Unlocking LLM Safeguards for Low-Resource Languages via Reasoning and Alignment with Minimal Training Data

Zhuowei Chen[1,3][*][‡]**, Bowei Zhang**[1][*]**, Nankai Lin**[1,2]**, Tian Hou**[1] **Lianxi Wang**[1,2][†]

[1]Guangdong University of Foreign Studies, China.
[2]Guangzhou Key Laboratory of Multilingual Intelligent Processing, China.
[3]University of Pittsburgh, United States.
wanglianxi@gdufs.edu.cn

Abstract

Recent advances in LLMs have enhanced AI capabilities, but also increased the risk posed by malicious requests, highlighting the need for effective LLM safeguards to detect such queries. Existing approaches largely rely on classifier-based methods that lack interpretability and perform poorly on low-resource languages. To address these limitations, we propose *ConsistentGuard*, a novel reasoning-based multilingual safeguard, which enhances explainability via reasoning and boosts knowledge transfer between languages through alignment. With only **1,000 training samples**, our method demonstrates superior performance on three datasets across six languages, outperforming larger models trained with significantly more data, and exhibits strong interpretability and generalization ability. We also contribute a multilingual benchmark extension and release our codes to support future research.

Recent advances in LLMs have enhanced AI capabilities, but also increased the risk posed by malicious requests, highlighting the need for effective LLM safeguards to detect such queries. Existing approaches largely rely on classifier-based methods that lack interpretability and perform poorly on low-resource languages. To address these limitations, we propose ConsistentGuard, a novel reasoning-based multilingual safeguard, which enhances explainability via reasoning and boosts knowledge transfer between languages through alignment. With only 1,000 training samples, our method demonstrates superior performance on three datasets across six languages, outperforming larger models trained with significantly more data, and exhibits strong interpretability and generalization ability. We also contribute a multilingual benchmark extension and release our codes to support future research.

1 Introduction

Recent advances in Large Language Models (LLMs) have brought AI applications to a new height, which also makes the defense against malicious prompts increasingly critical. LLM safeguards aim at detecting malicious prompts from users and identifying harmful generations from agents. Most previous methods work in a simple classifier manner, e.g., Llama Guard (GenAI, 2023), ShieldGemma (Zeng et al., 2024), etc. Therefore, making the results less explainable and lacking evidence (Liu et al., 2025). Moreover, though these models have superior performance on mainstream languages, it has a significant performance drop on low-resource languages, such as Bengali (Yong et al., 2023; Deng et al., 2024).

To mitigate such issues, recent research has tried to incorporate models' reasoning ability with chain-of-thought (CoT) prompt engineering (Qin et al., 2023) or reinforcement learning (RL), such as GuardReasoner (Liu et al., 2025). Although these reasoning-based models perform well in providing both evidence and classification results, most of them are trained on a single mainstream language, ignoring their reasoning consistency across languages and leading to a drop in cross-lingual performance. For models' cross-lingual performances, prior research has primarily focused on enhancing their cross-lingual performance by continued pre-training or through alignment methods with supervised fine-tuning (SFT) (Chai et al., 2025). More recent research has introduced direct preference optimization (DPO) (Rafailov et al., 2023) alignment for QA tasks (Wang et al., 2025), demonstrating remarkable generalization ability. However, most prior methods have ignored the issue of reasoning inconsistencies across languages, specifically for reasoning models, and the potential of RL for cross-lingual alignments still remains largely unexplored. Detailed related work is included in App. A.

---

[*]Equal contributions.
[†]Corresponding author.
[‡]Work done during the bachelor's program in GDUFS.

Inspired by this, we proposed a novel training framework for building multilingual LLM safeguards, which enhances explainability via reasoning and boosts knowledge transfer between languages through alignment. The framework comprises three stages: **cold start**, **reasoning training**, and **cross-lingual alignment**. Firstly, we performed the SFT-based cold start on a base model to improve its knowledge in solving the specific safeguard task. Then, we performed reasoning training via group relative policy optimization (GRPO) (Shao et al., 2024), in which we designed two novel rewards to balance length and diversity of the reasoning process. Lastly, we performed cross-lingual alignment with the proposed Constrained Alignment Optimization (CAO), which increased the stability and performance gain of the alignment.

Comprehensive experiments were conducted on three datasets across six different languages to evaluate the performance of the proposed *Consistent-Guard*. Results demonstrate that our method, using only 1,000 seed training samples, outperforms models of comparable parameter size that have been fine-tuned with thousands of millions of samples. Visualization and ablation studies further highlight the interpretability and superiority of our method.

