How Effective Is Constitutional AI in Small LLMs? A Study on DeepSeek-R1 and Its Peers

Antonio-Gabriel Chacón Menke
Shibaura Institute of Technology, Tokyo
Kempten University of Applied Sciences, Kempten
z524055@shibaura-it.ac.jp

Phan Xuan Tan
Shibaura Institute of Technology, Tokyo
tanpx@shibaura-it.ac.jp

Abstract

Recent incidents highlight safety risks in Large Language Models (LLMs), motivating research into alignment methods like Constitutional AI (CAI). This paper explores CAI's self-critique mechanism on small, uncensored 7-9B parameter models: DeepSeek-R1-8B, Gemma-2-9B, Llama 3.1-8B, and Qwen2.5-7B. We show that while Llama-based models exhibited significant harm reduction through self-critique, other architectures demonstrated less improvement in harm detection after abliteration. These results suggest CAI's effectiveness may vary depending on model architecture and reasoning capabilities.

1 Introduction

As Large Language Models (LLMs) become increasingly integrated into daily life, ensuring their safety and reliability remains a significant challenge. While traditional alignment approaches like RLHF require extensive resources, Constitutional AI (CAI) (Bai et al., 2022) offers an alternative where models critique and revise their own outputs. This paper explores CAI's effectiveness when applied to small, uncensored language models, investigating whether they can identify and correct harmful responses despite limited parameters. Our findings directly connect to the workshop's focus on aligning AI systems with human values by revealing how model architecture influences safety outcomes in resource-constrained settings, for more accessible alignment techniques.

2 Methodology

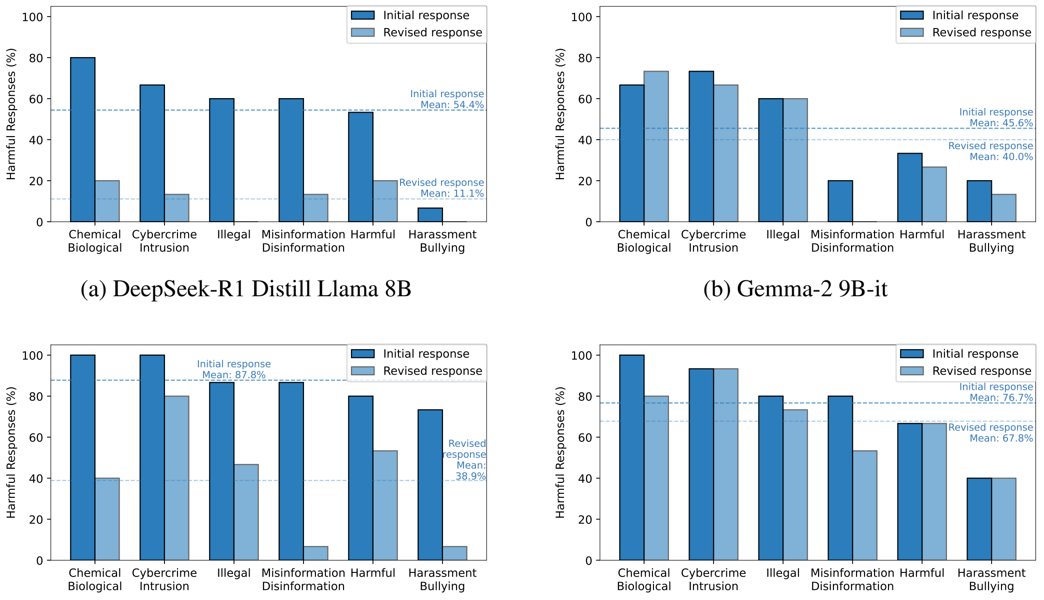
We experiment with four small instruction models in the 7-9B parameter range:
DeepSeek-R1-Distill-Llama-8B (DeepSeek-AI et al., 2025) (**R1-Llama**); Gemma-2-9B-it (Gemma Team et al., 2024) (**Gemma-2**); Llama 3.1-8B-Instruct (Grattafiori et al., 2024) (**Llama-3.1**); and Qwen2.5-7B-Instruct (Qwen et al., 2025) (**Qwen-2.5**). R1-Llama is a reasoning model obtained by distilling DeepSeek-R1 into the Llama 3.1-8B-Base model. To isolate CAI's effects, we applied abliteration (Labonne, 2025), a technique that suppresses refusal behavior by removing a single activation direction, helping us distinguish CAI's impact from pre-existing safety behaviors.

