EASE: Practical and Efficient Safety Alignment for Small Language Models

Haonan Shi, Guoli Wang, Tu Ouyang, An Wang

Case Western Reserve University
{haonan.shi3, guoli.wang, tu.ouyang, an.wang}@case.edu

Abstract

Small language models (SLMs) are increasingly deployed on edge devices, making their safety alignment crucial yet challenging. Current shallow alignment methods that rely on direct refusal of malicious queries fail to provide robust protection, particularly against adversarial jailbreaks. While deliberative safety reasoning alignment offers deeper alignment for defending against sophisticated attacks, effectively implanting such reasoning capability in SLMs with limited capabilities remains an open challenge. Moreover, safety reasoning incurs significant computational overhead as models apply reasoning to nearly all queries, making it impractical for resource-constrained edge deployment scenarios that demand rapid responses. We propose **EASE**, a novel framework that enables practical and **E**fficient safety **A**lignment for **S**mall languag**E** models. Our approach first identifies the optimal safety reasoning teacher that can effectively distill safety reasoning capabilities to SLMs. We then align models to selectively activate safety reasoning for dangerous adversarial jailbreak queries while providing direct responses to straightforward malicious queries and general helpful tasks. This selective mechanism enables small models to maintain robust safety guarantees against sophisticated attacks while preserving computational efficiency for benign interactions. Experimental results demonstrate that EASE reduces jailbreak attack success rates by up to 17% compared to shallow alignment methods while reducing inference overhead by up to 90% compared to deliberative safety reasoning alignment, making it practical for SLMs real-world edge deployments.

Introduction

Small language models (SLMs) (Qwen et al. 2025; Llama Team 2024; Javaheripi et al. 2023; Liu et al. 2024; Zhang et al. 2024) have demonstrated remarkable performance improvements, achieving comparable results to large language models (LLMs) across various tasks, including conversational AI (Gunter et al. 2024), code generation (Javaheripi et al. 2023), sentiment analysis, and domain-specific text processing while maintaining significant deployment advantages. Their compact architectures enable efficient deployment on resource-constrained environments, including mobile devices (Gunter et al. 2024) and edge computing platforms (KHIABANI et al. 2024), offering reduced computational overhead and enhanced privacy through local processing. These practical benefits have driven widespread adoption of SLMs across diverse real-world applications. However, as SLMs become increasingly prevalent in real-world deployments, ensuring their safety has emerged as a critical challenge that demands immediate attention.

Recent research shows that SLMs are more vulnerable to jailbreak attacks than LLMs (Yi et al. 2025; Zhang et al. 2025a). Some SLMs even fail to resist direct harmful queries (Yi et al. 2025). This highlights the critical need for better SLM safety measures. Safety alignment serves as a primary approach to improving model safety and has predominantly relied on refusal training methods (Llama Team 2024). These approaches employ training techniques such as supervised fine-tuning (SFT) (Liu et al. 2023; Taori et al. 2023) and preference-based optimization including Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al. 2022; Bai et al. 2022) and Direct Preference Optimization (DPO) (Rafailov et al. 2023; Liu, Sun, and Zheng 2024) to train models to directly refuse harmful queries. However, refusal training often leads to performance degradation on general tasks, which is particularly problematic for SLMs given their already limited model capacity, making extensive refusal training impractical. Moreover, refusal training constitutes shallow alignment where models rely on intuitive rejection of harmful queries without deeper reasoning. This approach exhibits poor generalization capabilities and fails to defend against diverse and sophisticated jailbreak attack strategies.

Chain-of-Thought reasoning has been proven to enable LLMs to solve complex problems more effectively. Research demonstrates that reasoning capabilities allow models to address problems that would otherwise require significantly larger architectures or remain unsolvable when generating direct answers (Li et al. 2024b). Building on this insight, several studies (Zhang et al. 2025c; Guan et al. 2024; Wang et al. 2025b; Zhang et al. 2025d; Mou et al. 2025) have developed deliberative alignment methods that incorporate safety reasoning for LLMs, successfully addressing the limited generalization issues of shallow alignment approaches. This approach enables LLMs to learn how to conduct safety reasoning analysis on queries to determine whether to respond or refuse. Models trained through deliberative alignment demonstrate enhanced safety robustness and can de-

velop deeper understanding of diverse harmful queries and sophisticated jailbreak attacks, enabling more reliable safety performance. However, directly applying deliberative alignment methods to SLMs presents two main challenges: (1) Given the limited capacity and capabilities of small models, how can they effectively learn safety reasoning abilities? (2) Deliberative alignment causes models to perform reasoning on all queries after safety alignment, generating more tokens in outputs and resulting in increased inference time costs. For SLMs, which are typically deployed in resource-constrained edge environments with requirements for rapid response (KHIABANI et al. 2024), the excessive inference overhead introduced by this alignment approach poses a significant deployment challenge.

To address these challenges, we propose **EASE**, a practical and **E**fficient safety **A**lignment framework for **S**mall Languag**E** models. EASE enables SLM to selectively apply safety reasoning only to adversarial jailbreak queries where shallow alignment fails to achieve adequate generalization, while providing direct responses to simple harmful queries and general tasks to reduce inference time overhead and preserve the low-cost and rapid-response characteristics of SLMs. EASE consists of two phases. In the first phase, to enable SLMs to acquire comprehensive safety reasoning knowledge, we employ knowledge distillation to leverage a more capable teacher model to implant safety reasoning capabilities into small models. In the second phase, we calibrate the safety reasoning boundaries of SLMs to enable adaptive safety reasoning activation. We identify vulnerable semantic regions where SLMs exhibit poor safety generalization and construct targeted training data comprising safety reasoning examples for queries in these vulnerable semantic regions and direct response examples for benign queries and direct harmful queries. This enables SLMs to activate safety reasoning only for adversarial queries in vulnerable semantic regions while providing direct responses to other queries to maintain efficiency.

Our main contributions are summarized as follows: (1) We propose EASE, a two-phase safety alignment framework that combines safety reasoning capability implantation with reasoning boundary calibration to achieve both enhanced safety and computational efficiency for SLMs. (2) We investigate optimal teacher model selection for safety reasoning knowledge distillation, revealing that Large Reasoning Models with smaller capability gaps are more effective teachers than conventional LLMs. (3) We develop a selective reasoning activation mechanism that enables SLMs to apply safety reasoning only to adversarial jailbreak queries in vulnerable semantic regions while maintaining direct responses for benign queries.