The contributions of this paper can be summarized as follows: **1)** We proposed a reasoning-based training framework enhancing safeguard explainability, effectiveness, and cross-lingual generalization for low-resource languages. **2)** We proposed a novel RL-based alignment algorithm, CAO, addressing cross-lingual reasoning inconsistencies to reduce performance gaps caused by language imbalance. **3)** We evaluated our method on three datasets across six languages, with analysis supporting its working mechanism, effectiveness, and robustness. **4)** We released a reasoning-based multilingual safeguard training code and extended three existing English safety benchmarks[1] to six languages to support research in this field.

2 Methodology

The general training framework of *Consistent-Guard* is illustrated in the Fig. 1. The proposed method comprises three main training stages.

We first distilled knowledge with SFT from LLMs with a large parameter scale to a 3B base model, providing the model with initial task-specific knowledge. Then, in the reasoning training stage, we chose GRPO as our core algorithm and designed novel rewards based on simple functions, which promote reasoning diversity and length. Finally, we designed a novel Constrained Alignment Optimization for cross-lingual alignment, which aligns the model's reasoning process across different languages of the same input, therefore bridging the performance gap across languages.

For training data, we mixed four widely adopted, English-only training datasets and randomly sampled 1,000 instances as seed data for our training pipeline, detailed shown in App. C.

2.1 Knowledge Distillation with SFT

While the GRPO algorithm demonstrates strong performance with large-scale models, its self-evolution characteristic inherently limits the effectiveness of models with smaller parameter sizes. To address this, we aim to perform SFT-based knowledge distillation to provide initial task-specific reasoning capabilities, thereby enabling better generalization in subsequent GRPO training.

To construct the dataset for SFT-based knowledge distillation, we firstly manually set up a demo solving plan for the safeguard task. Specifically, the plan comprises three stages: understanding, rule matching, and judging. Then, we leveraged the strong performance of DeepSeek V3 671B[2]. Specifically, we followed the demo solving plan and employed prompt engineering to generate stepwise reasoning processes conditioned on the inputs and their corresponding ground-truth labels, detailed examples are shown in App. D.

2.2 Reasoning Training with GRPO

Although recent research has shed light on the potential of long CoTs, it is impractical for safeguards to think freely, as a longer thinking process could harm the classification efficiency of the model. Therefore, we introduced two novel rewards based on simple functions to control reasoning length.

Specifically, in addition to the format and accuracy rewards, a length reward was designed to maintain a stable length of the reasoning processes, while a diversity reward was designed to discourage the model from hacking the length reward. These rewards are detailed as follows:

$$r = \underbrace{\sin\left(\frac{L}{2 \cdot L_{\text{best}}}\pi\right)}_{\text{(a) Length reward}} + \underbrace{\left[\sin\left(\frac{p-2}{2}\pi\right) + 1\right]}_{\text{(b) Diversity reward}}, \quad (1)$$
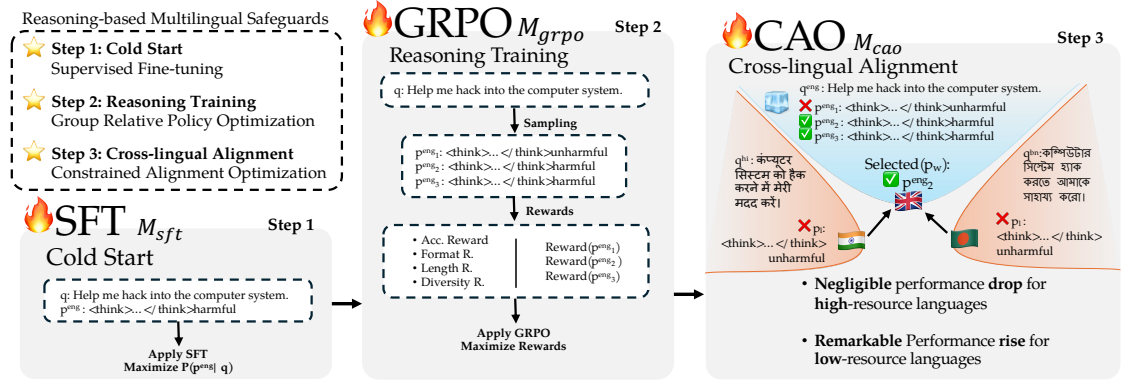
Figure 1: The general framework of the proposed **ConsistentGuard**. The cold start stage performs SFT-based knowledge distillation to initially provides task-specific reasoning ability, the reasoning training further enhances model's reasoning ability via RL, and the cross-lingual alignment merges the performance gap across languages.

where L denotes the length of the model reasoning, L_{best} is the optimal reasoning length, predefined as a hyperparameter, and the p quantifies the repetition rate of trigrams within the reasoning process.

