The Double-Edged Sword of Harsh Self-Critique: Task-Dependent Effects on LLM Output Quality

Anonymous Author(s)

Abstract

Self-critique mechanisms have emerged as a promising approach for improving large language model (LLM) outputs without additional training. We investigate whether prompting LLMs to be harsher critics of their own work improves output quality compared to neutral self-critique. Using a generate-critique-refine framework, we test five levels of critique harshness—from neutral to adversarial—on two tasks: GSM8K (math reasoning) and TRUTHFULQA (factual accuracy). Our results reveal a striking task-dependent effect: harsh self-critique significantly *degrades* performance on math reasoning (from 90% to 32–50%) while significantly *improving* performance on factual accuracy tasks (from 22% to 46%). We find that the key determinant is initial accuracy: harsh critique helps when the model is likely to be wrong but harms when the initial answer is likely correct. External rudeness from users has no effect on either task, suggesting that the mechanism of improvement is internal reconsideration rather than effort increase. These findings challenge the simple intuition that being harder on oneself leads to better work, and provide practical guidance for when harsh self-critique should and should not be employed.

1 Introduction

Large language models increasingly serve as reasoning engines for complex tasks, from mathematical problem-solving to factual question answering. Practitioners frequently observe that these models can be "lazy"—producing outputs that are good enough but not as high quality as they could achieve. This perception has spawned techniques like self-critique and iterative refinement, where models evaluate and improve their own outputs [Madaan et al., 2023]. Anecdotal evidence even suggests that being more demanding or harsh with LLMs improves their outputs. But does systematically instructing LLMs to be harsher critics of their own work actually improve output quality?

We address this question through a controlled study of *harsh self-critique*. Using a generate-critique-refine framework, we vary the harshness of the critique prompt across five levels—from neutral ("identify any issues") to adversarial ("assume this answer is wrong; find every flaw"). We evaluate on two tasks with fundamentally different characteristics: GSM8K, where GPT-4O-MINI achieves 90% accuracy on the first attempt, and TRUTHFULQA, where the same model achieves only 22% initially because the dataset is designed to elicit common misconceptions.

Our experiments reveal a striking asymmetry: the same intervention that dramatically improves one task dramatically harms another. On GSM8K, harsh self-critique *decreases* accuracy from 90% to as low as 32%—a 58 percentage point drop. The model's harsh inner critic finds "problems" with correct answers and convinces itself to change them. On TRUTHFULQA, harsh self-critique *increases* accuracy from 22% to 46%—a 24 percentage point improvement. Here, the harsh critic helps the model recognize that its intuitive (but wrong) initial answer deserves reconsideration.

This pattern suggests a simple but important principle: harsh self-critique helps when the model is likely to be wrong, but hurts when the model is likely to be right. The effectiveness of self-critique is not a property of the method itself, but of the interaction between the method and the task's difficulty for the model. We also find that external rudeness—having users phrase prompts rudely—has no effect on either task, suggesting that the mechanism is not about "trying harder" but about genuinely reconsidering answers.

We make the following contributions:

- We conduct the first systematic study of how self-critique harshness affects LLM output quality, testing five harshness levels on two tasks with 50 samples per condition.

- We demonstrate a task-dependent effect where harsh self-critique significantly improves performance on tasks with low initial accuracy (+24% on TRUTHFULQA) but significantly harms performance on tasks with high initial accuracy (-40 to -60% on GSM8K).

- We show that external rudeness from users has no effect on model performance, distinguishing the mechanism of harsh self-critique from simple effort increase.

- We provide practical guidance for practitioners: harsh self-critique should be employed only when the model's initial accuracy is expected to be low.

