Harsh Self-Critique and Budget Control Do Not Fix Lazy LLM Outputs

Anonymous Authors

Abstract

Large language models (LLMs) often produce short, shallow answers on reasoning tasks, and prompting is the cheapest lever to reduce this behavior. We examine whether harsher self-critique prompts and response-budget control improve reasoning accuracy compared to standard prompts. Using GSM8K and ARC-CHALLENGE test splits, we evaluate six prompt conditions on a fixed GPT-4.1 model with deterministic decoding and matched examples (n=50 per dataset). We measure accuracy, response length, and paired bootstrap confidence intervals versus a direct baseline. COT improves GSM8K accuracy from 0.52 to 0.74 (+0.22, 95% CI [0.08, 0.36]) and slightly improves ARC-CHALLENGE (0.88 to 0.90, +0.02, CI [-0.04, 0.08]). In contrast, HARSHCRITIC reduces accuracy on both datasets (GSM8K: 0.40; ARC-CHALLENGE: 0.32), and HARSHCRITIC+BUDGET performs worst (GSM8K: 0.14; ARC-CHALLENGE: 0.24). Rude and polite framing yield near-zero effects. These results show that harsh self-critique with budget control is not a reliable fix for lazy outputs; standard COT remains the most robust low-effort improvement in this setting.

1 Introduction

Why do lazy outputs matter? Reasoning benchmarks often require multi-step inference, yet LLMs frequently answer with short, shallow responses that miss key steps or rely on shortcuts. This behavior reduces reliability in settings where users need correct reasoning and verifiable answers. Evidence from in-context learning shows that models can exploit superficial cues instead of robust reasoning, a behavior described as "lazy learners" in prior work Tang et al. [2023].

What is missing? Prior prompting work shows that explicit reasoning, structured decomposition, or self-consistency can improve accuracy Wei et al. [2022], Zhou et al. [2023], Wang et al. [2023b]. Self-critique and refinement loops can also help in some tasks Madaan et al. [2023], Shinn et al. [2023], Wang et al. [2023a]. However, there is limited controlled evidence on whether making the critic harsher or constraining response budgets actually improves reasoning accuracy on standard benchmarks.

What do we do? We run a controlled prompt study that varies critic severity, response budget, and tone while holding model, data, and decoding fixed. Using GPT-4.1 on GSM8K and ARC-CHALLENGE (n=50 each), we compare six prompt conditions: DIRECT, COT, HARSHCRITIC, HARSHCRITIC+BUDGET, RUDE-DIRECT, and POLITE-DIRECT. We measure accuracy, response length, and paired bootstrap confidence intervals versus the direct baseline.

What do we find? COT provides the most reliable gain, improving GSM8K by +0.22 (95% CI [0.08, 0.36]) and slightly improving ARC-CHALLENGE by +0.02 (CI [-0.04, 0.08]). In contrast, HARSHCRITIC and HARSHCRITIC+BUDGET substantially reduce accuracy on both datasets, with the low-budget critic condition dropping GSM8K by -0.38 (CI [-0.52, -0.22]) and ARC-CHALLENGE by -0.64 (CI [-0.78, -0.50]). Tone changes have near-zero effects.

We make three contributions:

- We propose a controlled evaluation of critic harshness and response-budget control on GSM8K and ARC-CHALLENGE using a fixed model and deterministic decoding.
- We report accuracy, response length, and paired bootstrap confidence intervals showing that harsh-critic prompting hurts accuracy while CoT remains robust.
- We provide a focused discussion of limitations and follow-up directions for improving effort-inducing prompts.

2 Related Work

Effort-inducing prompting. Chain-of-thought prompting improves reasoning by eliciting intermediate steps Wei et al. [2022], and least-to-most prompting decomposes complex problems into simpler subproblems Zhou et al. [2023]. These methods motivate our CoT baseline and highlight that prompt structure can change reasoning behavior.