Related Works

Safety of Language Models To prevent LLMs from generating harmful content when faced with malicious queries, numerous safety alignment methods have been developed to align LLMs with safety requirements. Existing approaches primarily include SFT (Liu et al. 2023; Taori et al. 2023), DPO (Rafailov et al. 2023; Liu, Sun, and Zheng 2024) and RLHF (Ouyang et al. 2022; Bai et al. 2022). SFT finetunes models on curated datasets of safe responses to harmful prompts. Both RLHF and DPO leverage human preference data to align model outputs with human values. However, these alignment methods may lead to shallow alignment (Qi et al. 2025), causing models to primarily reject directly malicious queries while failing to deeply understand and refuse adversarial malicious queries, rendering these methods vulnerable to adversarial jailbreak attacks. To address this limitation, some advanced defensive methods such as machine unlearning (Liu et al. 2025a), representation engineering (Zou et al. 2024), and safeguard model (Ji et al. 2024; Liu et al. 2025b) approaches can further defend against adversarial jailbreak attacks. However, they often require additional external components, posing challenges for practical model deployment. Recent works (Zhang et al. 2025c,d) such as deliberative alignment (Guan et al. 2024) has demonstrated that enabling LLMs to perform safety reasoning on queries can enhance the model's deep understanding of malicious queries, thereby improving robustness against adversarial malicious queries. However, this approach leads to performing safety reasoning on all queries, which introduces computational overhead in inference time.

LLM Reasoning Reasoning capabilities enable LLMs to achieve stronger performance across numerous tasks. LLMs can currently acquire reasoning capabilities through two primary approaches. The first approach involves supervised learning on synthesized data, where methods include human annotation (Lightman et al. 2023), Monte Carlo Tree Search (MCTS) (Vodopivec, Samothrakis, and Ster 2017; Xie et al. 2024), and knowledge distillation (Huang et al. 2024) from more powerful LLMs/Large Reasoning Models (LRM). The second approach leverages reinforcement learning to enhance reasoning capabilities. Recent examples include OpenAI's o-series models (Jaech et al. 2024) and DeepSeek-R1 (DeepSeek-AI 2025), which employ reinforcement learning techniques to enhance step-by-step reasoning processes. For small models specifically, recent work from the DeepSeek-Distill series (DeepSeek-AI 2025) has demonstrated that small models can effectively obtain excellent reasoning performance from more powerful models through knowledge distillation. In our work, we investigate how small models can acquire better safety reasoning performance from larger models through knowledge distillation.

Methodology

We present EASE, a two-phase safety alignment framework designed specifically for SLMs, as illustrated in Figure 1. Our approach addresses two critical challenges: achieving more robust safety alignment performance while maintaining practical inference efficiency.

Notation: We denote the student model as M_s, safety reasoning teacher model as M_t, and safety reasoning implanted student model after Phase 1 as M_{reason}. In Phase 1, we use $\mathcal{D}_{\text{train}}$ for knowledge distillation training. In Phase 2, we employ $\mathcal{D}_{\text{diag}}$ to identify vulnerable jailbreak queries, then construct $\mathcal{D}_{\text{reason}}$ containing reasoning traces for vulnerable jailbreak queries and $\mathcal{D}_{\text{direct}}$ containing direct responses for

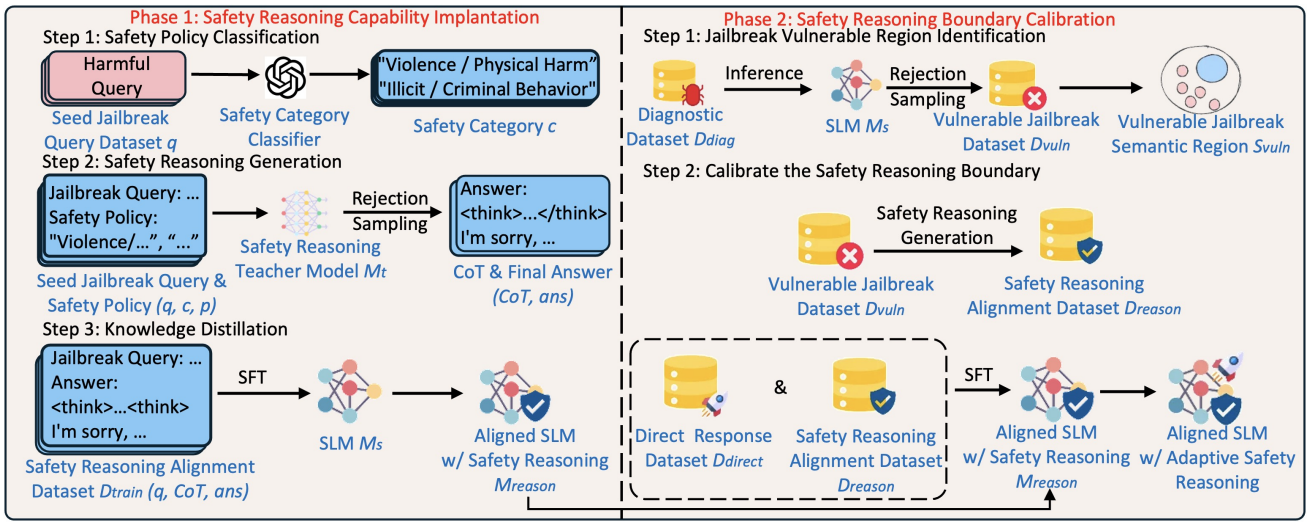

Figure 1: The workflow of our safety alignment method for small language models (EASE).

benign general tasks and straightforward jailbreak queries without reasoning traces to calibrate the safety reasoning boundary. The function f_{judge} is used for rejection sampling to filter training data.

Safety Reasoning Capability Implantation

To effectively and comprehensively implant safety reasoning capabilities into SLMs, we first teach the SLMs to learn how to utilize diverse safety policy knowledge and conduct safety reasoning analysis on jailbreak queries across multiple safety categories. We employ knowledge distillation with safety policy-guided deliberative reasoning to enable a teacher model to transfer safety reasoning capabilities to SLMs. The phase 1 of EASE consists of three key steps:

Safety Policy Classification We utilize existing category-specific safety policies from prior work (Guan et al. 2024; Wang et al. 2025b), where each policy defines clear policy objectives and response rules for handling requests within that safety category. For each safety category $c \in \mathcal{C}$, we employ the corresponding policy $p_c \in \mathcal{P}$. We then categorize seed jailbreak queries using a LLM classifier to assign each seed jailbreak query q in seed jailbreak query dataset to its corresponding category c.

