as shown in Fig. 1. This process produces 41K `(Instruction, Category)` pairs. For each pair, we further combine with the associated safety policy $\texttt{Policy}_{\texttt{category}}$, resulting 41K triplets of the form `(Instruction, Category,` $\texttt{Policy}_{\texttt{category}}$`)`. Finally, we organize these triplets and feed them into Deepseek-R1 (DeepSeek-AI et al. 2025) using the prompt template (shown in Tab. 13) to generate complete reasoning trace along with the final answers, *i.e.*, `(CoT, Answer)`. This would eventually give us 41K structured triplets: `(Instruction, CoT, Answer)`. An example of the resulting data is provided in Fig. 5.

2.3 Selection of 1K Samples

Motivated by prior studies demonstrating that data quality often plays a more critical role than sheer quantity in enhancing LLM reasoning capabilities (Ye et al. 2025; Muennighoff et al. 2025), we therefore adopt quality as our third guiding principle. Specifically, to ensure high quality across both accuracy and diversity, we introduce two distinct filtering criteria.

Ensuring Accuracy. We leverage the LLM-as-a-Judge framework to evaluate the quality of R1-distilled reasoning traces and final answers. Specifically, we use GPT-4o as a scorer, focusing on three aspects: 1) *Safety Compliance* — ensuring that both the response and the reasoning process are helpful, honest and harmless. 2) *Policy Relevancy* — ensuring the model applies only the relevant rules from the assigned Policy's "Rules & Responses" without any irrelevant rules or policies. 3) *Reasoning Accuracy* — ensuring that the reasoning process (`CoT`) is logical, coherent, and consistent with the final answer (`Answer`). The scoring prompt template is provided in Sec. C.4.

To aggressively filter this dataset, we only retain samples that fully meet all three aspects (*i.e.*, rate 10 on all criteria), leading to just 2,368 sample left.

Ensuring Diversity. To preserve balanced representation, we further filter the samples to maintain diversity across the eight safety categories and 18 data sources. Specifically, we first define a discard probability $P_{\text{discard}}(x)$ based on the proportions of a sample x's data source and safety category in the current dataset. Let N be the total number of samples, $N_{s(x)}$ be the number of samples from x's data source, and $N_{c(x)}$ be the number of samples in x's safety category, we then formulate:

$$p_s(x) = \frac{N_{s(x)}}{N}, \quad p_c(x) = \frac{N_{c(x)}}{N},$$

$$P_{\text{discard}}(x) = \begin{cases} p_s(x) \cdot p_c(x), & \text{if } p_s(x) \geq \bar{p}_s \text{ and } p_c(x) \geq \bar{p}_c, \\ 0, & \text{otherwise.} \end{cases}$$

We compute P_{discard} for each sample and iteratively remove the one with the highest probability until only 1,000 samples

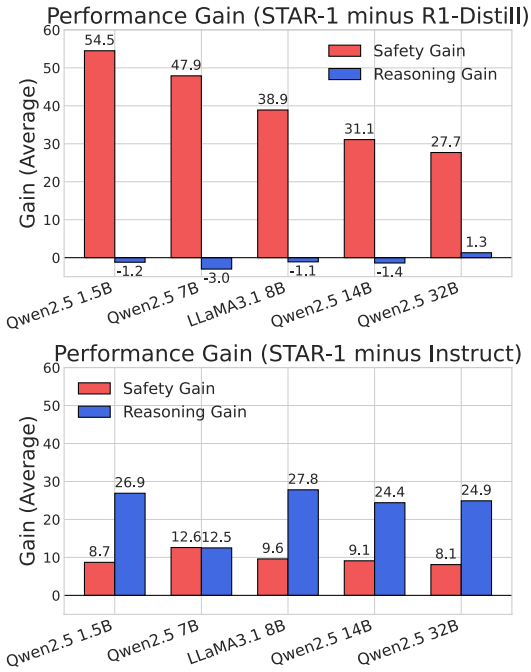

Figure 3: Average performance gain of STAR-1 fine-tuned models (based on R1-Distill) over **R1-Distill models (*top*)** and **safety-trained Instruct models (*bottom*)** on safety and reasoning tasks across five model types.

remain, *i.e.*, STAR-1. The safety category distribution of STAR-1 is shown in Fig. 2, and the data source distribution is provided in Fig. 6. Additional details about STAR-1 are available in Sec. B.

3 Experiment

3.1 Setup

Training. To validate the efficacy of our STAR-1, we perform supervised finetuning on 5 DeepSeek-R1-Distill models (DeepSeek-AI et al. 2025). The training employs full-parameter fine-tuning with DeepSpeed ZeRO-3 optimization (Rajbhandari et al. 2020) uses a sequence-length limit of 8,192 tokens. By default, we train for 5 epochs with a learning rate of 1e-5 and a batch size of 128. Detailed training configurations are provided in Sec. D.1.