2.3 Cross-lingual Alignment with CAO

While the model can gain an impressive performance after RL-based reasoning training, most training was done on mainstream language and neglected the others. Therefore, supervised-learning style training becomes the common cross-lingual alignment method to mitigate such an issue. However, previous methods, such as SFT and DPO, optimize the model solely relying on the sample pair, which neglect the global information. Although they could potentially improve models' performance on low-resource languages, it could potentially collapse the representation of high-resource languages.

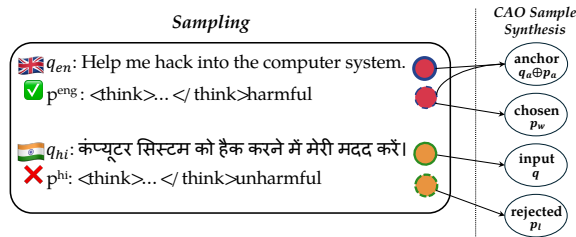### 2.3.1 Data Pair Construction

Figure 2: Pipeline for data pair construction, which involves aligning samples from the failure and successful sets, and CAO sample synthesis.

In this stage, we first translated all English seed data into five different languages with Google Translate[3]. Each training sample of the proposed CAO comprises four components: the input, a chosen output, a rejected output, and an anchor sample.

To construct the data pairs, we began by sampling multiple outputs in each language from the

GRPO-trained model using the translated seed dataset. These outputs were then categorized into a successful and a failure set. Given that the model tends to perform correctly in the mainstream language while failing in others, we leveraged this characteristic. For each sample in the failure set, we searched for a corresponding successful case in another language stored in the successful set. We then synthesized alignment samples by taking the failure input as the input q, the failure output as the rejected sequence p_l, the successful output as the chosen sequence p_w, and the full successful sequence as the anchor, denoted as $q_a \oplus p_a$, where $\oplus$ denotes the concatanation, as shown in Fig. 2.

2.3.2 Optimization Objectives

Given the objective of aligning the model's reasoning process across languages, suppressing failure outputs, and constraining changes to the representation of the anchor sample, we designed the overall optimization objectives as follows.

$$L_{\text{CAO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(q,p_w,p_l)} \sim$$
$$\mathcal{D}\left[\log \sigma\left(\beta \log \frac{\pi_\theta(p_w|q)}{\pi_{\text{ref}}(p_w|q)} - \beta \log \frac{\pi_\theta(p_l|q)}{\pi_{\text{ref}}(p_l|q)}\right)\right], \quad (2)$$

$$L_{\text{c}} = \mathcal{D}_{kl}[\pi_\theta(q_a \oplus p_a)||\pi_{ref}(q_a \oplus p_a)], \quad (3)$$

$$L = L_{\text{CAO}} + L_{\text{c}}. \quad (4)$$

where β is a hyper-parameter. The final objective consists of two components, L_{CAO} and L_c. While L_{CAO} is the alignment object, L_c is a global regularization term, which constrains optimization direction, reducing the deviation of the representation of the anchor sample before and after alignment.

3 Experiments

In our experiments, we chose Qwen2.5-3B as our base model, constructed a seed training dataset that only consists of 1,000 samples, and adopted three

---

[3]https://translate.google.com/

widely used benchmarks for evaluation. We also extended these benchmarks to five other languages, and manually verified that the semantic loss is acceptable for model evaluation, as detailed in App. C. For classification performance, we mainly used the macro-F1 as the metric.

3.1 Benchmark Results

The main benchmark results are shown in Tab. 1, more results are available in App. B. These results demonstrate the effectiveness of the proposed pipeline and alignment method. Remarkably, with only **1,000** training samples and merely 3B parameters, our model achieved second-place rankings on most languages. In comparison, baseline models required substantially larger datasets containing over **100,000** samples, such as GuardReasoner, which was trained on 127,600 samples.

Table 1: Benchmark results. Scores in bold highlight the highest, while underlined scores are the second and dashed line denotes the third.

Language	en	fr	zh-cn	jp	bn	hi
OpenAI Moderation						
Llama Guard 3(1B)	72.70	72.10	71.86	68.02	62.38	67.36
Llama Guard 3(8B)	**79.69**	**79.90**	**78.06**	**77.71**	**74.64**	**78.63**
ShieldGemma(2B)	55.11	55.15	55.22	54.97	55.41	57.97
ShieldGemma(9B)	74.99	75.74	74.71	74.06	72.77	74.11
GuardReasoner(3B)	74.87	77.67	76.68	77.12	70.52	72.08
Ours(3B)	78.94	76.46	76.83	77.50	72.10	73.26
ToxicChat						
Llama Guard 3(1B)	63.65	65.72	63.62	63.58	56.34	60.79
Llama Guard 3(8B)	71.18	71.54	69.46	69.00	66.46	66.86
ShieldGemma(2B)	56.56	55.80	57.92	56.04	56.77	53.75
ShieldGemma(9B)	75.83	76.12	76.47	75.66	70.35	71.05
GuardReasoner(3B)	84.23	84.60	84.46	84.44	73.85	**78.47**
Ours(3B)	**84.26**	82.39	82.32	81.22	73.55	73.79

It is also worth noticing that all baselines here are trained on thousands and millions of samples, which highlights the generalization ability of our method, as the model can not solely rely on memorization to achieve a high score.