Safety Reasoning Generation We combine seed jailbreak queries with their classified categories and corresponding policies to create triplets: (q, c, p_c). We employ a capable large teacher model M_t to generate deliberative CoT traces:

$$(CoT, ans) = M_t(q, c, p_c) \tag{1}$$

, where *CoT* represents the reasoning trace and *ans* is the final decision. Crucially, we apply context distillation (Snell, Klein, and Zhong 2022; Askell et al. 2021) by providing the full context but only retaining the seed jailbreak queries and generated reasoning in the knowledge distillation data: $\mathcal{D}_{\text{distilled}} = \{(q, CoT, ans)\}$. Subsequently, we perform rejection sampling using LLaMA-Guard-3-8B as safety judge:

$$f_{\text{judge}} : \mathcal{R} \rightarrow \{0, 1\} \tag{2}$$

to evaluate whether the generated responses are harmful, where $f_{\text{judge}}(ans) = 1$ indicates harmful output. By filtering out samples with harmful responses, we obtain our training dataset consisting of safe refusals:

$$\mathcal{D}_{\text{train}} = \{(q, CoT, ans) \in \mathcal{D}_{\text{distilled}} : f_{\text{judge}}(ans) = 0\} \tag{3}$$

Knowledge Distillation We fine-tune the shallow aligned SLM M_s using the context-distilled training samples by minimizing the cross-entropy loss:

$$\mathcal{L}_{CE} = -\mathbb{E}_{(q,y)\sim\mathcal{D}_{\text{train}}}\big[\log P_{M_s}(y \mid q)\big], \quad y = (CoT, ans). \tag{4}$$

Enabling the model to learn safety reasoning patterns without requiring explicit policy access during inference.

However, the critical question for effective safety reasoning distillation to SLMs is: **what type of language models serves as the most effective teacher?** Given the limited capabilities of SLMs, simply using the most powerful available model as a teacher may not yield optimal results. We investigate how different teacher model characteristics – including the model sizes and reasoning ability – impact the distillation effectiveness for safety reasoning tasks.

To identify optimal teacher models for safety reasoning distillation, we conduct systematic comparisons across different teacher types and scales using Qwen2.5-1.5B-Instruct as student SLM. We sample 10K seed jailbreak queries from the STAR-41K dataset (Wang et al. 2025b) for safety reasoning distillation and evaluate on 250 randomly selected prompts from the test dataset of WildJailbreak (Jiang et al. 2024) using Llama-Guard-3-8B. Our investigation reveals two key insights regarding teacher model selection. While existing work on safety reasoning for alignment has employed both LLMs (Mou et al. 2025; Zhang et al. 2025c,b; Wang et al. 2025a) and LRMs (Guan et al. 2024; Wang et al. 2025b) as teachers, there has been limited systematic investigation into which type of model serves as a more effective teacher for SLMs.

We find that LRMs significantly outperform LLMs as teachers for safety reasoning distillation, as demonstrated

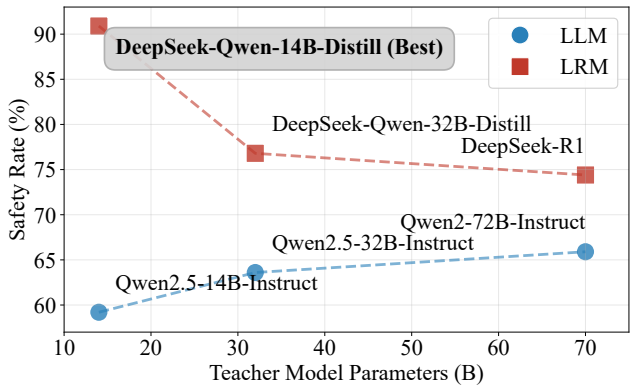

Figure 2: LRMs outperform LLMs as safety reasoning teachers, with smaller LRMs showing better distillation effectiveness due to reduced capability gaps with SLMs.

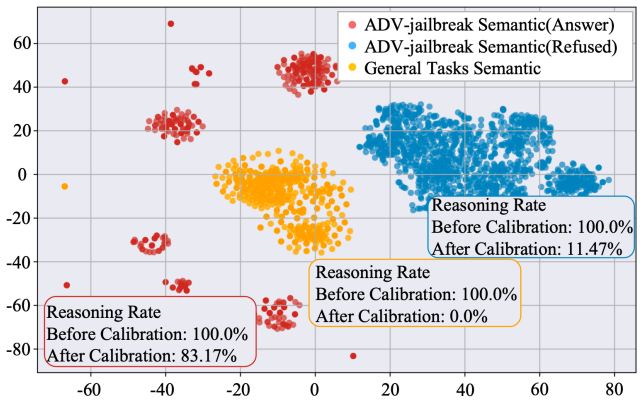

Figure 3: Model M_s intermediate layer activations reveal semantic clustering by query type through t-SNE visualization. Red clusters correspond to multiple vulnerable jailbreak tactics semantic regions, blue points represent refused adversarial queries, and yellow points show benign queries. Post-calibration reasoning rates demonstrate adaptive safety reasoning activation across different semantic regions, confirming effective boundary calibration.

in Figure 2. This superiority stems from LRMs' enhanced reasoning capabilities, which enable them to generate more structured and transferable safety reasoning patterns. Even when we employ specifically designed prompts (Mou et al. 2025; Zhang et al. 2025d,c,b; Wang et al. 2025a) to guide LLMs to produce CoT formatted responses for harmful queries (e.g., *"You first think about the reasoning process as an internal monologue and then provide the user with the answer...Respond in the following CoT format: <think>...</think>[Final Answer]"*), their performance remains substantially inferior to LRMs in teaching effective safety reasoning to small models.

We observe that smaller LRMs demonstrate superior safety reasoning distillation effectiveness compared to larger LRMs. This counterintuitive finding likely reflects the capacity gap between student and teacher models - SLMs struggle to learn safety reasoning from overly powerful teachers whose reasoning complexity exceeds their learning capacity. Similar phenomena have been observed in other reasoning tasks (Li et al. 2025). However, teacher model capability cannot be reduced arbitrarily. When we experiment with DeepSeek-Qwen-7B-Distill (DeepSeek-AI 2025) as a teacher model, the model's limited capabilities result in degraded safety reasoning quality, occasionally producing harmful responses to malicious queries. Even using the rejection sampling to filter out such harmful training examples, this degradation in teacher quality significantly impairs the safety reasoning distillation performance for SLMs, indicating a critical threshold for minimum teacher competence.