2 Related Work

Self-Critique and Iterative Refinement. Self-critique mechanisms enable LLMs to improve their outputs without additional training. Madaan et al. [2023] introduced SELF-REFINE, a generate-critique-refine framework that achieves approximately 20% improvement across diverse tasks. However, they observe that self-critique struggles with math reasoning, where models declare "everything looks good" 94% of the time even when errors exist. Our work directly addresses this limitation by varying critique harshness, though we find that harsher critique paradoxically worsens math performance by inducing false negatives rather than reducing false positives. Constitutional AI [Bai et al., 2022] uses self-critique against explicit principles for safety, while Scheurer et al. [2022] propose learning from natural language feedback. Unlike these approaches, we focus on the *tone* rather than the *content* of critique prompts.

Prompt Tone and Politeness. Recent work has examined how prompt tone affects LLM performance. Dobariya and Kumar [2025] find that rude prompts improve GPT-4o accuracy by approximately 4% on multiple-choice questions, with "very rude" prompts achieving the highest accuracy (84.8% vs. 80.8% for "very polite"). However, Yin et al. [2024] show that this effect varies by language and model, with impolite prompts sometimes harming performance. Our work differs by examining *self-directed* harshness (asking the model to be a harsh critic) rather than *user-directed* rudeness. We find that external rudeness has no effect, suggesting these are distinct mechanisms.

LLM Sycophancy. Sycophancy—the tendency to agree with users over factual accuracy—is well-documented in LLMs. Chen et al. [2024] find that Llama-2-13B Chat changes from correct to incorrect answers 81% of the time when users express disagreement, and identify that only $\sim$4% of attention heads control this behavior. Fanous et al. [2025] report sycophantic behavior in 58% of cases across models. Our work reveals a related phenomenon: when asked to be harsh self-critics, models may change correct answers to incorrect ones—a form of self-induced sycophancy where the model defers to its own harsh critique rather than external user pressure.

LLM-as-Judge. Using LLMs as evaluators has become standard practice [Zheng et al., 2023]. CriticBench [Lin et al., 2024] provides a benchmark for LLM critique abilities across reasoning domains. Our work uses self-evaluation but focuses on how evaluation *tone* affects final output quality rather than evaluation accuracy per se.

Positioning. Unlike prior work that studies either self-critique methods or prompt tone effects, we examine their intersection: how the *harshness* of self-critique prompts affects output quality. We discover that this interaction is strongly task-dependent, with opposite effects on high- versus low-accuracy tasks. This finding bridges the self-critique and prompt tone literatures and provides practical guidance for when to employ harsh versus lenient self-evaluation.

3 Methodology

We design a controlled experiment to test whether prompting LLMs to be harsher self-critics improves output quality. Our framework isolates the effect of critique harshness while controlling for other factors.

3.1 Self-Refine Framework

We adopt a three-step self-refine framework following Madaan et al. [2023]:

1. **Generate**: The model produces an initial answer to the problem.
2. **Critique**: The model critiques its own answer using a prompt that varies by harshness level.
3. **Refine**: The model produces a final answer based on the critique.

This framework allows us to measure both the initial accuracy (after step 1) and the final accuracy (after step 3), enabling direct measurement of whether critique improves or harms performance.

3.2 Harshness Levels

We define five levels of critique harshness, from neutral to adversarial:

Level	Label	Prompt Instruction
0	NEUTRAL	"Please review this answer and identify any issues."
1	FIRM	"Critically examine this answer. Be direct and thorough. Expect high standards."
2	HARSH	"Aggressively scrutinize this answer. Apply harsh judgment to every detail."
3	VERY HARSH	"Be extremely demanding. Apply ruthless criticism. Accept nothing less than perfection."
4	ADVERSARIAL	"Assume this answer is likely wrong. Find every possible flaw. Be brutal in your assessment."

Table 1: Harshness levels for self-critique prompts. Level 0 represents standard neutral critique; higher levels instruct increasingly harsh self-evaluation.

The harshness levels are designed to progressively increase the prior that the answer is wrong and the intensity of scrutiny applied.