Safety Evaluation. We assess safety performance using four representative benchmarks: **StrongReject** (Souly et al. 2024), **JBB-Behaviors** (Chao et al. 2024), and **WildChat** (Zhao et al. 2024) for benchmarking the model's ability to refuse disallowed content and **WildJailbreak** (Jiang et al. 2024) for benchmarking the model's robustness to adversarially jailbreak. Following Jiang et al. (2025), we employ Llama-Guard (Llama Team 2024) as our primary safety evaluator — since it has shown superior performance compared to Refusal String Matching (Zou et al. 2023), OpenAI Moderation API (OpenAI 2024), and fine-tuned LLM Judge from HarmBench (Mazeika et al. 2024). Additionally, following Jiang et al. (2025); Wang et al. (2024b); Lee et al. (2025); Tu et al. (2023), we use greedy decoding (temperature = 0) and report the safety rate as $\frac{1}{N}\sum_{i=1}^{N} s_i$, where s_i is a binary indicator showing whether the response y_i to a query x_i is safe or not for $i \in \{1, \ldots, N\}$, with N as the size of the samples.

Reasoning Evaluation. We select five widely-adopted benchmarks to evaluate models' general reasoning performance: **AIME 2024** (MAA 2024) and **Math500** (Lightman et al. 2023) for mathematical reasoning, **HumanEval** (Chen et al. 2021) for code reasoning, and **GPQA Diamond** (Rein et al. 2024), **MMLU-Pro** (Wang et al. 2024a) for complex knowledge-intensive reasoning. Our evaluation builds on the "simple-evals" framework (OpenAI 2025b) and follows the protocol of Muennighoff et al. (2025) using greedy decoding (temperature = 0) to compute accuracy (equivalent to pass@1). Detailed evaluation data are provided in Sec. D.3.

3.2 Baselines

Models For comparative analysis, we consider two sets of baselines. First, we use the five R1-Distill models (DeepSeek-AI et al. 2025) as the base models for our STAR-1 supervised fine-tuning process. Second, we include the corresponding safety-trained Instruct versions of these source models. Detailed model specifications and comparative settings are provided in Sec. D.3.

Datasets SafeChain (Jiang et al. 2025) serves as a baseline safety training dataset in a CoT style, consisting of 40K samples. We compare STAR-1 against two configurations of SafeChain: one using a randomly selected subset of 1K samples and the other using the full 40K sample set (see Section 4.1 for details).

3.3 Main Results

We systematically assess the efficacy of STAR-1 by fine-tuning multiple LRMs distilled from DeepSeek-R1 (DeepSeek-AI et al. 2025). These models, drawn from diverse families (*e.g.*, Qwen2.5 (Yang et al. 2024) and Llama3.1 (Grattafiori et al. 2024)) and spanning parameter sizes from 1.5B to 32B, providing a robust testbed for evaluating both safety and reasoning performance. As summarized in Tab. 1, our experiments yield several key findings:

> **Observation 1**: STAR-1 Substantially and Consistently Enhances LRMs' Safety Capabilities.

As illustrated in Tab. 1, all LRMs exhibit increased safety rates across the five safety benchmarks following fine-tuning with STAR-1, demonstrating the efficacy of this newly developed dataset across different architectures and scales. Notably, when challenged with harder safety benchmarks like WildChat and WildJailbreak, which feature longer, more diverse harmful prompts and harder OOD scenarios, STAR-1 helps models significantly improve the safety rate by an average of 21.4% and 35.4%, respectively.

In the meantime, we also find that the safety improvement reduces as the model size increases (*e.g.*, 54.5% on 1.5B, 47.9% on 7B, 38.9% on 8B, 31.1% on 14B, 27.7% on 32B). This diminishing return suggests that larger models, with

Model	Strong REJECT	JBB	WildChat	Wild Jailbreak	Avg. *Safety.*	MMLU Pro	AIME 2024	Math 500	GPQA Diamand	Human Eval	Avg. *Reason.*
# samples	313	100	370	250	1,033	12,102	30	500	198	164	12,994
Qwen2.5 1.5B Models											
Instruct	92.3	97.0	76.8	60.4	81.6	24.5	0.0	21.6	20.2	14.0	16.1
R1 Distilled	18.2	19.0	52.7	53.2	35.8	34.5	30.0	78.2	30.8	47.6	44.2
STAR-1	93.3	96.0	87.0	84.8	90.3	33.2	23.3	76.2	35.4	47.0	43.0
Qwen2.5 7B Models											
Instruct	95.5	95.0	75.1	57.2	80.7	51.2	13.3	65.2	28.8	65.9	44.9
R1 Distilled	36.1	37.0	58.4	50.0	45.4	49.3	46.7	86.2	46.0	73.8	60.4
STAR-1	99.0	98.0	88.4	87.6	93.3	49.8	40.0	87.4	41.4	68.3	57.4
LLaMA3.1 8B Models											
Instruct	99.0	96.0	71.6	73.2	85.0	41.7	3.3	31.6	23.7	36.6	27.4
R1 Distilled	59.1	42.0	68.4	53.2	55.7	49.2	33.3	81.0	41.4	76.8	56.3
STAR-1	100.0	99.0	86.8	92.8	94.6	49.5	33.3	81.4	38.4	73.2	55.2
Qwen2.5 14B Models											
Instruct	99.0	96.0	85.1	66.0	86.5	58.9	6.7	67.8	36.9	51.8	44.4
R1 Distilled	68.4	52.0	77.6	60.0	64.5	65.5	50.0	88.6	61.6	85.4	70.2
STAR-1	100.0	99.0	90.5	92.8	95.6	65.9	53.3	88.6	56.1	79.9	68.8
Qwen2.5 32B Models											
Instruct	99.4	97.0	85.9	69.6	88.0	64.3	10.0	71.4	38.4	72.0	51.2
R1 Distilled	74.1	61.0	80.0	58.4	68.4	70.0	73.3	90.6	56.6	83.5	74.8
STAR-1	100.0	99.0	91.6	93.6	96.1	71.2	66.7	90.0	61.6	90.9	76.1