Safety Reasoning Boundary Calibration

Jailbreak Vulnerable Region Identification Following the safety reasoning distillation phase, we observe that SLMs apply safety reasoning processes broadly across diverse query types, including benign general tasks and straightforward harmful queries that do not require safety deliberation. This comprehensive reasoning approach creates significant efficiency challenges for SLM deployment scenarios where computational resources are limited and real-time responses are critical. We find that shallow aligned SLMs are already capable of refusing to answer easy-to-align jailbreak queries, but struggle with hard-to-align adversarial jailbreak queries that are particularly resistant to conventional shallow safety alignment techniques. Thus, we only need SLMs to perform safety reasoning on hard-to-align adversarial jailbreak queries to enhance their safety robustness. Therefore, we still need to further adjust the M_{reason} obtained from safety reasoning capability implantation phase of our safety alignment process.

To identify which jailbreak data are hard-to-align for the model when only using shallow alignment method, we first need to determine which jailbreak tactics are difficult for shallow-aligned models to recognize and refuse. Jiang *et al.* (Jiang et al. 2024) demonstrated that adversarial jailbreak behaviors, automatically mined from real-world user–chatbot interactions (Zhao et al. 2024; Zheng et al. 2024), comprise 105,438 tactic instances, which can be semantically grouped into 5,688 distinct semantic regions $\mathcal{S}_{\text{adv}}$ (e.g., fictitious scenario, assign personality, and code by pseudonym). We can leverage a dataset containing data that employ these jailbreak tactics to test which jailbreak tactics pose vulnerabilities for shallow aligned SLM M_s.

To operationalize this idea, we use the adversarial harmful subset of the WildJailbreak training dataset (Jiang et al. 2024), which contains adversarially rewritten queries generated via combinations of real-world jailbreak tactics. This subset serves as a diagnostic dataset $\mathcal{D}_{\text{diag}}$ to evaluate the failure modes of shallow aligned SLM M_s. Using the same safety judge model f_{judge} described earlier, we identify adversarial prompts where M_s fails to refuse harmful requests:

$$\mathcal{D}_{\text{vuln}} = \{q \in \mathcal{D}_{\text{diag}} : f_{\text{judge}}(M_s(q)) = 1\} \quad (5)$$

We can systematically identify which jailbreak tactics M_s is vulnerable to based on the tactics present in these non-

refused responses. As Figure 3 shown, these vulnerable data reveal the vulnerable semantic regions $\mathcal{S}_{\text{vuln}} \subset \mathcal{S}_{\text{adv}}$ of the jailbreak space where M_{s} demonstrates insensitivity and where shallow alignment fails to establish robust safeguards.

Calibrate the Safety Reasoning Boundary Having identified the vulnerable semantic regions, we now focus on training M_{reason} to selectively apply safety reasoning in these problematic areas. Our objective is to enable the model to activate safety reasoning when encountering semantic patterns characteristic of $\mathcal{S}_{\text{vuln}}$, while maintaining efficient direct responses for benign queries.

For queries in $\mathcal{S}_{\text{vuln}}$, we construct reasoning data $\mathcal{D}_{\text{reason}}$ using $\mathcal{D}_{\text{vuln}}$ in the form $(q, \textit{CoT}, \textit{ans})$, where reasoning traces and final answers are generated through safety reasoning from the teacher model M_t. Simultaneously, to enable the model to learn when direct responses are appropriate, we construct a direct-response dataset $\mathcal{D}_{\text{direct}}$ comprising two components: (1) vanilla harmful queries without jailbreak tactics that M_{s} can already refuse correctly, and (2) general task queries with benign intent. Each example in $\mathcal{D}_{\text{direct}}$ follows the format $(q, \textit{ans})$, where *ans* represents either a refusal or direct answer obtained from LLMs without intermediate reasoning processes.

We combine $\mathcal{D}_{\text{reason}}$ and $\mathcal{D}_{\text{direct}}$ to form our calibration dataset for safety reasoning boundary calibration. Subsequently, we perform SFT on this combined dataset for M_{reason} by minimizing the cross-entropy loss:

$$\mathcal{L}_{\text{calibration}} = -\,\mathbb{E}_{(q,y)\sim\mathcal{D}_{\text{reason}}\cup\mathcal{D}_{\text{direct}}}\big[\log P_{M_{\text{reason}}}(y \mid q)\big], \quad (6)$$

where $y = (\textit{CoT}, \textit{ans})$ for $(q, y) \in \mathcal{D}_{\text{reason}}$ and $y = \textit{ans}$ for $(q, y) \in \mathcal{D}_{\text{direct}}$. This supervised fine-tuning objective enables the model to recognize semantic patterns across different query types and determine whether safety reasoning is necessary or direct output suffices.

Evaluations

Experiment Setup

Models and Datasets We select three base SLMs for safety alignment: Qwen2.5-1.5B-Instruct, Qwen2.5-3B-Instruct, and Llama3.2-3B-Instruct (Llama Team 2024). Additionally, we employ DeepSeek-Qwen-14B-Distill as the safety reasoning teacher model in the safety alignment process. Regarding the datasets utilized in EASE: In the Safety Reasoning Capability Implantation phase, we employ 10k data samples from STAR-41K as seed jailbreak queries. During the Safety Reasoning Boundary Calibration phase, we utilize 10k adversarial harmful data samples from the training dataset of WildJailbreak as the diagnostic dataset. Through rejection sampling, we select 1,500 vulnerable samples to generate the safety reasoning dataset for Phase 2. Furthermore, we incorporate 1,750 general task data samples from the UltraFeedback dataset (Cui et al. 2024) and 1,000 vanilla harmful query data samples from STAR-41K as the direct response dataset.

Baselines We compare EASE with two other safety alignment approaches. First is the refusal training method, where we employed all the harmful seed jailbreak data and direct final answers used in EASE as the alignment dataset for model safety alignment. The second method is the current state-of-the-art deliberative alignment (Guan et al. 2024) approach proposed by Guan *et al.*, which also leverages safety reasoning for model safety alignment.