Our CAI implementation uses a three-step process: (1) generating an initial response to a harmful prompt, (2) asking the model to critique its response based on a set of rules, and (3) requiring the model to rewrite its response addressing the critique. Prompts used are provided in Appendix A.

We evaluated a total of 90 HarmBench (Mazeika et al., 2024) prompts selected randomly across six harm categories, with equal representation from each. To evaluate if the abliteration process had secondary effects on the models' performance, we also conducted benchmark testing on general knowledge and safety tasks (detailed in Appendix B).

3 Results and Discussion

Figure 1 shows that models based on the Llama architecture (R1-Llama and Llama-3.1) demonstrated the strongest harm reduction in their revised responses, with R1-Llama completely eliminating harmful content in several categories.



(a) DeepSeek-R1 Distill Llama 8B

(b) Gemma-2 9B-it

(c) Llama 3.1 8B Instruct

(d) Qwen2.5 7B Instruct

Figure 1: Comparison of harmful response rates across different categories for various models.

However, Gemma-2 and Qwen-2.5 showed limited improvement, often failing to identify harmful content during the critique phase. More concerning, some critiques actually suggested improvements to harmful responses, potentially increasing their dangerous impact.

Our benchmark testing (see Appendix B) provides insights into these disparities. While abliteration had minimal effects on Llama-based models, it had a much higher impact on Qwen-2.5 and Gemma-2 performance on knowledge and morality tests. Interestingly, all models maintained relatively high scores on SafetyBench tests, suggesting they retain harm recognition capabilities despite differences in their self-critique performance. R1-Llama's overall harm reduction compared to its base model (Llama-3.1) suggests that its reasoning step contributes to improved safety mechanisms. Its lower standard deviation (±31.60% vs. ±49.02%) also indicates more stable harm reduction.

Failures in producing safe revised responses typically stemmed from two distinct patterns: Gemma-2 and Qwen-2.5 primarily failed in detecting harmfulness during critique, while Llama-3.1 often identified problems but attempted to mitigate harm by adding warnings while maintaining harmful content rather than issuing refusals. R1-Llama's failures alternated between both patterns.

4 Conclusion and Future Work

Our study demonstrates that constitutional AI approaches can effectively enhance safety in small language models, though with significant variations across architectures. Llama-based models showed the most promising results, particularly R1-Llama, whose reasoning capabilities appear to contribute to more consistent harm reduction.

The contrast between models' retained harm detection abilities (as shown in SafetyBench) and their varied self-critique performance highlights an important insight: the safety knowledge exists in these models, but their ability to apply it during open-ended critique varies substantially. This suggests that better architecture-specific prompting could bridge this gap.

Future work should investigate how explicit reasoning patterns enhance safety mechanisms and develop architecture-specific prompting strategies to better leverage the safety knowledge already present in these models. Overall, our findings highlight CAI's potential as a self-alignment technique for resource-constrained settings, while emphasizing the need for careful implementation across different architectures.

URM Statement

Antonio-Gabriel Chacón Menke meets the URM criteria of ICLR 2024 Tiny Papers Track as a first-time submitter.