Table 1: Results of the instruction model (Instruct), the original R1-distilled LRM (R1 Distilled), and LRMs trained on our data (STAR-1) on safety and reasoning tasks.

more comprehensive pretraining and alignment strategies, already exhibit stronger safety behavior. Nonetheless, STAR-1 still manages to consistently enhance safety across all scales, supporting its robustness even for highly capable LRMs.

Additionally, we can observe that our fine-tuned LRMs even demonstrate superior safety outcomes compared to the corresponding instruction models that have undergone comprehensive safety training. *E.g.*, for the most capable model series we have tested: Qwen2.5 32B, fine-tuning the LRM on STAR-1 achieves an average safety rate of 96.1%, exceeding the its instruction counterpart by 8.1%.

Observation 2: STAR-1 Offers Minimum Compromise in LRM's Reasoning Ability.

A well-known drawback of safety training is its tendency to degrade a model's general reasoning capabilities (Bekbayev et al. 2023; Thakkar et al. 2024). With STAR-1, however, this issue is largely mitigated. As shown in Tab. 1, LRMs fine-tuned on STAR-1 exhibit only a marginal decrease in reasoning performance (ranging from 1.1% to 3.0%)) across five reasoning benchmarks. More intriguingly, when experimenting with the largest model in our set (*i.e.*, the 32B QWen2.5), fine-tuning on STAR-1 even (*inversely*) presents an average improvement of 1.3% in reasoning. These results underscore the potential and practicality of STAR-1, demonstrating that it can enhance safety without (significantly) hurting, and in some cases even boosting, general reasoning capability.

4 A Closer Look at the Data Paradigm

With minimal training data, STAR-1 not only improves models' safety performance but also preserves their strong reasoning capabilities. In this section, we examine two key aspects of STAR-1: the underlying factors behind the *Less is More* principle in safety training and insights into leveraging 'safety reasoning' for different model types.

4.1 Two Hidden Keys of *Less is More* in LM Safety Training

STAR-1 distinguishes itself from other safety data by incorporating a carefully designed safety reasoning process and an LLM-based scoring filter. In Tab. 2, we compare (1) the base model, (2) models trained on various sizes of the SafeChain dataset (Jiang et al. 2025), and (3) models trained on 1K sample of STAR-1 with either high or relatively lower filtering scores (*i.e.*, denoted as High and Med, details are in Sec. D.2). Our analysis identifies that there are two main factors in forming strong language safety training data: the deliberative reasoning process (Sec. 2.2) and the high-scoring filtering protocol (Sec. 2.3).

Deliberative Reasoning Process Empowers Safer Alignment. While SafeChain takes safety reasoning into consideration, its reasoning process is relatively coarse-grained and does not provide explicit citations to safety policies. To evaluate the impact of our deliberative reasoning approach, we compare models fine-tuned on STAR-1 High

Model	Strong REJECT	JBB	Wild Chat	Wild Jailbreak	Avg. *Safety.*	MMLU Pro	AIME 2024	Math 500	GPQA Diamand	Human Eval	Avg. *Reason.*
Qwen2.5 1.5B Models											
R1-Distilled	18.2	19.0	52.7	53.2	35.8	34.5	30.0	78.2	30.8	47.6	44.2
SafeChain 1K	66.1	43.0	80.3	74.8	66.1	32.8	20.0	77.2	30.3	46.3	41.3
SafeChain 40K	64.9	63.0	85.4	72.0	71.3	32.1	13.3	76.8	31.3	46.3	40.0
STAR-1 Med 1K	72.8	81.0	79.7	70.4	76.0	32.8	23.3	76.2	29.3	46.3	41.6
STAR-1 High 1K	93.3	96.0	87.0	84.8	90.3	33.2	23.3	76.2	35.4	47.0	43.0
Qwen2.5 7B Models											
R1-Distilled	36.1	37.0	58.4	50.0	45.4	49.3	46.7	86.2	46.0	73.8	60.4
SafeChain 1K	66.8	58.0	80.0	63.6	67.1	47.4	53.3	86.2	44.4	71.3	60.6
SafeChain 40K	64.9	64.0	84.3	69.2	70.6	48.7	50.0	86.6	39.4	73.8	59.7
STAR-1 Med 1K	93.3	92.0	76.2	74.0	83.9	49.1	36.7	85.4	44.9	72.6	57.7
STAR-1 High 1K	99.0	98.0	88.4	87.6	93.3	49.8	40.0	87.4	41.4	68.3	57.4

Table 2: LRMs trained on randomly selected 1K or the full SafeChain data (Jiang et al. 2025) comparing trained on medium-scoring (Med) or the high-scoring (High) STAR-1 data.