Decoding and selection for reasoning. Self-consistency aggregates multiple reasoning traces to select the most consistent answer Wang et al. [2023b], while self-evaluation guided beam search uses evaluation signals during decoding Xie et al. [2023]. Our study keeps decoding fixed to isolate the impact of prompt-level interventions.

Self-critique and refinement. Self-refinement loops can improve outputs across tasks Madaan et al. [2023], and Reflexion uses self-reflection memory to improve agent performance Shinn et al. [2023]. Dual-critique prompting studies how critique can reduce errors under inductive instructions Wang et al. [2023a]. Unlike these approaches, we focus on a single-pass critic prompt and explicitly test whether critic harshness and budget control improve reasoning accuracy.

Lazy learners and shortcut reliance. Tang et al. show that LLMs can rely on prompt shortcuts, leading to performance collapse on anti-shortcut data Tang et al. [2023]. This motivates our focus on prompt interventions that might reduce shallow or shortcut-driven outputs.

3 Methodology

Problem setup. We test whether prompt-level interventions that increase self-critique severity and constrain response budgets improve reasoning accuracy relative to a direct baseline. We hold the model, data, and decoding constant and vary only the prompt framing and token budget.

Datasets and sampling. We use the GSM8K and ARC-CHALLENGE test splits (1,319 and 1,172 examples, respectively). For cost control, we sample a fixed evaluation subset of 50 examples per dataset with seed 42. For GSM8K, final answers are parsed using the canonical `####` delimiter.

Prompt conditions. We evaluate six conditions on the same examples: DIRECT (baseline), CoT, HARSHCRITIC, HARSHCRITIC+BUDGET (harsh critic with reduced budget), RUDE-DIRECT, and POLITE-DIRECT. The harsh-critic prompts instruct the model to critique and revise its answer; the low-budget variant reduces the maximum output tokens.

Model and decoding. We query GPT-4.1 via the OpenAI API with temperature 0.0. The baseline uses a maximum of 256 output tokens, while HARSHCRITIC+BUDGET uses 128. All other parameters are held fixed. Inference is API-based; available GPUs were not used.

Evaluation metrics. We report exact-match accuracy for GSM8K and multiple-choice letter accuracy for ARC-CHALLENGE. We also compute average response length (word count) as a proxy for effort. For statistical comparisons, we report paired bootstrap confidence intervals for accuracy differences versus DIRECT.

4 Results

Main accuracy results. Table 1 shows accuracy for all prompt conditions. CoT is the strongest overall improvement, lifting GSM8K from 0.52 to 0.74, while RUDE-DIRECT yields the highest ARC-CHALLENGE accuracy (0.92). Both harsh-critic variants reduce accuracy substantially on both datasets.

Condition	GSM8K Acc.	ARC-CHALLENGE Acc.
DIRECT	0.52	0.88
COT	**0.74**	0.90
HARSHCRITIC	0.40	0.32
HARSHCRITIC+BUDGET	0.14	0.24
RUDE-DIRECT	0.48	**0.92**
POLITE-DIRECT	0.48	0.88

Table 1: Accuracy on GSM8K and ARC-CHALLENGE (n=50 each). Best results per dataset are in **bold**.

Condition	GSM8K Diff vs. DIRECT	ARC-CHALLENGE Diff vs. DIRECT
COT	+0.22 [0.08, 0.36]	+0.02 [-0.04, 0.08]
HARSHCRITIC	-0.12 [-0.26, 0.02]	-0.56 [-0.72, -0.40]
HARSHCRITIC+BUDGET	-0.38 [-0.52, -0.22]	-0.64 [-0.78, -0.50]
RUDE-DIRECT	-0.04 [-0.12, 0.04]	+0.04 [0.00, 0.10]
POLITE-DIRECT	-0.04 [-0.12, 0.04]	0.00 [0.00, 0.00]

Table 2: Paired bootstrap confidence intervals (95%) for accuracy differences versus DIRECT.