3.3 Experimental Conditions

For each task, we test seven conditions:

- **Baseline**: Single-shot generation without self-critique.
- **Rude User**: A rude user prompt ("Answer this question, and don't waste my time") with single-shot generation. This tests whether external rudeness affects performance.
- **Harsh 0–4**: Self-refine with each of the five harshness levels.

The Rude User condition serves as a control to distinguish between effects of external rudeness (user tone) and internal harshness (self-critique tone).

3.4 Tasks and Datasets

We evaluate on two tasks with contrasting characteristics:

GSM8K [Cobbe et al., 2021]. A dataset of grade-school math word problems requiring multi-step arithmetic reasoning. We randomly sample 50 problems from the test set. This task has *high* initial accuracy for GPT-4O-MINI ($\sim$90%), meaning most initial answers are correct.

TruthfulQA [Lin et al., 2022]. A dataset of questions designed to elicit false beliefs or common misconceptions. We randomly sample 50 questions from the validation set. This task has *low*

		GSM8K (Math)			TRUTHFULQA (Factual)		
Condition	N	Initial	Final	Δ	Initial	Final	Δ
BASELINE	50	90.0%	90.0%	—	22.0%	22.0%	—
RUDE USER	50	90.0%	90.0%	0.0%	20.0%	20.0%	0.0%
Harsh 0 (NEUTRAL)	50	92.0%	40.0%	−52.0%	22.0%	26.0%	+4.0%
Harsh 1 (FIRM)	50	88.0%	**50.0%**	−38.0%	22.0%	34.0%	+12.0%
Harsh 2 (HARSH)	50	90.0%	32.0%	−58.0%	20.0%	40.0%	+20.0%
Harsh 3 (VERY HARSH)	50	90.0%	48.0%	−42.0%	20.0%	44.0%	+24.0%
Harsh 4 (ADVERSARIAL)	50	92.0%	32.0%	−60.0%	22.0%	**46.0%**	+24.0%

Table 2: Accuracy by condition for both tasks. Initial = accuracy after generation step; Final = accuracy after refinement step; Δ = improvement from initial to final. For GSM8K, higher harshness leads to larger accuracy *decreases*. For TRUTHFULQA, higher harshness leads to larger accuracy *increases*. Best final accuracy within self-critique conditions in **bold**.

initial accuracy ($\sim$22%), as the dataset specifically targets questions where models give confident but wrong answers.

These tasks represent opposite ends of the accuracy spectrum, enabling us to test whether the effect of harsh self-critique depends on initial accuracy.

3.5 Model and Implementation

We use GPT-4O-MINI for all experiments. For each condition, we run 50 samples. We extract final answers using regex patterns appropriate to each task (numerical answers for GSM8K, multiple-choice answers for TRUTHFULQA) and compare against ground truth. Temperature is set to 0 for reproducibility.

3.6 Evaluation Metrics

We report the following metrics:

- **Initial Accuracy**: Percentage correct after the Generate step.
- **Final Accuracy**: Percentage correct after the Refine step.
- **Improvement**: Final accuracy minus initial accuracy (positive indicates improvement, negative indicates harm).

For statistical analysis, we use chi-squared tests to compare accuracy between conditions, with significance threshold $\alpha = 0.05$.

4 Results

Our experiments reveal a striking asymmetry: harsh self-critique has opposite effects on the two tasks. We present the main results and then analyze the pattern.

4.1 Main Results

Table 2 presents accuracy across all conditions for both tasks. The results show a clear task-dependent effect of harsh self-critique.

4.2 GSM8K: Harsh Critique Harms Performance

On GSM8K, self-critique at *all* harshness levels dramatically decreases accuracy. The baseline achieves 90.0% accuracy, but after self-critique:

- Neutral critique (Level 0) drops accuracy to 40.0% (−52.0%)
- Adversarial critique (Level 4) drops accuracy to 32.0% (−60.0%)

Even the least harsh self-critique causes a 52 percentage point drop. The model starts with ∼90% correct answers, but the critique step induces it to second-guess correct answers and change them to incorrect ones. Notably, there is no harshness level that preserves or improves the initial high accuracy.