Evaluation To evaluate the effectiveness of different safety alignment methods, we employ four established benchmarks: StrongReject (Souly et al. 2024), WildJailbreak test set (Jiang et al. 2024), Do-Anything-Now (DAN) (Shen et al. 2024), and WildChat (Zhao et al. 2024). StrongReject contains 313 forbidden prompts across six harmful categories, WildJailbreak provides 2,000 adversarial queries for sophisticated attack evaluation, DAN comprises 1,405 real-world jailbreak prompts from online platforms, and WildChat offers 370 malicious queries selected from real user-ChatGPT conversations. To further test the robustness of our safety alignment method against adversarial jailbreaks, following previous works (Zhang et al. 2025c,d), we leverage state-of-the-art jailbreak methods PAIR (Chao et al. 2025), PAP (Zeng et al. 2024) and Human-Jailbreaks (Li et al. 2024a) for evaluation on AdvBench (Zou et al. 2023). Across all evaluation tasks, we utilize Llama-Guard-3-8B to measure attack success rates (ASR).

To measure the impact of different safety alignment methods on helpful performance and generation efficiency in general tasks, we evaluate models on MMLU (Hendrycks et al. 2021) for knowledge assessment, HellaSwag (Zellers et al. 2019) for commonsense reasoning, and GSM8K (Cobbe et al. 2021) for mathematical problem solving using the lm-evaluation-harness framework (Gao et al. 2024).

Main Results

Safety Performance Evaluation Safety reasoning significantly enhances the safety performance of SLMs and enables better generalization of safety capabilities across diverse datasets and jailbreak tactics through reasoning mechanisms. Table 1 presents the safety performance of SLMs across different safety datasets following various safety alignment approaches. Given identical seed jailbreak prompts in the alignment datasets, we observe that both Deliberative alignment and our proposed EASE method substantially outperform Refusal Training in terms of safety performance. This superiority is maintained even when confronted with datasets containing diverse jailbreak techniques, such as WildJailbreak and DAN, where both reasoning-based methods achieve notably better safety performance. Specifically, the ASR of reasoning-based approaches are approximately 50% of those achieved by Refusal Training, demonstrating their enhanced robustness against sophisticated adversarial attacks.

Among these, EASE achieves better safety performance compared to Deliberative Alignment. This improvement stems from our targeted approach to safety reasoning alignment for SLMs, where we select a smaller and more suitable safety reasoning teacher, Deepseek-Qwen-14B-Distill. The reduced capability gap between the teacher model and SLMs facilitates more effective knowledge transfer, enabling SLMs to better acquire safety reasoning capabilities.

Model	Method	Safety (ASR ↓)				General (ACC ↑)		
		StrongReject	WildJailbreak	DAN	WildChat	MMLU	Hellaswag	GSM8K
Qwen2.5-1.5B -Instruct	Instruct	4.79%	42.79%	8.41%	32.16%	**58.08%**	**60.28%**	63.84%
	Refusal Training	1.60%	23.26%	6.10%	17.30%	57.50%	58.29%	63.00%
	Deliberative Alignment	0.96%	11.70%	**2.15%**	12.16%	57.44%	59.32%	62.70%
	EASE (Ours)	**0.96%**	**6.90%**	3.45%	**11.89%**	57.71%	59.40%	**64.29%**
Qwen2.5-3B -Instruct	Instruct	2.24%	46.70%	13.06%	29.19%	**64.60%**	70.20%	**75.97%**
	Refusal Training	1.60%	18.50%	7.01%	19.46%	61.91%	68.30%	74.22%
	Deliberative Alignment	0.32%	11.20%	3.16%	12.97%	63.51%	69.60%	75.16%
	EASE (Ours)	**0.32%**	**5.35%**	**2.55%**	**12.16%**	63.95%	**70.40%**	75.59%
Llama3.2-3B -Instruct	Instruct	4.79%	32.50%	12.22%	40.27%	**58.37%**	**66.70%**	**70.19%**
	Refusal Training	0.68%	12.73%	9.41%	26.12%	56.86%	63.21%	66.22%
	Deliberative Alignment	**0.00%**	11.10%	5.56%	21.89%	57.02%	65.85%	69.91%
	EASE (Ours)	0.32%	**3.35%**	**4.35%**	**8.38%**	57.24%	66.17%	69.83%

Table 1: Comparison of safety and general task performance. For safety tasks, ASR denotes attack success rate (%); for general tasks, ACC denotes accuracy (%). **Bold** indicates the best performance, underline indicates the second-best performance.

EASE Robustness Analysis We further examine EASE's specific robustness against advanced adversarial attacks. Table 2 demonstrates EASE's performance on AdvBench under various sophisticated attack methods. EASE shows strong resilience, maintaining low ASRs of 0-6% across different attacks, compared to the original models' ASRs of 22-44%. This analysis validates EASE's effectiveness in defending against state-of-the-art jailbreak techniques.

Models	None	HumanJB	PAP	PAIR
Qwen2.5-1.5B-Instruct	0.0%	22.0%	28.0%	40.0%
Qwen2.5-1.5B-EASE(Ours)	0.0%	6.0%	0.0%	2.0%
Qwen2.5-3B-Instruct	0.0%	44.0%	36.0%	42.0%
Qwen2.5-3B-EASE(Ours)	0.0%	0.0%	2.0%	2.0%
Llama3.2-3B-Instruct	4.0%	12.0%	40.0%	36.0%
Llama3.2-3B-EASE(Ours)	0.0%	2.0%	4.0%	0.0%

Table 2: Performance(ASR) comparison against different attack scenarios on AdvBench dataset.

Impact on General Task Capabilities As shown in Table 1, our safety alignment method EASE demonstrates the ability to maintain strong performance on general tasks while simultaneously improving safety capabilities in SLMs. The general task performance of SLMs tends to experience substantial degradation following alignment with Refusal Training methods. In contrast, our approach achieves a much better trade-off between safety enhancement and general capability preservation.

Computational Efficiency Analysis Our method also strikes a better balance between safety and efficiency compared to other safety alignment approaches, making it particularly well-suited for SLM deployment. As shown in Table 3, we observe that the original Instruct versions exhibit higher token generation on more sophisticated jailbreak datasets due to insufficient safety mechanisms that allow them to respond to numerous adversarial queries. While deliberative alignment achieves strong safety performance, it suffers from a critical efficiency limitation: the method requires reasoning for every query, regardless of whether it is benign or adversarial. This universal reasoning requirement significantly increases response lengths. For example, HellaSwag responses increase by an average of 298 tokens, substantially degrading inference efficiency. EASE addresses this limitation through selective reasoning activation. Our method triggers safety reasoning only when adversarial jailbreak queries are detected, allowing the model to respond efficiently to general tasks (e.g., Hellaswag) and simple harmful queries (e.g., StrongReject) with token counts similar to the original model. This targeted approach enables EASE to maintain both safety guarantees and general capability while achieving response efficiency comparable to original models – a significant improvement over deliberative alignment's consistently extended responses.