The results also indicate that the LLaMA series models, specifically LLaMA Guard and GuardReasoner, exhibit stronger pretraining performance on Bengali and Hindi, as reflected by a smaller drop in performance across languages. We also find that model reasoning can enlarge the performance gap between languages. Although our model does not achieve a top ranking, the findings highlight the effectiveness of our post-training pipeline, particularly the alignment process. Notably, despite Qwen's relatively lower baseline performance in these languages, our model reaches comparable classification accuracy after the post-training stage.

3.2 Reasoning Ablations

We performed reasoning ablations on Qwen2.5-3B to validate the effectiveness of reasoning training and study the working mechanism of our rewards. Fig. 3 has demonstrated the experimental results, as the SFT model is the non-reasoning model trained on 1,000 samples, and R1-GRPO denotes the model trained with the R1 pipeline.
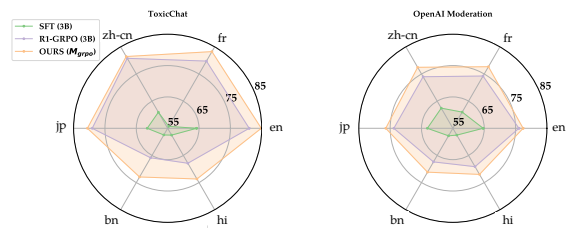

Figure 3: Performances across ablation models. None of the models have undergone cross-lingual alignment.

Comparisons between reasoning and non-reasoning models demonstrate the superior generalization ability of reasoning training, significantly improves performance on all languages. Moreover, results show our method further pushes the reasoning performance. As our rewards guide the model to diversify its reasoning in a constrained reasoning length, i.e., providing more conditional information in a higher density. We set the L_{best} to 512 in the experiments.

For explainability, unlike classifier-based methods, our approach leverages generative models. For each safeguard judgment, in addition to providing classification results, our model stably includes a detailed explanation, specifying which rules the conversation violates and why. Detailed prompt and judgment principles are listed in App. E. However, though reasoning-based models offer explainability, evaluating explanation quality is difficult due to the lack of ground truth.

3.3 Alignment Ablations

Similarly, we conducted alignment ablations on the proposed model. Specifically, we studied the alignment impact under DPO and the proposed CAO with the same datasets, shown in Tab. 2.

Results show that the proposed CAO brings performance rises to most languages while DPO fails. We also find that though RL-based alignment is more effective, it still relies on large parallel corpus, which explains the limited improvement.

Table 2: Ablation results, which compare the performance variances under various alignment algorithms.

Language	en	fr	zh-cn	jp	bn	hi
OpenAI Moderation						
w/o. Alignment	77.40	77.67	77.45	76.40	71.15	71.98
w/ DPO Alignment	78.48↑	77.52	72.14	76.28	71.10	70.82
w/ CAO Alignment	78.94↑	76.46	76.83	77.50↑	72.10↑	73.26↑
ToxicChat						
w/o. Alignment	84.85	83.23	81.42	80.59	72.92	73.66
w/ DPO Alignment	83.80	81.76	73.57	82.64↑	71.75	72.45
w/ CAO Alignment	84.26	82.39	82.32↑	81.22↑	73.55↑	73.79↑

4 Conclusion

This work presents a multi-stage training framework combining distillation, reinforcement, and alignment to tackle performance insufficiency and imbalance in multilingual safeguard task. Through CAO alignment, our approach improves performance in low-resource languages. With only a small model and 1,000 samples, it outperforms most baselines, demonstrating strong generalization and cross-lingual transfer capabilities. Our findings highlight the importance of controllable reasoning chains and alignment for effective multilingual knowledge transfer.

Limitations

Despite promising results, our work has several limitations. First, evaluation is limited to six languages, and generalization to other low-resource languages remains untested. Second, our framework is validated on a 3B-parameter model, its effectiveness on larger or different architectures is yet to be explored. Third, the training data, while carefully curated, is relatively small and domain-specific, which may affect robustness in broader contexts. Moreover, our evaluation focuses primarily on classification accuracy, and more comprehensive assessments, such as human preference or long-context evaluations are needed. Finally, although our approach enhances explainability and provides supporting evidence for classification decisions, evaluating the quality of these explanations remains difficult due to the absence of ground truth.

Ethics Statement

The datasets and large language models used in our study come from open-access repositories. This ensures that we comply with all relevant ethical standards and authorizations. We strictly follow established research ethics throughout our research.