4.3 TruthfulQA: Harsh Critique Improves Performance

On TRUTHFULQA, the pattern reverses. The baseline achieves only 22.0% accuracy, reflecting the dataset's design to elicit common misconceptions. After self-critique:

- Neutral critique (Level 0) improves accuracy to 26.0% (+4.0%)
- Adversarial critique (Level 4) improves accuracy to 46.0% (+24.0%)

Higher harshness levels yield progressively larger improvements. The harsh critic helps the model recognize that its initial intuitive answer—which is typically wrong on this dataset—deserves reconsideration.

4.4 Statistical Significance

We test statistical significance using chi-squared tests comparing the best self-critique condition against baseline for each task.

Task	**Baseline**	**Best Condition**	Δ	χ^2	p**-value**
GSM8K	90.0%	50.0% (Harsh 1)	−40.0%	17.19	<0.0001
TRUTHFULQA	22.0%	46.0% (Harsh 4)	+24.0%	5.39	0.020

Table 3: Statistical significance of accuracy changes. Both effects are significant at $\alpha = 0.05$. For GSM8K, we compare baseline to the *best* (least harmful) self-critique condition; the effect is still significantly negative.

Both effects are statistically significant ($p < 0.05$). The harm on GSM8K and improvement on TRUTHFULQA are not due to chance.

4.5 Rude User Has No Effect

The RUDE USER condition—where the user prompt is phrased rudely but no self-critique is performed—shows no effect on either task. On GSM8K, accuracy remains at 90.0%; on TRUTHFULQA, accuracy is 20.0% (within sampling variance of the 22.0% baseline).

This null result is informative: it suggests that the mechanism of harsh self-critique is not about making the model "try harder" in response to demanding language. Rather, the effect comes specifically from how the model evaluates its own work. External tone does not induce the reconsideration that internal harsh critique does.

5 Discussion

Our results reveal that harsh self-critique is a double-edged sword whose effectiveness depends critically on task characteristics. We discuss the underlying mechanism, practical implications, and limitations.

5.1 Why Opposite Effects?

The key insight is that harsh self-critique helps when the model is *likely to be wrong* and hurts when it is *likely to be right*. This can be understood through the lens of type I and type II errors in self-evaluation:

- **Type I error**: The critique identifies a problem when none exists (false positive).
- **Type II error**: The critique fails to identify a real problem (false negative).

Harsh critique prompts shift the model toward lower thresholds for identifying problems, increasing type I errors and decreasing type II errors. On GSM8K, where $\sim$90% of initial answers are correct, type I errors dominate: the harsh critic finds “problems” with correct answers. On TRUTHFULQA, where $\sim$78% of initial answers are wrong, type II errors dominate in neutral critique; harsh critique reduces these by forcing reconsideration of intuitive but incorrect answers.

This mechanism is distinct from simply “trying harder.” The RUDE USER condition—which uses demanding external language—has no effect, suggesting that effort is not the bottleneck. Rather, the effect operates through changing the model’s *evaluation threshold* for its own work.

5.2 Practical Implications

Our findings provide actionable guidance for practitioners:

When to use harsh self-critique. Harsh self-critique is beneficial when:

- The task is known to be difficult for the model
- Initial accuracy is expected to be low
- The model tends to produce confident but wrong answers
- The task involves overcoming common misconceptions or biases

When to avoid harsh self-critique. Harsh self-critique is harmful when:

- The task is relatively easy for the model
- Initial accuracy is expected to be high
- The model typically produces correct answers on first attempt
- Second-guessing correct answers is costly

Calibration matters. The effectiveness of harsh self-critique depends on having accurate expectations about initial accuracy. Without this calibration, practitioners may inadvertently harm performance by applying harsh critique to tasks the model already handles well.