Statistical comparisons. Table 2 reports paired bootstrap confidence intervals relative to DIRECT. COT provides a reliable gain on GSM8K (+0.22, 95% CI [0.08, 0.36]) and a small, non-significant change on ARC-CHALLENGE. In contrast, HARSHCRITIC and HARSHCRITIC+BUDGET show large negative deltas on both datasets, and the low-budget critic condition is the worst overall.

Length-effort trade-offs. Direct prompts are extremely short (about 2 words on average). COT responses are longer (roughly 103–123 words) and more accurate. HARSHCRITIC produces the longest responses (roughly 155–191 words) while still reducing accuracy, and HARSHCRITIC+BUDGET reduces length (about 85–99 words) but further harms accuracy. This pattern suggests that longer outputs are not inherently better when the critique signal is weak.

Figures. Figure 1 and Figure 2 visualize the accuracy pattern across conditions, and Figure 3 shows the question length distribution for the evaluation samples.

5 Discussion

Why does harsh critique hurt? The harsh-critic conditions produce longer responses but lower accuracy, which suggests that the critic prompt can introduce unnecessary edits or distract from the original reasoning. Without an external error signal, the model may revise correct answers into incorrect ones. This aligns with the intuition that critique is most effective when it is grounded in verifiable feedback rather than tone alone.

Limits of tone changes. Rude and polite framing yield near-zero differences across both datasets. This indicates that tone is not a stable lever for reasoning quality under controlled conditions, despite common anecdotes that rudeness improves performance.

Limitations. Our evaluation uses small subsets (n=50 per dataset) and a single model (GPT-4.1), so the estimates have wide uncertainty and may not generalize to other models or tasks. We do not test multi-round self-refine or self-consistency, and we do not measure human preference or perceived "laziness" beyond word count.

Broader implications. Prompt-level interventions are attractive because they are cheap and easy to deploy, but our results show that harsher critique and strict budgets can degrade performance. Practitioners should validate prompt changes empirically rather than rely on intuition about effort or tone.

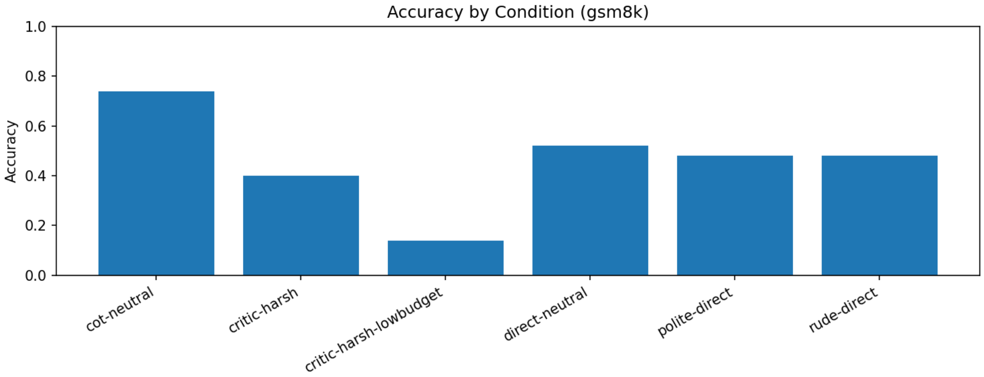

Figure 1: Accuracy by condition on GSM8K (n=50). CoT improves accuracy, while harsh-critic variants reduce performance.