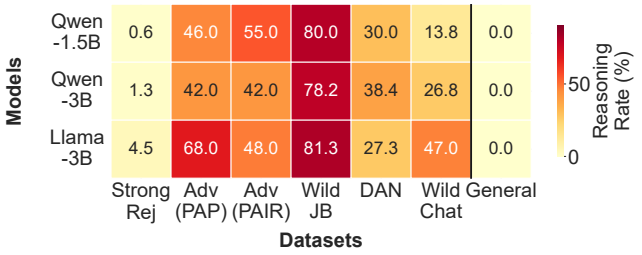

Figure 4: EASE adaptive reasoning rates across datasets showing selective activation based on threat sophistication. General tasks including MMLU, HellaSwag, and GSM8K.

Adaptive Safety Reasoning Mechanism To demonstrate EASE's adaptive reasoning activation mechanism, we analyze the safety reasoning rates across different query types, as shown in Figure 4. The results reveal EASE's adap-

Model	Method	Safety (Tokens ↓)				General (Tokens ↓)		
		StrongReject	WildJailbreak	DAN	WildChat	MMLU	Hellaswag	GSM8K
Qwen-2.5-1.5B -Instruct	Instruct	66.8	666.4	179.1	451.9	51.4	29.0	193.3
	Refusal Training	23.9	309.7	90.1	175.6	62.9	30.0	188.3
	Deliberative Alignment	307.6	542.2	343.2	488.0	284.3	317.0	336.5
	EASE *(vs. Deliberative)*	**44.3** *(-86%)*	**367.3** *(-32%)*	**246.0** *(-28%)*	**242.3** *(-50%)*	**52.0** *(-82%)*	**31.0** *(-90%)*	**224.7** *(-33%)*
Qwen2.5-3B -Instruct	Instruct	218.2	820.6	219.7	430.7	50.6	37.0	279.6
	Refusal Training	18.9	249.7	83.3	162.8	49.8	27.3	276.9
	Deliberative Alignment	311.9	559.1	352.5	479.1	291.7	331.2	345.1
	EASE *(vs. Deliberative)*	**38.6** *(-88%)*	**342.7** *(-39%)*	**207.3** *(-41%)*	**298.4** *(-38%)*	**51.2** *(-82%)*	**35.1** *(-89%)*	**295.1** *(-15%)*
Llama3.2-3B -Instruct	Instruct	98.2	492.9	1245.9	1189.6	29.9	27.2	228.9
	Refusal Training	21.6	60.7	77.1	102.2	42.5	28.9	252.9
	Deliberative Alignment	311.0	646.6	346.5	601.8	330.6	340.3	303.3
	EASE *(vs. Deliberative)*	**50.8** *(-84%)*	**333.4** *(-48%)*	**202.5** *(-42%)*	**292.8** *(-51%)*	**60.8** *(-82%)*	**32.5** *(-90%)*	**222.6** *(-27%)*

Table 3: Comparison of generation efficiency across datasets. EASE (Ours) shows significant token reduction compared to state-of-the-art defend method *Deliberative Alignment* while maintaining safety performance.

tive behavior patterns that align with query complexity and threat levels. For simple direct attacks like StrongReject, EASE maintains low reasoning rates (0.6-4.5%), indicating efficient handling without unnecessary computational overhead. In contrast, adversarial datasets trigger substantially higher reasoning rates. When facing AdvBench dataset augmented with PAIR attacks, EASE demonstrates high safety reasoning rates (55.0-81.3%), proving the effectiveness of its adaptive safety reasoning mechanism against state-of-the-art jailbreak methods. Similarly, WildJailbreak, which contains exclusively adversarial jailbreak tactics, shows consistently high reasoning rates (78.2-81.3%), confirming EASE's ability to detect and perform safety reasoning to complex threats. Real-world malicious datasets (DAN and WildChat) show intermediate safety reasoning rates (13.8-47.0%), reflecting their mixed nature of containing both adversarial jailbreak queries and simpler harmful content. This demonstrates EASE's practical applicability in realistic deployment scenarios. Crucially, general tasks maintain 0% reasoning activation across all models, confirming that EASE preserves computational efficiency for benign queries.

Ablation Study

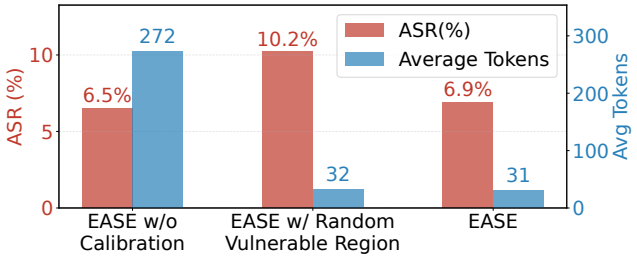

Figure 5: Ablation study demonstrating EASE's optimal balance between safety (ASR on WildJailbreak) and efficiency (average tokens on HellaSwag).

The Effectiveness of Safety Reasoning Calibration To validate our safety reasoning calibration approach, we compare EASE with two variants: (1) EASE w/o Reasoning Boundary Calibration, which applies reasoning to all queries, and (2) EASE w/ Random Vulnerable Region Selection, which randomly selects reasoning training data. As shown in Figure 5, without calibration, the model suffers from excessive computational overhead despite good safety performance. Random selection leads to degraded safety outcomes. Our targeted approach successfully balances both objectives, demonstrating that vulnerable region identification is essential for achieving optimal safety-efficiency trade-offs.

Conclusion

In this paper, we present EASE, a practical and efficient safety alignment framework for small language models that addresses the critical challenge of balancing safety robustness with computational efficiency, which is essential for the resource-constrained deployment scenarios that small language models are specifically designed for. Through a two-phase methodology combining safety reasoning knowledge distillation and safety reasoning boundary calibration, EASE enables selective safety reasoning activation for adversarial attacks while preserving computational efficiency for general tasks. Our experimental evaluation across multiple safety benchmarks shows that EASE maintains strong safety performance while significantly improving inference efficiency compared to existing deliberative methods. The framework successfully preserves general task capabilities while providing robust protection against sophisticated jailbreak attacks. EASE demonstrates that adaptive safety reasoning can achieve both safety and efficiency objectives for small language models, establishing a practical approach for safe deployment in resource-constrained environments.