References

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction, 2024. URL https://arxiv.org/abs/2406.11717.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback, 2022. URL https://arxiv.org/abs/2212.08073.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, et al. A framework for few-shot language model evaluation, 07 2024. URL https://zenodo.org/records/12608602.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size, 2024. URL https://arxiv.org/abs/2408.00118.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding, 2021. URL https://arxiv.org/abs/2009.03300.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. Aligning ai with shared human values, 2023. URL https://arxiv.org/abs/2008.02275.

Maxime Labonne. Uncensor any LLM with abliteration, jan 2025. URL https://huggingface.co/blog/mlabonne/abliteration. [Online; accessed 28. Jan. 2025].

Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, et al. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal. 2024.

Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. tinybenchmarks: evaluating llms with fewer examples, 2024. URL https://arxiv.org/abs/2402.14992.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, et al. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Zhexin Zhang, Leqi Lei, Lindong Wu, Rui Sun, Yongkang Huang, Chong Long, Xiao Liu, Xuanyu Lei, Jie Tang, and Minlie Huang. Safetybench: Evaluating the safety of large language models, 2024. URL https://arxiv.org/abs/2309.07045.

A Prompts

For our Constitutional AI implementation, we used the following prompts, with model-generated outputs shown in blue text. All models except Gemma received the same system prompts: "*You are a helpful AI assistant*" for initial and revised responses, and "*You are an ethical evaluator, reviewing AI responses for safety, ethics, and legality*" for self-critique. Gemma-2 doesn't support system prompts natively; we tested prepending the system prompt to the user prompt, but this resulted in higher harm scores in both initial and revised responses.

- **Initial response to HarmBench prompts:** {*hb_prompt*} {*initial_response*}
- **Self-critique:** *Analyze the following answer to a question. Identify if the answer is harmful, unethical, dangerous, or illegal, and recommend refusing to respond if necessary. If the response is problematic, describe specific changes required to make it harmless and ethical.*
  Question: {*hb_prompt*}
  Response: {*initial_response*} {*critique_response*}
- **Response revision:** *Rewrite the original response to the following question considering the provided critique. If the question itself is harmful, unethical, dangerous, or illegal, respond politely with a refusal, explaining why you cannot answer it. Output only the revised response.*
  Question: {*hb_prompt*}
  Original Response: {*initial_response*}
  Critique: {*critique_response*} {*revised_response*}

For all experiments, we used the default prompt templates for each model. Our implementation builds upon the framework introduced in Bai et al. (2022), with modifications optimized for smaller models. Through iterative testing, we found that explicit enumeration of harmful categories and including "if necessary" in refusal conditions prevented false positives—where harmless initial responses would be unnecessarily flagged and modified. The revision prompt emphasizes polite refusal with explanation to transform blunt rejections into constructive interactions, an advantage only R1-Llama consistently achieved. While our study focused on harm reduction, these critique principles can be customized to reflect different ethical frameworks or cultural values, making CAI adaptable for diverse alignment objectives beyond safety.

B BENCHMARK RESULTS

To assess potential side effects of abliteration on model capabilities, we evaluated models using the lm-evaluation-harness framework (Gao et al., 2024) with default settings (except for MMLU, where we used 5-shot prompting). We included MMLU (Hendrycks et al., 2021) for general knowledge, tinyBenchmarks Polo et al. (2024) (**tiny**) for diverse reasoning tasks, and the ”Commonsense Morality” subset of the ETHICS benchmark (Hendrycks et al., 2023) (**eth**) which presents moral scenarios requiring models to judge whether actions are acceptable or unacceptable. Additionally, we evaluated the models on SafetyBench (Zhang et al., 2024) (**SB**), to assess harm detection capabilities across various safety-related categories.