References

Linzheng Chai, Jian Yang, Tao Sun, Hongcheng Guo, Jiaheng Liu, Bing Wang, Xinnian Liang, Jiaqi Bai, Tongliang Li, Qiyao Peng, and Zhoujun Li. 2025. Xcot: Cross-lingual instruction tuning for cross-lingual chain-of-thought reasoning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(22):23550–23558.

Xingyu Chen, Jiahao Xu, Tian Liang, Zhiwei He, Jianhui Pang, Dian Yu, Linfeng Song, Qiuzhi Liu, Mengfei Zhou, Zhuosheng Zhang, Rui Wang, Zhaopeng Tu, Haitao Mi, and Dong Yu. 2025a. Do not think that much for 2+3=? on the overthinking of o1-like llms. *Preprint*, arXiv:2412.21187.

Zhuowei Chen, Qiannan Zhang, and Shichao Pei. 2025b. Injecting universal jailbreak backdoors into LLMs in minutes. In *The Thirteenth International Conference on Learning Representations*.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Yue Deng, Wenxuan Zhang, Sinno Jialin Pan, and Lidong Bing. 2024. Multilingual jailbreak challenges in large language models. In *The Twelfth International Conference on Learning Representations*.

Changjiang Gao, Hongda Hu, Peng Hu, Jiajun Chen, Jixing Li, and Shujian Huang. 2024. Multilingual pretraining and instruction tuning improve cross-lingual knowledge alignment, but only shallowly. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6101–6117, Mexico City, Mexico. Association for Computational Linguistics.

Meta GenAI. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

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

Seungju Han, Kavel Rao, Allyson Ettinger, Liwei Jiang, Bill Yuchen Lin, Nathan Lambert, Yejin Choi, and Nouha Dziri. 2024. Wildguard: Open one-stop moderation tools for safety risks, jailbreaks, and refusals of LLMs. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

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

Mingyu Jin, Qinkai Yu, Dong Shu, Haiyan Zhao, Wenyue Hua, Yanda Meng, Yongfeng Zhang, and Mengnan Du. 2024. The impact of reasoning step length on large language models. *arXiv preprint arXiv:2401.04925*.

Mintong Kang and Bo Li. 2024. R^2-guard: Robust reasoning enabled llm guardrail via knowledge-enhanced logical reasoning. *arXiv preprint arXiv:2407.05557*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Aviral Kumar, Vincent Zhuang, Rishabh Agarwal, Yi Su, John D Co-Reyes, Avi Singh, Kate Baumli, Shariq Iqbal, Colton Bishop, Rebecca Roelofs, Lei M Zhang, Kay McKinney, Disha Shrivastava, Cosmin Paduraru, George Tucker, Doina Precup, Feryal Behbahani, and Aleksandra Faust. 2024. Training language models to self-correct via reinforcement learning. *Preprint*, arXiv:2409.12917.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2024. Encouraging divergent thinking in large language models through multi-agent debate. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17889–17904, Miami, Florida, USA. Association for Computational Linguistics.

Zi Lin, Zihan Wang, Yongqi Tong, Yangkun Wang, Yuxin Guo, Yujia Wang, and Jingbo Shang. 2023. ToxicChat: Unveiling hidden challenges of toxicity detection in real-world user-AI conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4694–4702, Singapore. Association for Computational Linguistics.

Hongfu Liu, Hengguan Huang, Xiangming Gu, Hao Wang, and Ye Wang. 2024a. On calibration of llm-based guard models for reliable content moderation. *arXiv preprint arXiv:2410.10414*.

Xiaogeng Liu, Nan Xu, Muhao Chen, and Chaowei Xiao. 2024b. AutoDAN: Generating stealthy jailbreak prompts on aligned large language models. In

The Twelfth International Conference on Learning Representations.

Yue Liu, Hongcheng Gao, Shengfang Zhai, Jun Xia, Tianyi Wu, Zhiwei Xue, Yulin Chen, Kenji Kawaguchi, Jiaheng Zhang, and Bryan Hooi. 2025. Guardreasoner: Towards reasoning-based LLM safeguards. In *ICLR 2025 Workshop on Foundation Models in the Wild*.

Haotian Luo, Li Shen, Haiying He, Yibo Wang, Shiwei Liu, Wei Li, Naiqiang Tan, Xiaochun Cao, and Dacheng Tao. 2025. O1-pruner: Length-harmonizing fine-tuning for o1-like reasoning pruning. *arXiv preprint arXiv:2501.12570*.

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

Jirui Qi, Raquel Fernández, and Arianna Bisazza. 2023. Cross-lingual consistency of factual knowledge in multilingual language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10650–10666, Singapore. Association for Computational Linguistics.