Acknowledgments

We thank all the constructive feedback provided by the AAAI '26 reviewers. This work has been supported by an ONR grant N00014-23-1-2137 and an NSF award CNS-2442976.

References

Askell, A.; Bai, Y.; Chen, A.; Drain, D.; Ganguli, D.; Henighan, T.; Jones, A.; Joseph, N.; Mann, B.; DasSarma, N.; et al. 2021. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*.

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

Chao, P.; Robey, A.; Dobriban, E.; Hassani, H.; Pappas, G. J.; and Wong, E. 2025. Jailbreaking black box large language models in twenty queries. In *2025 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, 23–42. IEEE.

Cobbe, K.; Kosaraju, V.; Bavarian, M.; Chen, M.; Jun, H.; Kaiser, L.; Plappert, M.; Tworek, J.; Hilton, J.; Nakano, R.; et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Cui, G.; Yuan, L.; Ding, N.; Yao, G.; He, B.; Zhu, W.; Ni, Y.; Xie, G.; Xie, R.; Lin, Y.; Liu, Z.; and Sun, M. 2024. ULTRAFEEDBACK: boosting language models with scaled AI feedback. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

DeepSeek-AI. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. arXiv:2501.12948.

Gao, L.; Tow, J.; Abbasi, B.; Biderman, S.; Black, S.; DiPofi, A.; Foster, C.; Golding, L.; Hsu, J.; Le Noac'h, A.; Li, H.; McDonell, K.; Muennighoff, N.; Ociepa, C.; Phang, J.; Reynolds, L.; Schoelkopf, H.; Skowron, A.; Sutawika, L.; Tang, E.; Thite, A.; Wang, B.; Wang, K.; and Zou, A. 2024. The Language Model Evaluation Harness.

Guan, M. Y.; Joglekar, M.; Wallace, E.; Jain, S.; Barak, B.; Helyar, A.; Dias, R.; Vallone, A.; Ren, H.; Wei, J.; et al. 2024. Deliberative alignment: Reasoning enables safer language models. *arXiv preprint arXiv:2412.16339*.

Gunter, T.; Wang, Z.; Wang, C.; Pang, R.; Narayanan, A.; Zhang, A.; Zhang, B.; Chen, C.; Chiu, C.-C.; Qiu, D.; et al. 2024. Apple intelligence foundation language models. *arXiv preprint arXiv:2407.21075*.

Hendrycks, D.; Burns, C.; Basart, S.; Zou, A.; Mazeika, M.; Song, D.; and Steinhardt, J. 2021. Measuring Massive Multitask Language Understanding. In *International Conference on Learning Representations*.

Huang, Z.; Zou, H.; Li, X.; Liu, Y.; Zheng, Y.; Chern, E.; Xia, S.; Qin, Y.; Yuan, W.; and Liu, P. 2024. O1 Replication Journey–Part 2: Surpassing O1-preview through Simple Distillation, Big Progress or Bitter Lesson? *arXiv preprint arXiv:2411.16489*.

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*.

Javaheripi, M.; Bubeck, S.; Abdin, M.; Aneja, J.; Bubeck, S.; Mendes, C. C. T.; Chen, W.; Del Giorno, A.; Eldan, R.; Gopi, S.; et al. 2023. Phi-2: The surprising power of small language models. *Microsoft Research Blog*, 1(3): 3.

Ji, J.; Chen, B.; Lou, H.; Hong, D.; Zhang, B.; Pan, X.; Qiu, T.; Dai, J.; and Yang, Y. 2024. Aligner: Efficient Alignment by Learning to Correct. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Jiang, L.; Rao, K.; Han, S.; Ettinger, A.; Brahman, F.; Kumar, S.; Mireshghallah, N.; Lu, X.; Sap, M.; Choi, Y.; et al. 2024. Wildteaming at scale: From in-the-wild jailbreaks to (adversarially) safer language models. *Advances in Neural Information Processing Systems*, 37: 47094–47165.

KHIABANI, Y. S.; Atif, F.; Hsu, C.; Stahlmann, S.; Michels, T.; Kramer, S.; Heidrich, B.; Sarfraz, M. S.; Merten, J.; and Tafazzoli, F. 2024. Optimizing Small Language Models for In-Vehicle Function-Calling. In *NeurIPS 2024 Workshop on Fine-Tuning in Modern Machine Learning: Principles and Scalability*.

Li, N.; Han, Z.; Steneker, I.; Primack, W.; Goodside, R.; Zhang, H.; Wang, Z.; Menghini, C.; and Yue, S. 2024a. Llm defenses are not robust to multi-turn human jailbreaks yet. *arXiv preprint arXiv:2408.15221*.

Li, Y.; Yue, X.; Xu, Z.; Jiang, F.; Niu, L.; Lin, B. Y.; Ramasubramanian, B.; and Poovendran, R. 2025. Small models struggle to learn from strong reasoners. *arXiv preprint arXiv:2502.12143*.

Li, Z.; Liu, H.; Zhou, D.; and Ma, T. 2024b. Chain of thought empowers transformers to solve inherently serial problems. *arXiv preprint arXiv:2402.12875*, 1.

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*.

Liu, S.; Yao, Y.; Jia, J.; Casper, S.; Baracaldo, N.; Hase, P.; Yao, Y.; Liu, C. Y.; Xu, X.; Li, H.; et al. 2025a. Rethinking machine unlearning for large language models. *Nature Machine Intelligence*, 1–14.

Liu, W.; Zeng, W.; He, K.; Jiang, Y.; and He, J. 2023. What makes good data for alignment? a comprehensive study of automatic data selection in instruction tuning. *arXiv preprint arXiv:2312.15685*.

Liu, Y.; Gao, H.; Zhai, S.; Xia, J.; Wu, T.; Xue, Z.; Chen, Y.; Kawaguchi, K.; Zhang, J.; and Hooi, B. 2025b. Guardreasoner: Towards reasoning-based llm safeguards. *arXiv preprint arXiv:2501.18492*.