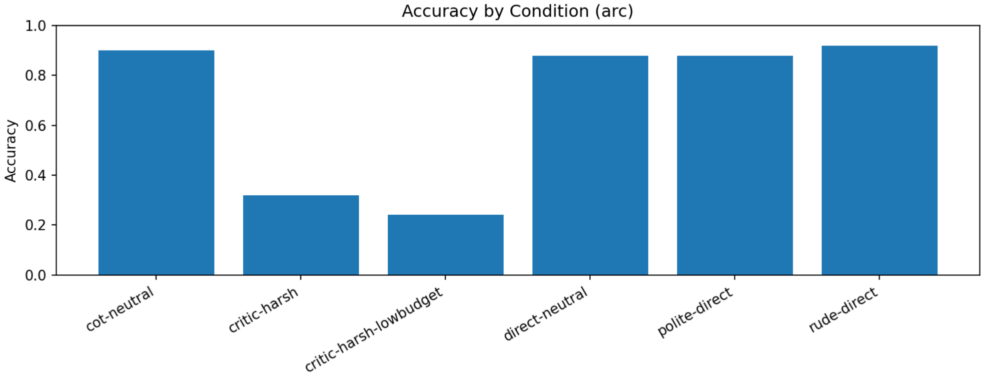

Figure 2: Accuracy by condition on ARC-Challenge (n=50). Harsh-critic prompting sharply reduces accuracy despite longer outputs.

6 Conclusion

We evaluated harsh self-critique prompting and response-budget control on GSM8K and ARC-Challenge with a fixed GPT-4.1 model. The main result is negative: harsh-critic prompting reduces accuracy, and the low-budget variant performs worst. In contrast, CoT remains the most reliable low-effort improvement. The key takeaway is that harsher self-critique does not fix lazy LLM outputs in this setting and can actively harm accuracy.

Future work should test self-consistency and least-to-most prompting on the same samples, run multi-round self-refine to see if iterative feedback recovers the critic losses, and incorporate human or LLM-judge evaluations of output quality beyond accuracy.

References

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.

Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *arXiv preprint arXiv:2303.11366*, 2023.

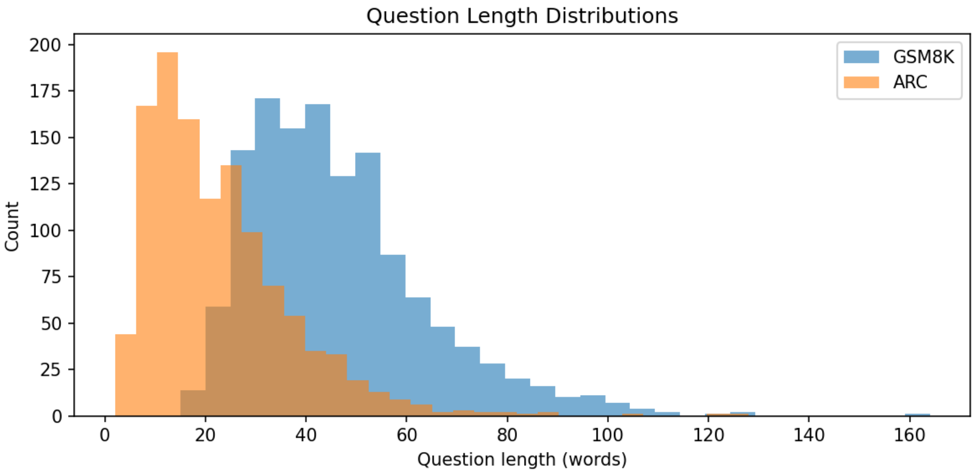

Figure 3: Question length distribution for the evaluation samples.

Ruixiang Tang, Dehan Kong, Longtao Huang, and Hui Xue. Large language models can be lazy learners: Analyze shortcuts in in-context learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4645–4657, 2023.

Rui Wang, Hongru Wang, Fei Mi, Boyang Xue, Yi Chen, Kam-Fai Wong, and Ruifeng Xu. Enhancing large language models against inductive instructions with dual-critique prompting. *arXiv preprint arXiv:2305.13733*, 2023a.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations (ICLR)*, 2023b. arXiv:2203.11171.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*, 2022.

Yuxi Xie, Kenji Kawaguchi, Yiran Zhao, James Xu Zhao, Min-Yen Kan, Junxian He, and Michael Qizhe Xie. Self-evaluation guided beam search for reasoning. *arXiv preprint arXiv:2305.00633*, 2023.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V. Le, and Ed H. Chi. Least-to-most prompting enables complex reasoning in large language models. In *International Conference on Learning Representations (ICLR)*, 2023. arXiv:2205.10625.