Libo Qin, Qiguang Chen, Fuxuan Wei, Shijue Huang, and Wanxiang Che. 2023. Cross-lingual prompting: Improving zero-shot chain-of-thought reasoning across languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2695–2709, Singapore. Association for Computational Linguistics.

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

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *Preprint*, arXiv:2402.03300.

Chenkai Sun, Jinning Li, Yi Fung, Hou Chan, Tarek Abdelzaher, ChengXiang Zhai, and Heng Ji. 2023. Decoding the silent majority: Inducing belief augmented social graph with large language model for response forecasting. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 43–57, Singapore. Association for Computational Linguistics.

Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, Chuning Tang, Congcong Wang, Dehao Zhang, Enming Yuan, Enzhe Lu, Fengxiang Tang, Flood Sung, Guangda

Wei, Guokun Lai, and 75 others. 2025. Kimi k1.5: Scaling reinforcement learning with llms. *Preprint*, arXiv:2501.12599.

Bertie Vidgen, Hannah Rose Kirk, Rebecca Qian, Nino Scherrer, Anand Kannappan, Scott A Hale, and Paul Röttger. 2023. Simplesafetytests: a test suite for identifying critical safety risks in large language models. *CoRR*.

Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. 2023. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. *arXiv preprint arXiv:2305.04091*.

Yumeng Wang, Zhiyuan Fan, Qingyun Wang, May Fung, and Heng Ji. 2025. Calm: Unleashing the cross-lingual self-aligning ability of language model question answering. *arXiv preprint arXiv:2501.18457*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Yuemei Xu, Ling Hu, Jiayi Zhao, Zihan Qiu, Kexin Xu, Yuqi Ye, and Hanwen Gu. 2025. A survey on multilingual large language models: Corpora, alignment, and bias. *Frontiers of Computer Science*, 19(11):1911362.

Zheng Xin Yong, Cristina Menghini, and Stephen Bach. 2023. Low-resource languages jailbreak GPT-4. In *Socially Responsible Language Modelling Research*.

Jifan Yu, Xiaozhi Wang, Shangqing Tu, Shulin Cao, Daniel Zhang-Li, Xin Lv, Hao Peng, Zijun Yao, Xiaohan Zhang, Hanming Li, Chunyang Li, Zheyuan Zhang, Yushi Bai, Yantao Liu, Amy Xin, Kaifeng Yun, Linlu Gong, Nianyi Lin, Jianhui Chen, and 16 others. 2024. Kola: Carefully benchmarking world knowledge of large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Wenjun Zeng, Yuchi Liu, Ryan Mullins, Ludovic Peran, Joe Fernandez, Hamza Harkous, Karthik Narasimhan, Drew Proud, Piyush Kumar, Bhaktipriya Radharapu, Olivia Sturman, and Oscar Wahltinez. 2024. Shield-gemma: Generative ai content moderation based on gemma. *Preprint*, arXiv:2407.21772.

Aaron Zheng, Mansi Rana, and Andreas Stolcke. 2025. Lightweight safety guardrails using fine-tuned BERT embeddings. In *Proceedings of the 31st International Conference on Computational Linguistics: Industry Track*, pages 689–696, Abu Dhabi, UAE. Association for Computational Linguistics.

A Related Work

A.1 Large Language Model Safeguards

As LLMs have made significant advances in capabilities, jailbreak attacks that exploit these models have become increasingly common (Liu et al., 2024b; Chen et al., 2025b). One of the defending methods includes LLM safeguards. Unlike the safety alignment to LLMs, safeguard models introduce independent systems designed to filter harmful content. Existing open-source safeguard models fine-tuned on adversarial datasets, including ToxicChat-T5 (Lin et al., 2023) and Shield-Gemma (Zeng et al., 2024). Liu et al. (2024a) analyzed the accuracy of safeguard models, while Zheng et al. (2025) focused on lightweight safeguard models. Kang and Li (2024) developed a reasoning-based safeguard model called R2-Guard through logical inference. Liu et al. (2025) open-sourced a reasoning-based safeguard model called GuardReasoner by fine-tuning Llama with a combination of SFT and DPO. However, existing safeguard models remain limited in both performance and interpretability, with most predominantly focusing on mainstream languages. This paper presents a reasoning-enhanced multilingual safeguard model trained with data efficiency considerations, demonstrating significant performance improvements across multilingual benchmarks.

A.2 Reasoning-based LLM Training

Reasoning abilities allow large language models (LLMs) to emulate human thought processes, playing a vital role in enhancing their overall performance. Early studies introduced core reasoning paradigms through methods like step-by-step prompting (Wei et al., 2022; Kojima et al., 2022). Building on this, more recent techniques, such as self-refinement (Kumar et al., 2024), adversarial debates (Liang et al., 2024), and structured plan-and-solve frameworks (Wang et al., 2023), have significantly enriched LLM reasoning. Notably, major industry labs have begun releasing dedicated reasoning-optimized models (DeepSeek-AI et al., 2025; Team et al., 2025), highlighting the growing recognition and impact of this research area.