Liu, Z.; Sun, X.; and Zheng, Z. 2024. Enhancing llm safety via constrained direct preference optimization. *arXiv preprint arXiv:2403.02475*.

Liu, Z.; Zhao, C.; Iandola, F.; Lai, C.; Tian, Y.; Fedorov, I.; Xiong, Y.; Chang, E.; Shi, Y.; Krishnamoorthi, R.; et al.

2024. Mobilellm: Optimizing sub-billion parameter language models for on-device use cases. In *Forty-first International Conference on Machine Learning*.

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

Mou, Y.; Luo, Y.; Zhang, S.; and Ye, W. 2025. SaRO: Enhancing LLM Safety through Reasoning-based Alignment. *arXiv preprint arXiv:2504.09420*.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744.

Qi, X.; Panda, A.; Lyu, K.; Ma, X.; Roy, S.; Beirami, A.; Mittal, P.; and Henderson, P. 2025. Safety Alignment Should be Made More Than Just a Few Tokens Deep. In *The Thirteenth International Conference on Learning Representations*.

Qwen; :; Yang, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Li, C.; Liu, D.; Huang, F.; Wei, H.; Lin, H.; Yang, J.; Tu, J.; Zhang, J.; Yang, J.; Yang, J.; Zhou, J.; Lin, J.; Dang, K.; Lu, K.; Bao, K.; Yang, K.; Yu, L.; Li, M.; Xue, M.; Zhang, P.; Zhu, Q.; Men, R.; Lin, R.; Li, T.; Tang, T.; Xia, T.; Ren, X.; Ren, X.; Fan, Y.; Su, Y.; Zhang, Y.; Wan, Y.; Liu, Y.; Cui, Z.; Zhang, Z.; and Qiu, Z. 2025. Qwen2.5 Technical Report. arXiv:2412.15115.

Rafailov, R.; Sharma, A.; Mitchell, E.; Manning, C. D.; Ermon, S.; and Finn, C. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36: 53728–53741.

Shen, X.; Chen, Z.; Backes, M.; Shen, Y.; and Zhang, Y. 2024. "Do Anything Now": Characterizing and Evaluating In-The-Wild Jailbreak Prompts on Large Language Models. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, CCS '24, 1671–1685. New York, NY, USA: Association for Computing Machinery. ISBN 9798400706363.

Snell, C.; Klein, D.; and Zhong, R. 2022. Learning by distilling context. *arXiv preprint arXiv:2209.15189*.

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. *Advances in Neural Information Processing Systems*, 37: 125416–125440.

Taori, R.; Gulrajani, I.; Zhang, T.; Dubois, Y.; Li, X.; Guestrin, C.; Liang, P.; and Hashimoto, T. B. 2023. Stanford alpaca: An instruction-following llama model.

Vodopivec, T.; Samothrakis, S.; and Ster, B. 2017. On monte carlo tree search and reinforcement learning. *Journal of Artificial Intelligence Research*, 60: 881–936.

Wang, H.; Qin, Z.; Shen, L.; Wang, X.; Cheng, M.; and Tao, D. 2025a. Leveraging reasoning with guidelines to elicit and utilize knowledge for enhancing safety alignment. *arXiv preprint arXiv:2502.04040*, 3.

Wang, Z.; Tu, H.; Wang, Y.; Wu, J.; Mei, J.; Bartoldson, B. R.; Kailkhura, B.; and Xie, C. 2025b. Star-1: Safer alignment of reasoning llms with 1k data. *arXiv preprint arXiv:2504.01903*.

Xie, Y.; Goyal, A.; Zheng, W.; Kan, M.-Y.; Lillicrap, T. P.; Kawaguchi, K.; and Shieh, M. 2024. Monte carlo tree search boosts reasoning via iterative preference learning. *arXiv preprint arXiv:2405.00451*.

Yi, S.; Cong, T.; He, X.; Li, Q.; and Song, J. 2025. Beyond the Tip of Efficiency: Uncovering the Submerged Threats of Jailbreak Attacks in Small Language Models. *arXiv preprint arXiv:2502.19883*.

Zellers, R.; Holtzman, A.; Bisk, Y.; Farhadi, A.; and Choi, Y. 2019. Hellaswag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*.

Zeng, Y.; Lin, H.; Zhang, J.; Yang, D.; Jia, R.; and Shi, W. 2024. How Johnny Can Persuade LLMs to Jailbreak Them: Rethinking Persuasion to Challenge AI Safety by Humanizing LLMs. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 14322–14350. Bangkok, Thailand: Association for Computational Linguistics.

Zhang, P.; Zeng, G.; Wang, T.; and Lu, W. 2024. Tinyllama: An open-source small language model. *arXiv preprint arXiv:2401.02385*.

Zhang, W.; Xu, H.; Wang, Z.; He, Z.; Zhu, Z.; and Ren, K. 2025a. Can Small Language Models Reliably Resist Jailbreak Attacks? A Comprehensive Evaluation. *arXiv preprint arXiv:2503.06519*.

Zhang, Y.; Li, M.; Han, W.; Yao, Y.; Cen, Z.; and Zhao, D. 2025b. Safety is Not Only About Refusal: Reasoning-Enhanced Fine-tuning for Interpretable LLM Safety. *arXiv preprint arXiv:2503.05021*.

Zhang, Y.; Zhang, S.; Huang, Y.; Xia, Z.; Fang, Z.; Yang, X.; Duan, R.; Yan, D.; Dong, Y.; and Zhu, J. 2025c. Stair: Improving safety alignment with introspective reasoning. *arXiv preprint arXiv:2502.02384*.

Zhang, Z.; Loye, X. Q.; Huang, V. S.-J.; Yang, J.; Zhu, Q.; Cui, S.; Mi, F.; Shang, L.; Wang, Y.; Wang, H.; et al. 2025d. How Should We Enhance the Safety of Large Reasoning Models: An Empirical Study. *arXiv preprint arXiv:2505.15404*.

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.; Gonzalez, J. E.; Stoica, I.; and Zhang, H. 2024. LMSYS-Chat-1M: A Large-Scale Real-World LLM Conversation Dataset. In *The Twelfth International Conference on Learning Representations*.

Zou, A.; Phan, L.; Wang, J.; Duenas, D.; Lin, M.; Andriushchenko, M.; Kolter, J. Z.; Fredrikson, M.; and Hendrycks, D. 2024. Improving alignment and robustness with circuit breakers. *Advances in Neural Information Processing Systems*, 37: 83345–83373.

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*.