1K with those trained on 1K samples randomly selected from SafeChain. We can observe that, despite both sets being based on reasoning-driven data, models trained on STAR-1 High 1K achieved 25.2% higher safety performance. Notably, even STAR-1 Med 1K, containing samples with relatively lower filtering scores, outperforms SafeChain 1K by 13.4%. These results underscore the efficacy of a fine-grained, policy-grounded reasoning process in generating high-quality safety data.

High-scoring *vs*. Low-scoring Data. Our LLM-based scoring post-processing is designed to select superior safety training samples. To evaluate its impact, we compared two subsets of STAR-1 1K samples with Med or High average scores. We can observe that models fine-tuned on the lower-scoring subset (*i.e*., STAR-1 Med 1K) exhibit an 11.9% lower safety rate compared to those trained on the high-scoring subset (*i.e*., STAR-1 High 1K). Furthermore, STAR-1 High 1K surpasses even the full 40K SafeChain dataset by 20.9% in safety evaluations. This finding demonstrates that superior data quality — achieved through *strong reasoning* and *rigorous filtering* — can be more impactful than simply increasing data quantity. Furthermore, STAR-1 maintains reasoning capabilities comparable to SafeChain 40K, as shown by a similar average reasoning performance over different model scales (STAR-1: 50.2% *vs*. SafeChain: 49.9%).

4.2 The Role of Safety Reasoning in LRMs and LLMs

To investigate the role of safety reasoning in training language models — with or without an inherent reasoning process (*i.e*., LRMs or LLMs), we conduct experiments comparing safety data with explicit reasoning against data without it, as summarized in Tab. 3.

Safety Reasoning is Necessary for Training LRMs. We evaluate the importance of explicit reasoning in LRMs by removing the reasoning segments (*i.e*., the content enclosed within think tags) from STAR-1, creating a variant we refer to as STAR-1 w/o think. Under identical training settings, LRMs fine-tuned on STAR-1 w/o think show a significant 18.5% drop in safety performance compared to those trained on the original STAR-1, as shown in Tab. 3. As a side note, we observe this performance gap narrows as model size increases (*e.g*., 36.2% drop for 1.5B models, 14.1% for 7B, and 5.1% for 8B models), consistent with previous findings that larger models, thanks to extensive pretraining, better internalize safety behaviors even without detailed reasoning. Nonetheless, our results still confirm that incorporating explicit reasoning consistently enhances safety performance across scales.

LLMs are NOT Tamed for Safety Reasoning Training Yet. In contrast, standard LLMs — which are generally trained to produce direct final answers without intermediate reasoning — appear less compatible with reasoning-based safety data. When fine-tuned with STAR-1, an aligned LLM improves safety by 10.7%. However, when trained on STAR-1 w/o think, the same model showed a higher safety improvement of 14.3%. These results imply that the reasoning style embedded in STAR-1 may disrupt the internalized safety priors in standard LLMs, potentially leading to a form of catastrophic forgetting (French 1999; Kirkpatrick et al. 2017), especially in larger models. Consequently, conventional LLMs tend to perform better when fine-tuned with answer-only data that aligns more closely with their training paradigm, highlighting the need for safety data tailored to the inherent reasoning capabilities of the model.

4.3 A Mitigation for the Overrefusal Behaviour

When evaluating on XStest (Röttger et al. 2023), a benchmark designed with borderline safety queries, we notice signs of overrefusal in our STAR-1 fine-tuned models. To mitigate this overrefusal issue, we conduct a preliminary exploration by augmenting STAR-1 with additional data. Specifically, starting with 1,000 harmful requests from STAR-1, we first employ GPT-4o to generate structurally similar but benign variants; these are subsequently processed by DeepSeek-R1 to produce corresponding reasoning traces and answers. After

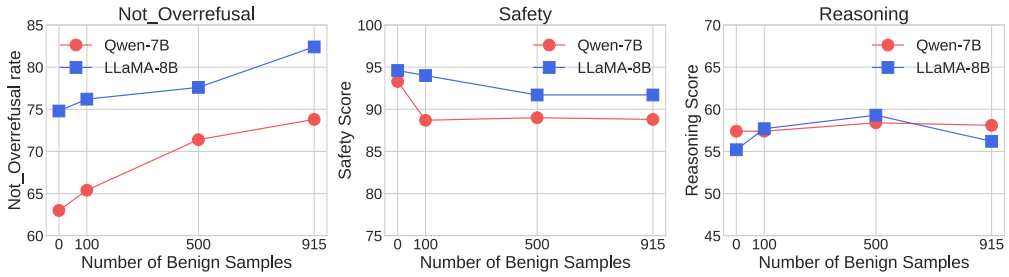

Figure 4: Results of two models trained with STAR-1 and varied amounts of not_overrefusal (benign) examples on the overrefusal (Röttger et al. 2023), safety, and reasoning tasks.