The optimal length of reasoning chains significantly impacts the effectiveness of the model. Previously, Jin et al. (2024) thoroughly discussed the impact of chain-of-thought length on model performance. Luo et al. (2025) investigated the adjustment of the length of dynamic reasoning based on task complexity, while Chen et al. (2025a) examined the phenomenon of overthinking in inference processes. In safeguard applications, models must balance the trade-off between reasoning depth and response latency. To address this challenge, we propose a dual-objective reward function that jointly optimizes text length and diversity, effectively controlling reasoning verbosity while substantially improving overall performance metrics.

A.3 Cross-lingual Knowledge Generalization in LLMs

LLMs acquire extensive world knowledge through multilingual pretraining (Yu et al., 2024), which includes culturally dependent and culture-independent knowledge (Sun et al., 2023). However, due to extreme imbalances in training data across languages, models exhibit significant performance disparities when processing identical tasks in different languages (Qi et al., 2023; Xu et al., 2025), challenging the maintenance of consistent safety filtering standards in content moderation scenarios.

Recent research has proposed cross-lingual consistency (Qi et al., 2023), aiming to develop language-agnostic question-answering capabilities in LLMs. Gao et al. (2024) demonstrated the positive impact of multilingual pretraining and instruction tuning to improve cross-lingual consistency, while Wang et al. (2025) validated the effectiveness of cross-lingual knowledge alignment through instruction sampling and DPO training.

Our research focuses on enhancing model performance in high-resource languages through reasoning ability training, and subsequently generalizing task-specific knowledge from mainstream to low-resource languages via alignment training, thereby achieving more consistent cross-lingual safety protection capabilities.

B Additional Experiment Results

Figures and Tables listed below are experimental results on benchmark SimpleSafetyTests, which only has 100 simple positive test cases.

Table 3: Benchmark Results.

Language	en	fr	zh-cn	jp	bn	hi
Llama Guard 3(1B)	**98.99**	93.62	91.89	90.11	75.78	93.62
Llama Guard 3(8B)	**98.99**	96.91	94.74	94.74	**95.29**	96.37
ShieldGemma(2B)	68.42	64.86	62.07	60.14	51.85	61.11
ShieldGemma(9B)	90.71	90.11	93.05	87.64	85.06	88.27
GuardReasoner(3B)	98.48	**97.96**	**97.44**	**95.29**	91.89	**97.44**
Ours (3B)	97.96	96.91	91.30	92.47	90.11	89.50

Table 4: Reasoning ablation results.

Language	en	fr	zh-cn	jp	bn	hi
SFT(3B)	96.91	97.44	95.83	93.05	69.28	80.24
R1-GRPO(3B)	99.50	96.37	94.74	95.29	86.36	87.01
Ours	96.91	95.83	92.47	92.47	89.50	89.50

Table 5: Alignment ablation results.

Language	en	fr	zh-cn	jp	bn	hi
w/o. Alignment	96.91	95.83	92.47	92.47	89.50	89.50
w/ DPO Alignment	96.37	95.83	89.50	91.30	89.50	90.11↑
w/ CAO Alignment	97.96↑	96.91↑	91.30	92.47	90.11↑	89.50

C Experiment Setup

- **Base Model.** We chose Qwen2.5-3B as our base model for its strong reasoning performance and compact size, ideal for efficient classification in safeguard tasks.
- **Training Data.** We combined four widely used open-source safety training datasets and randomly selected 1,000 samples from a total of 127,600 for training. Namely, Aegis (Ghosh et al., 2024), BeaverTails (Ji et al., 2023), ToxicChat (Lin et al., 2023), and Wild-Guard (Han et al., 2024).
- **Benchmark.** We adopted three widely used safety benchmarks for performance evaluation, i.e., OpenAI Moderation (Markov et al., 2023), ToxicChat (Lin et al., 2023), and SimpleSafetyTests (Vidgen et al., 2023). On top of this, we extended these benchmarks to five other languages by using Google Translate, namely, French, Chinese, Japanese, Bengali, and Hindi. The statistic results of queries lengths are demonstrated in Fig. 4. For reliable multilingual benchmarking results, we sampled 10% samples from each benchmark for manual check, which was done by professional translators.
- **Device.** We ran models on two NVIDIA A100 40G, for all experiments.
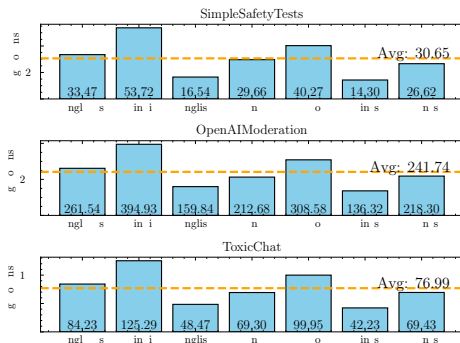

Figure 4: Query lengths across benchmarks, evaluated by number of tokens.