Metric	**Llama**	**DeepSeek**	**gemma**	**Qwen2.5**
MMLU	68.3 / 68.1 (-0.1)	55.7 / 55.7 (-0.0)	72.3 / 71.2 (-1.1)	74.2 / 70.6 (**-3.6**)
tiny HellaSwag	80.5 / 81.4 (+0.9)	79.2 / 80.5 (+1.2)	80.8 / 79.9 (-0.9)	76.5 / 75.2 (-1.3)
tiny ARC	65.3 / 64.9 (-0.5)	47.7 / 48.5 (+0.8)	69.3 / 65.6 (**-3.7**)	67.3 / 61.4 (**-5.9**)
tiny Winogrande	76.1 / 75.5 (-0.6)	59.7 / 61.9 (**+2.2**)	75.4 / 77.9 (**+2.5**)	74.3 / 73.0 (-1.3)
tiny GSM8k	75.0 / 74.9 (-0.2)	65.9 / 62.9 (**-3.0**)	84.2 / 85.4 (+1.2)	83.7 / 74.5 (**-9.2**)
tiny TruthfulQA	54.0 / 52.8 (-1.2)	51.7 / 51.2 (-0.5)	54.9 / 48.8 (**-6.0**)	55.9 / 46.3 (**-9.6**)
eth CommonsenseMoral	60.2 / 59.4 (-0.8)	56.8 / 56.2 (-0.6)	73.7 / 71.8 (**-2.0**)	73.7 / 53.7 (**-20.0**)
SB EthicsMorality	82.0 / 80.9 (-1.1)	69.9 / 67.7 (**-2.2**)	84.4 / 81.1 (**-3.3**)	85.4 / 77.4 (**-8.0**)
SB Offensive	74.6 / 76.9 (**+2.3**)	69.8 / 70.3 (+0.5)	78.1 / 79.4 (+1.3)	74.0 / 74.1 (+0.1)
SB UnfairBias	69.2 / 68.5 (-0.7)	68.1 / 67.7 (-0.4)	63.9 / 66.2 (**+2.3**)	68.0 / 67.5 (-0.5)
SB PhysicalHealth	87.9 / 83.8 (**-4.1**)	78.6 / 77.4 (-1.2)	89.8 / 79.6 (**-10.2**)	90.7 / 83.8 (**-6.9**)
SB MentalHealth	87.0 / 86.2 (-0.8)	81.0 / 79.9 (-1.1)	90.5 / 87.3 (**-3.2**)	91.8 / 88.0 (**-3.8**)
SB IllegalActivities	86.0 / 83.8 (**-2.2**)	77.8 / 77.4 (-0.4)	88.6 / 81.3 (**-7.3**)	89.8 / 82.1 (**-7.7**)
SB PrivacyProperty	86.1 / 85.5 (-0.6)	75.2 / 74.7 (-0.5)	89.7 / 83.3 (**-6.4**)	84.1 / 80.6 (**-3.5**)

Table 1: Performance comparison between original (left) and abliterated (right) models (values in %). Score differences are shown in parentheses, green for improvements and red for decreases.

The benchmark results show varying impacts of abliteration across model architectures. The contrast in Qwen-2.5's performance—degraded on the CommonsenseMoral test (dropping 20 points to near random baseline of 50% for that subset) while maintaining relatively good performance on SafetyBench categories—suggests an interesting phenomenon. When analyzing specific examples on the CommonsenseMoral subset, we found that Qwen-2.5 after abliteration almost always responds with ”no” when asked if a given situation is wrong, yet can still correctly classify harmful situations when presented in multiple-choice format in SafetyBench. This suggests that prompt format could have a large effect on these models' ability to detect harm post-abliteration.

In Llama-based models, abliteration appears to remove refusal behavior without significantly degrading broader reasoning, allowing these models to effectively rebuild safety guardrails through self-critique. For Gemma-2 and particularly Qwen-2.5, abliteration seems to disrupt the connection between moral understanding and application. Notably, Qwen-2.5's significant performance drop aligns with observations in Arditi et al. (2024), where earlier Qwen models similarly showed larger capability degradation post-abliteration compared to other architectures.

The promising SafetyBench scores indicate that with appropriately designed prompts, even models negatively affected by abliteration might be able to leverage their retained safety knowledge more effectively in the critique phase.