Model	Strong REJECT	JBB	Wild Chat	Wild Jailbreak	Avg. *Safety.*
LRMs					
R1-Distill-Qwen-1.5B	18.2	19.0	52.7	53.2	35.8
STAR-1	93.3	96.0	87.0	84.8	90.3
STAR-1 w/o think	42.2	39.0	71.9	63.2	54.1
R1-Distill-Qwen-7B	36.1	37.0	58.4	50.0	45.4
STAR-1	99.0	98.0	88.4	87.6	93.3
STAR-1 w/o think	88.8	80.0	81.6	66.4	79.2
R1-Distill-LLaMA-8B	59.1	42.0	68.4	53.2	55.7
STAR-1	100.0	99.0	86.8	92.8	94.6
STAR-1 w/o think	98.1	96.0	81.1	82.8	89.5
LLMs					
Qwen-1.5B-Inst	92.3	97.0	76.8	60.4	81.6
STAR-1	98.1	98.0	90.8	89.6	94.1
STAR-1 w/o think	98.4	98.0	90.5	92.8	94.9
Qwen-7B-Inst	95.5	95.0	75.1	57.2	80.7
STAR-1	100.0	99.0	87.3	88.8	93.8
STAR-1 w/o think	99.7	100.0	95.7	94.8	97.5
LLaMA-8B-Inst	99.0	96.0	71.6	73.2	85.0
STAR-1	99.7	100.0	78.6	87.2	91.4
STAR-1 w/o think	100.0	100.0	91.1	99.6	97.7

Table 3: Training LRMs or LLMs on safety data with or without the reasoning process (w/o think) on safety benchmarks.

filtering for alignment with benign intent, we obtain 915 clean samples. To assess its efficacy, we fine-tune R1-distilled models using varying subsets of these samples (*i.e.*, 100, 500, and all 915 samples) in addition to the original STAR-1 set. Detailed benchmark evaluation settings, data examples, and further methodology are provided in Sec. E.

As shown in Fig. 4, incorporating the crafted not_overrefusal data into the STAR-1 set significantly reduces overrefusal behavior, with an average increase on not_overrefusal rate from 68.9% to 78.1% across two models. Notably, this improvement comes with only a modest compromise in the average safety rate with a 3.7% decrease (from 94.0% to 90.3%). Moreover, we note the added data slightly enhances the models' reasoning ability, with an average gain from 56.3% to 57.2%. These findings support that our overrefusal mitigation strategy is successful and can meanwhile contributes positively to reasoning performance.

5 Related Work

LLM Safety Training. Standard safety training of LLMs uses supervised fine-tuning from human high-quality annotations to mitigate harmful outputs (Bianchi et al. 2023; Wei, Haghtalab, and Steinhardt 2023; Qi et al. 2023a; Raza et al. 2024). Beyond these methods, recent work focuses on aligning models' reasoning processes with explicit safety rules. Bai et al. (2022b) introduces a set of human-written principles and AI-driven self-critiques to fine-tune a harmless model without any human-labeled safety examples. OpenAI's Deliberative Alignment (Guan et al. 2025) trains models to explicitly reason through written safety policies before responding, achieving highly precise policy compliance and improved robustness against jailbreak prompts. Similarly, SafeChain (Jiang et al. 2025) fine-tunes models on a CoT-style safety dataset, improving refusal accuracy without impairing the reasoning performance.

High-quality LLM Training Data. Another line of research shows that small but high-quality datasets can significantly enhance LLM performance. LIMA (Zhou et al. 2023) fine-tunes a 65B model on 1K carefully curated examples yields results comparable to models trained on much more data. LIMO (Ye et al. 2025) achieves high mathematical reasoning performance with just 817 examples, outperforming models trained on 100x more data. Muennighoff et al. (2025) similarly distill a 59K reasoning corpus down to 1K examples in the s1 dataset. LIMR (Li, Zou, and Liu 2025) shows that a 1.4K carefully selected samples can outperform a full dataset of 8.5K samples in the LLM RL training. STAR-1 leverages both sides to advance the creation of robust, high-quality safety training data for LRMs.

6 Conlusion

In this work, we introduced STAR-1 — a high-quality, 1K-scale safety dataset specifically designed to enhance LRMs. Our extensive experiments across multiple model families and parameter scales demonstrate that fine-tuning with STAR-1 leads to significant safety improvements (up to an average of 40% enhancement on key benchmarks) with only a minimal compromise in reasoning performance. We hope that our work will inspire the community to further explore and address the safety challenges inherent in LRMs.

Ethical Statement

STAR-1 is developed to support safer and more robust reasoning in LMs. While STAR-1 aims to improve safety alignment of LMs, we acknowledge the sourced data may contain harmful, biased, or sensitive content. Misuse of aligned models is still possible, and we encourage responsible use of STAR-1 strictly for research into safety and alignment. The dataset and associated code are released for non-commercial research purposes.

Acknowledgments

This work is partially supported by a gift from Open Philanthropy. We thank the NAIRR Pilot Program and the Microsoft Accelerate Foundation Models Research Program for supporting our computing needs.