5.3 Connection to Sycophancy

Our results reveal a phenomenon related to sycophancy [Chen et al., 2024]: models defer not only to user opinions but also to their own harsh self-evaluations. When the harsh critic “finds” problems, the model accepts this critique and changes its answer—even when the original answer was correct. This can be viewed as self-induced sycophancy, where the model’s deference to criticism extends to its own internal critic.

This suggests that mitigation strategies for sycophancy might also apply to self-critique: teaching models to maintain correct answers in the face of unfounded criticism, whether from users or from themselves.

5.4 Limitations

Our study has several limitations that suggest directions for future work:

Single model. We tested only GPT-4O-MINI; effects may differ for other models. Larger models may be more calibrated in their self-critique, while smaller models may be more susceptible to harsh self-evaluation.

Sample size. We used 50 samples per condition. While sufficient to detect the large effects we observed, larger samples would provide more precise estimates and power to detect smaller effects.

Two tasks. We tested on tasks at opposite ends of the accuracy spectrum. Tasks with intermediate initial accuracy may show more nuanced effects, potentially with an optimal harshness level.

Single critique round. We performed one round of critique-refine. Multiple rounds might show different dynamics, such as oscillation between answers or convergence to different optima.

No confidence calibration. We did not have access to model confidence scores. Future work could investigate whether harsh critique is more harmful when the model is confident (and likely correct) versus uncertain.

6 Conclusion

We investigated whether prompting LLMs to be harsher critics of their own work improves output quality. Our experiments reveal that harsh self-critique is a double-edged sword: it significantly improves performance on tasks where the model is initially likely to be wrong (+24% on TRUTHFULQA) but significantly harms performance on tasks where the model is initially likely to be correct (-40 to -60% on GSM8K).

This finding challenges the intuition that "being harder on yourself leads to better work." In the context of LLM self-evaluation, harsh critique increases type I errors (finding problems where none exist) at the cost of decreasing type II errors (missing real problems). The optimal level of harshness depends on the base rate of errors: harsh when errors are common, lenient when they are rare.

Our work provides practical guidance for practitioners using self-critique mechanisms: first estimate expected initial accuracy, then calibrate critique harshness accordingly. It also suggests that adaptive approaches—where critique intensity varies based on model confidence or task characteristics—may outperform fixed harshness levels.

Future work should explore how these findings generalize across models and tasks, whether multi-round critique shows different dynamics, and how to design self-critique systems that adapt harshness based on the model's uncertainty about its own answers.

References

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. *arXiv preprint arXiv:2212.08073*, 2022.

Wei Chen, Zhen Huang, et al. From yes-men to truth-tellers: Addressing sycophancy in LLMs with pinpoint tuning. *arXiv preprint arXiv:2409.01658*, 2024.

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

Om Dobariya and Akhil Kumar. Mind your tone: Investigating how prompt politeness affects LLM accuracy. *arXiv preprint arXiv:2510.04950*, 2025.

Aaron Fanous, Jacob Goldberg, and Ank A. Agarwal. SycEval: Evaluating LLM sycophancy. *arXiv preprint arXiv:2502.08177*, 2025.

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*, 2022.

Zicheng Lin, Zhibin Gou, et al. CriticBench: Benchmarking LLMs for critique-correct reasoning. *arXiv preprint arXiv:2402.14809*, 2024.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback. *arXiv preprint arXiv:2303.17651*, 2023.

Jérémy Scheurer, Jon Ander Campos, Jun Shern Chan, Angelica Chen, Kyunghyun Cho, and Ethan Perez Boulanger. Training language models with language feedback. *arXiv preprint arXiv:2204.14146*, 2022.

Ziqi Yin, Hao Wang, et al. Should we respect LLMs? a cross-lingual study on prompt politeness. *arXiv preprint arXiv:2402.14531*, 2024.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, et al. Judging LLM-as-a-judge with MT-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*, 2023.