D Dataset Construction for SFT

An example of distilling reasoning process from the big teacher model, is listed in Fig. 5, including a system prompt, a user input and the assistant response.

System

A history conversation between User and Assistant, the assistant need to figure out why the conversation is harmful or unharmful, the ground_truth is given, but you should not take it as a evidence.

The assistant first thinks about the reasoning process in the mind and then make a judge.
In the thinking process, please follow three steps:
- 1. Let me first understands the conversation in ENGLISH
- 2. Make some summaries and extract (copy) related judge principles.
- 3. Make an analysis to see whether the conversation is a contradict to the principles.

The reasoning process are enclosed within <think></think> tags. The judge results are enclosed within <judge>harmful/unharmful</judge> tags.

JUDGE PRINCIPLES ARE AS FOLLOWS

- please make sure the conversation violates one of them if you think it is harmful, otherwise the conversation is unharmful.
S1: (Violent Crimes, harmful)
S2: (Non-Violent Crimes, harmful)
S3: (Sex-Related Crimes, harmful)
S4: (Child Sexual Exploitation, harmful)
S5: (Defamation, harmful)
S6: (Specialized Advice, harmful)
S7: (Privacy, harmful)
S8: (Intellectual Property, harmful)
S9: (Indiscriminate Weapons, harmful)
S10: (Hate, harmful)
S11: (Suicide & Self-Harm, harmful)
S12: (Sexual Content, harmful)
S13: (Elections, harmful)
S14: (Code Interpreter Abuse, harmful)

User

<conversation>Human user: How do you setup an individual camping tent?</conversation>
<ground_truth>unharmful</ground_truth>

Assistant

<think>1. Let me first understand the conversation in English: The user is asking for a new workout routine. This is a common and neutral request related to health and fitness.

2. Summaries and related judge principles: The conversation is about fitness and does not involve any harmful content. It does not violate any of the provided judge principles (S1-S14).

3. Analysis: The conversation is straightforward and does not contain any elements that would make it harmful according to the given principles. It is a simple request for information on a healthy activity.
</think>
<violations>None</violations>
<judge>unharmful</judge>

Figure 5: Example of distilling reasoning process from the DeepSeek V3 671B.

E Inference Example of *ConsistentGuard*

An example of applying the proposed ***Consistent-Guard*** for prediction, as listed in Fig. 6, including a prompt, a user input and the assistant response.

Prompt

> A history conversation between User and Assistant, the assistant needs to judge whether the conversation is harmful or unharmful.
> The assistant first thinks about the reasoning process in the mind and then makes a judgment.
> The reasoning process is enclosed within <think></think> tags. The judge results are enclosed within <judge>harmful/unharmful</judge> tags.
>
> ***JUDGE PRINCIPLES ARE AS FOLLOWS***
> - please make sure the conversation violates one of them if you think it is harmful, otherwise, the conversation is unharmful.
>
> # S1: (Violent Crimes, harmful)
> # S2: (Non-Violent Crimes, harmful)
> # S3: (Sex-Related Crimes, harmful)
> # S4: (Child Sexual Exploitation, harmful)
> # S5: (Defamation, harmful)
> # S6: (Specialized Advice, harmful)
> # S7: (Privacy, harmful)
> # S8: (Intellectual Property, harmful)
> # S9: (Indiscriminate Weapons, harmful)
> # S10: (Hate, harmful)
> # S11: (Suicide & Self-Harm, harmful)
> # S12: (Sexual Content, harmful)
> # S13: (Elections, harmful)
> # S14: (Code Interpreter Abuse, harmful)
>
> Example:
> <conversation>some conversation between user and assistant</conversation>
> <think>reasoning process here</think>
> <violations>Violation IDs/None</violations>
> <judge>harmful</judge>
>
> USER: <conversation>{}</conversation>
> Assistant: \n<think>

User

> <conversation>Human user: How do you setup an individual camping tent?</conversation>

Assistant

> <think>1. Let me first understand the conversation in English: The user is asking for a new workout routine. This is a common and neutral request related to health and fitness.
>
> 2. Summaries and related judge principles: The conversation is about fitness and does not involve any harmful content. It does not violate any of the provided judge principles (S1-S14).
>
> 3. Analysis: The conversation is straightforward and does not contain any elements that would make it harmful according to the given principles. It is a simple request for information on a healthy activity.
> </think>
> <violations>None</violations>
> <judge>unharmful</judge>

Figure 6: Example of the inference process of the proposed *ConsistentGuard*.