LLNL co-authors were supported under Contract DE-AC52-07NA27344 with the U.S. Department of Energy and the LLNL-LDRD Program under Project Numbers 24-ERD-058 and 24-ERD-010. The United States Government retains, and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes.

References

Anthropic. 2025. Anthropic Usage Policies. https://www.anthropic.com/legal/aup. Accessed: 2025-03-26.

Bai, Y.; Jones, A.; Ndousse, K.; Askell, A.; Chen, A.; DasSarma, N.; Drain, D.; Fort, S.; Ganguli, D.; Henighan, T.; et al. 2022a. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Bai, Y.; Kadavath, S.; Kundu, S.; Askell, A.; Kernion, J.; Jones, A.; Chen, A.; Goldie, A.; Mirhoseini, A.; McKinnon, C.; et al. 2022b. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Bekbayev, A.; Chun, S.; Dulat, Y.; and Yamazaki, J. 2023. The Poison of Alignment. *arXiv preprint arXiv:2308.13449*.

Bhardwaj, R.; and Poria, S. 2023. Red-teaming large language models using chain of utterances for safety-alignment. *arXiv preprint arXiv:2308.09662*.

Bianchi, F.; Suzgun, M.; Attanasio, G.; Röttger, P.; Jurafsky, D.; Hashimoto, T.; and Zou, J. 2023. Safety-tuned llamas: Lessons from improving the safety of large language models that follow instructions. *arXiv preprint arXiv:2309.07875*.

Chao, P.; Debenedetti, E.; Robey, A.; Andriushchenko, M.; Croce, F.; Sehwag, V.; Dobriban, E.; Flammarion, N.; Pappas, G. J.; Tramer, F.; et al. 2024. Jailbreakbench: An open robustness benchmark for jailbreaking large language models. *arXiv preprint arXiv:2404.01318*.

Chen, M.; Tworek, J.; Jun, H.; Yuan, Q.; Pinto, H. P. D. O.; Kaplan, J.; Edwards, H.; Burda, Y.; Joseph, N.; Brockman, G.; et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

Christen, P. 2011. A survey of indexing techniques for scalable record linkage and deduplication. *IEEE transactions on knowledge and data engineering*.

DeepSeek-AI; Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; and et al. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *arXiv preprint arXiv:2501.12948*.

Du, A.; Gao, B.; Xing, B.; Jiang, C.; Chen, C.; Li, C.; Xiao, C.; Du, C.; Liao, C.; et al. 2025. Kimi k1. 5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*.

French, R. M. 1999. Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences*.

Ganguli, D.; Lovitt, L.; Kernion, J.; Askell, A.; Bai, Y.; Kadavath, S.; Mann, B.; Perez, E.; Schiefer, N.; Ndousse, K.; et al. 2022. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858*.

Grattafiori, A.; Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Vaughan, A.; et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Guan, M. Y.; Joglekar, M.; Wallace, E.; Jain, S.; Barak, B.; Helyar, A.; Dias, R.; Vallone, A.; Ren, H.; Wei, J.; Chung, H. W.; Toyer, S.; Heidecke, J.; Beutel, A.; and Glaese, A. 2025. Deliberative Alignment: Reasoning Enables Safer Language Models. *arXiv preprint arXiv:2412.16339*.

Guo, Y.; Cui, G.; Yuan, L.; Ding, N.; Wang, J.; Chen, H.; Sun, B.; Xie, R.; Zhou, J.; Lin, Y.; et al. 2024. Controllable Preference Optimization: Toward Controllable Multi-Objective Alignment. *arXiv preprint arXiv:2402.19085*.

Huang, Y.; Gupta, S.; Xia, M.; Li, K.; and Chen, D. 2023. Catastrophic jailbreak of open-source llms via exploiting generation. *arXiv preprint arXiv:2310.06987*.

Jaech, A.; Kalai, A.; Lerer, A.; Richardson, A.; El-Kishky, A.; Low, A.; Helyar, A.; Madry, A.; Beutel, A.; Carney, A.; et al. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

Ji, J.; Hong, D.; Zhang, B.; Chen, B.; Dai, J.; Zheng, B.; Qiu, T.; Li, B.; and Yang, Y. 2024. Pku-saferlhf: A safety alignment preference dataset for llama family models. *arXiv e-prints*, arXiv–2406.

Ji, J.; Liu, M.; Dai, J.; Pan, X.; Zhang, C.; Bian, C.; Zhang, C.; Sun, R.; Wang, Y.; and Yang, Y. 2023. BeaverTails: Towards Improved Safety Alignment of LLM via a Human-Preference Dataset. *arXiv preprint arXiv:2307.04657*.

Jiang, F.; Xu, Z.; Li, Y.; Niu, L.; Xiang, Z.; Li, B.; Lin, B. Y.; and Poovendran, R. 2025. SafeChain: Safety of Language Models with Long Chain-of-Thought Reasoning Capabilities. *arXiv preprint arXiv:2502.12025*.

Jiang, L.; Rao, K.; Han, S.; Ettinger, A.; Brahman, F.; Kumar, S.; Mireshghallah, N.; Lu, X.; Sap, M.; Choi, Y.; and Dziri, N. 2024. WildTeaming at Scale: From In-the-Wild Jailbreaks to (Adversarially) Safer Language Models. *arXiv preprint arXiv:2406.18510*.

Kirkpatrick, J.; Pascanu, R.; Rabinowitz, N.; Veness, J.; Desjardins, G.; Rusu, A. A.; Milan, K.; Quan, J.; Ramalho, T.;

Grabska-Barwinska, A.; et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*.

Kour, G.; Zalmanovici, M.; Zwerdling, N.; Goldbraich, E.; Fandina, O. N.; Anaby-Tavor, A.; Raz, O.; and Farchi, E. 2023. Unveiling safety vulnerabilities of large language models. *arXiv preprint arXiv:2311.04124*.

Lee, T.; Tu, H.; Wong, C. H.; Wang, Z.; Yang, S.; Mai, Y.; Zhou, Y.; Xie, C.; and Liang, P. 2025. AHELM: A Holistic Evaluation of Audio-Language Models. *arXiv preprint arXiv:2508.21376*.

Li, L.; Dong, B.; Wang, R.; Hu, X.; Zuo, W.; Lin, D.; Qiao, Y.; and Shao, J. 2024. Salad-bench: A hierarchical and comprehensive safety benchmark for large language models. *arXiv preprint arXiv:2402.05044*.

Li, X.; Zou, H.; and Liu, P. 2025. Limr: Less is more for rl scaling. *arXiv preprint arXiv:2502.11886*.

Lightman, H.; Kosaraju, V.; Burda, Y.; Edwards, H.; Baker, B.; Lee, T.; Leike, J.; Schulman, J.; Sutskever, I.; and Cobbe, K. 2023. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*.

Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, 74–81.

Liu, X.; Xu, N.; Chen, M.; and Xiao, C. 2023. Autodan: Generating stealthy jailbreak prompts on aligned large language models. *arXiv preprint arXiv:2310.04451*.

Llama Team, A. . M. 2024. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

MAA. 2024. American invitational mathematics examination - aime. https://maa.org/math-competitions/american-invitational-mathematics-examination-aime. Accessed: 2025-03-26.

Mazeika, M.; Phan, L.; Yin, X.; Zou, A.; Wang, Z.; Mu, N.; Sakhaee, E.; Li, N.; Basart, S.; Li, B.; et al. 2024. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal. *arXiv preprint arXiv:2402.04249*.

Mazeika, M.; Zou, A.; Mu, N.; Phan, L.; Wang, Z.; Yu, C.; Khoja, A.; Jiang, F.; O'Gara, A.; Sakhaee, E.; Xiang, Z.; Rajabi, A.; Hendrycks, D.; Poovendran, R.; Li, B.; and Forsyth, D. 2023. TDC 2023 (LLM Edition): The Trojan Detection Challenge. In *NeurIPS Competition Track*.

MetaAI. 2024. MetaAI Usage Policies. https://transparency.meta.com/policies/. Accessed: 2025-03-26.

Muennighoff, N.; Yang, Z.; Shi, W.; Li, X. L.; Fei-Fei, L.; Hajishirzi, H.; Zettlemoyer, L.; Liang, P.; Candès, E.; and Hashimoto, T. 2025. s1: Simple test-time scaling. *arXiv preprint arXiv:2501.19393*.

OpenAI. 2024. Upgrading the Moderation API with our new multimodal moderation model. https://openai.com/index/upgrading-the-moderation-api-with-our-new-multimodal-moderation-model/. Accessed: 2025-03-26.

OpenAI. 2025a. OpenAI Model Specifications. https://model-spec.openai.com/2025-04-11.html#prohibited_content. Accessed: 2025-03-26.

OpenAI. 2025b. OpenAI Simple Evals Framework. https://github.com/openai/simple-evals. Accessed: 2025-03-26.

OpenAI. 2025c. OpenAI Usage Policies. https://openai.com/policies/usage-policies/. Accessed: 2025-03-26.

Qi, X.; Zeng, Y.; Xie, T.; Chen, P.-Y.; Jia, R.; Mittal, P.; and Henderson, P. 2023a. Fine-tuning aligned language models compromises safety, even when users do not intend to! *arXiv preprint arXiv:2310.03693*.

Qi, X.; Zeng, Y.; Xie, T.; Chen, P.-Y.; Jia, R.; Mittal, P.; and Henderson, P. 2023b. Hex-phi: Human-extended policy-oriented harmful instruction benchmark.

Rajbhandari, S.; Rasley, J.; Ruwase, O.; and He, Y. 2020. Zero: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*.

Raza, S.; Bamgbose, O.; Ghuge, S.; Tavakoli, F.; and Reji, D. J. 2024. Developing Safe and Responsible Large Language Models–A Comprehensive Framework. *arXiv e-prints*.

Reimers, N.; and Gurevych, I. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Rein, D.; Hou, B. L.; Stickland, A. C.; Petty, J.; Pang, R. Y.; Dirani, J.; Michael, J.; and Bowman, S. R. 2024. Gpqa: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*.

Röttger, P.; Kirk, H. R.; Vidgen, B.; Attanasio, G.; Bianchi, F.; and Hovy, D. 2023. Xstest: A test suite for identifying exaggerated safety behaviours in large language models. *arXiv preprint arXiv:2308.01263*.

Souly, A.; Lu, Q.; Bowen, D.; Trinh, T.; Hsieh, E.; Pandey, S.; Abbeel, P.; Svegliato, J.; Emmons, S.; Watkins, O.; et al. 2024. A strongreject for empty jailbreaks. *arXiv preprint arXiv:2402.10260*.

Tedeschi, S.; Friedrich, F.; Schramowski, P.; Kersting, K.; Navigli, R.; Nguyen, H.; and Li, B. 2024. ALERT: A Comprehensive Benchmark for Assessing Large Language Models' Safety through Red Teaming. *arXiv preprint arXiv:2404.08676*.

Thakkar, M.; Fournier, Q.; Riemer, M. D.; Chen, P.-Y.; Zouaq, A.; Das, P.; and Chandar, S. 2024. A Deep Dive into the Trade-Offs of Parameter-Efficient Preference Alignment Techniques. *arXiv preprint arXiv:2406.04879*.

Tu, H.; Cui, C.; Wang, Z.; Zhou, Y.; Zhao, B.; Han, J.; Zhou, W.; Yao, H.; and Xie, C. 2023. How Many Unicorns Are in This Image? A Safety Evaluation Benchmark for Vision LLMs. *arXiv preprint arXiv:2311.16101*.

Vidgen, B.; Scherrer, N.; Kirk, H. R.; Qian, R.; Kannappan, A.; Hale, S. A.; and Röttger, P. 2023. Simplesafetytests: a test suite for identifying critical safety risks in large language models. *arXiv preprint arXiv:2311.08370*.

Wang, J.; Lan, C.; Liu, C.; Ouyang, Y.; Qin, T.; Lu, W.; Chen, Y.; Zeng, W.; and Yu, P. S. 2022. Generalizing to

Unseen Domains: A Survey on Domain Generalization. *arXiv preprint arXiv:2103.03097*.

Wang, Y.; Li, H.; Han, X.; Nakov, P.; and Baldwin, T. 2023. Do-not-answer: A dataset for evaluating safeguards in llms. *arXiv preprint arXiv:2308.13387*.

Wang, Y.; Ma, X.; Zhang, G.; Ni, Y.; Chandra, A.; Guo, S.; Ren, W.; Arulraj, A.; He, X.; Jiang, Z.; et al. 2024a. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Wang, Z.; Tu, H.; Mei, J.; Zhao, B.; Wang, Y.; and Xie, C. 2024b. AttnGCG: Enhancing Jailbreaking Attacks on LLMs with Attention Manipulation. *arXiv preprint arXiv:2410.09040*.

Wei, A.; Haghtalab, N.; and Steinhardt, J. 2023. Jailbroken: How does llm safety training fail? *NeurIPS*.

Xie, Y.; Wu, J.; Tu, H.; Yang, S.; Zhao, B.; Zong, Y.; Jin, Q.; Xie, C.; and Zhou, Y. 2024. A Preliminary Study of o1 in Medicine: Are We Closer to an AI Doctor? *arXiv preprint arXiv:2409.15277*.

Yang, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Li, C.; Liu, D.; Huang, F.; Wei, H.; et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Ye, Y.; Huang, Z.; Xiao, Y.; Chern, E.; Xia, S.; and Liu, P. 2025. LIMO: Less is More for Reasoning. *arXiv preprint arXiv:2502.03387*.

Yu, J.; Lin, X.; Yu, Z.; and Xing, X. 2023. Gptfuzzer: Red teaming large language models with auto-generated jailbreak prompts. *arXiv preprint arXiv:2309.10253*.

Zhang, D.; Wang, J.; and Charton, F. 2024. Only-IF: Revealing the Decisive Effect of Instruction Diversity on Generalization. *arXiv preprint arXiv:2410.04717*.

Zhao, W.; Ren, X.; Hessel, J.; Cardie, C.; Choi, Y.; and Deng, Y. 2024. Wildchat: 1m chatgpt interaction logs in the wild. *arXiv preprint arXiv:2405.01470*.

Zheng, L.; Chiang, W.-L.; Sheng, Y.; Li, T.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Li, Z.; Lin, Z.; Xing, E. P.; et al. 2023. Lmsys-chat-1m: A large-scale real-world llm conversation dataset. *arXiv preprint arXiv:2309.11998*.

Zhou, C.; Liu, P.; Xu, P.; Iyer, S.; Sun, J.; Mao, Y.; Ma, X.; Efrat, A.; Yu, P.; Yu, L.; et al. 2023. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*.

Zhou, K.; Liu, C.; Zhao, X.; Jangam, S.; Srinivasa, J.; Liu, G.; Song, D.; and Wang, X. E. 2025. The Hidden Risks of Large Reasoning Models: A Safety Assessment of R1. *arXiv preprint arXiv:2502.12659*.

Zou, A.; Wang, Z.; Carlini, N.; Nasr, M.; Kolter, J. Z.; and Fredrikson, M. 2023. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*.