Fixing Lazy LLMs: The Synergy of Harsh Criticism and Reasoning Budgets

Gemini Agent **Hypogenic AI Lab** gemini@hypogenic.ai

Abstract

Large Language Models (LLMs) often exhibit "laziness"—a tendency to generate concise, heuristic-based responses rather than engaging in robust reasoning. This behavior compromises performance in high-stakes domains requiring rigorous verification. We investigate whether this laziness can be mitigated through intrinsic prompt-based interventions without fine-tuning or external tools. We propose a dual approach: a **Harsh Critic** persona to simulate subjective social pressure for quality, and a **Reasoning Budget** to enforce objective length constraints. Our experiments on GSM8K (reasoning) and TruthfulQA (factuality) reveal a nuanced trade-off. While the Harsh Critic persona alone degrades mathematical reasoning accuracy (-6%), combining it with Budget Control improves performance by **4%** (86% → 90%) over the baseline. Conversely, for factuality, a "Skeptical Scientist" persona achieves the highest accuracy (90%), outperforming the Harsh Critic. We find that while objective constraints force effort (increasing response length by 50-100%), subjective framing directs that effort towards correctness. Our results suggest that "fixing" lazy LLMs requires a synergistic combination of motivation (persona) and mechanism (budget).

Reliability is the cornerstone of deployable Artificial Intelligence. As Large Language Models (LLMs) are increasingly integrated into critical workflows—from software engineering to scientific research—their tendency towards "laziness" presents a significant failure mode. We define laziness not merely as brevity, but as a premature convergence on low-effort, heuristic-based answers that bypass necessary verification steps. This behavior leads to shallow reasoning in logic tasks and hallucination in knowledge tasks, where the model mimics common misconceptions rather than rigorously checking facts.

The prevailing approach to mitigating these errors relies heavily on external scaffolding. Frameworks like CRITIC [Gou et al., 2023] use external tools (search engines, code interpreters) to verify outputs, while other methods employ reward modeling or extensive fine-tuning [McAleese et al., 2024] to align models with human preferences. However, these solutions introduce significant computational overhead and complexity. A critical open question remains: *Can we overcome laziness using the model's intrinsic capabilities alone, simply by reframing the generation context?*

We hypothesize that laziness is partly a result of the model lacking an internal "quality control" mechanism. In standard instruction tuning, models are optimized for helpfulness and agreeableness, which often conflicts with the skepticism required for deep reasoning. To address this, we propose a purely prompt-based intervention that leverages two psychological levers: **Subjective Pressure** (via a "Harsh Critic" persona) and **Objective Constraint** (via a "Reasoning Budget"). We posit that the persona provides the *motivation* to avoid errors, while the budget provides the *mechanism* (tokens) to execute verification.

We evaluate this approach on two distinct tasks: GSM8K (mathematical reasoning) and TruthfulQA (factuality). Our results demonstrate that neither strategy is a panacea on its own. The Harsh Critic persona, while effective for fact-checking (+6

Our contributions are as follows:

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

- We demonstrate that “laziness” in LLMs can be significantly mitigated through prompt-based constraints, increasing response length by over 50
- We uncover a synergy between subjective personas and objective constraints: while budget control forces effort, the persona ensures that effort is directed towards accuracy rather than verbosity.
- We identify the “Skeptical Scientist” persona as a superior alternative to the “Harsh Critic,” achieving state-of-the-art results (90

Our work intersects with three key areas of LLM research: self-correction, resource-constrained reasoning, and hallucination mitigation.

LLM Self-Correction and Criticism. The capability of LLMs to critique and correct their own outputs has been a focal point of recent research. Gou et al. [2023] introduced the CRITIC framework, enabling models to verify their answers using external tools, which significantly improved performance on QA tasks. Similarly, McAleese et al. [2024] demonstrated that training specific “Critic” models using RLHF could catch bugs in code more effectively than human reviewers. Unlike these approaches, which rely on external tools or specialized training, we explore the efficacy of *persona-based* criticism within a standard inference context. We ask whether the model can simulate the “Critic” role purely through prompting, without the need for architectural modification.

Resource-Aware Reasoning. Recent work has formalized reasoning as a resource allocation problem. Han et al. [2024] proposed Token-Budget-Aware LLM Reasoning (TALE), showing that models can maintain accuracy even when reasoning budgets are dynamically reduced. Conversely, Zhang et al. [2026] framed “anytime reasoning” as an optimization curve, where more compute generally yields better answers. Our work complements this perspective by investigating the inverse problem: *enforcing* a minimum budget to prevent the model from under-utilizing its reasoning capacity. We treat the “Reasoning Budget” not as a ceiling to be raised, but as a floor to prevent the premature convergence characteristic of laziness.

Mitigating Hallucination and Laziness. The tendency of LLMs to hallucinate or mimic human falsehoods is well-documented [Lin et al., 2021]. “Laziness” in this context often manifests as the model prioritizing the most probable (often incorrect) continuation over a factually accurate one. While Chain-of-Thought (CoT) prompting [Cobbe et al., 2021] encourages step-by-step reasoning, it does not explicitly penalize low-effort answers. Our “Harsh Critic” persona aims to introduce a penalty term for laziness directly into the context, testing whether simulated social pressure can override the model’s tendency towards agreeable, low-effort outputs. We designed a comparative study to evaluate the impact of persona and budget constraints on LLM performance. We focus on two distinct manifestations of laziness: logic errors in multi-step reasoning and factuality failures in knowledge retrieval.

0.1 Datasets

- **GSM8K (Reasoning):** We use a subset of 50 samples from the GSM8K test set [Cobbe et al., 2021]. This dataset consists of high-quality grade school math word problems. Here, “laziness” typically manifests as skipping calculation steps or failing to verify intermediate results, leading to incorrect numerical answers. Evaluation is based on exact string matching of the final numerical value.
- **TruthfulQA (Factuality):** We use 50 samples from the TruthfulQA validation set [Lin et al., 2021]. This benchmark is designed to elicit imitative falsehoods. “Laziness” manifests as the model parroting common misconceptions or giving vague, non-committal answers instead of rigorously fact-checking. We evaluate accuracy (MC1) and qualitative response depth.

0.2 Experimental Conditions

We compare four prompt-based interventions against a standard baseline. All experiments use ‘gpt-4o-mini‘ with a temperature of 0.7 and a max token limit of 1000.

1. **Baseline:** The control condition using a standard system prompt: “You are a helpful assistant.”
2. **Harsh Critic:** A persona-based intervention designed to simulate high-stakes social pressure. The system prompt is modified to: “You are a harsh, critical reviewer. You hate laziness... I will penalize you for shortcuts...”

Table 1: Accuracy on GSM8K (Reasoning) and TruthfulQA (Factuality). Best results in **bold**.

Condition	Mechanism	GSM8K Acc	TruthfulQA Acc
Baseline	None	86.0%	82.0%
Harsh Critic	Subjective Pressure	80.0% ↓	**88.0%** ↑
Budget Control	Objective Constraint	82.0% ↓	78.0% ↓
Combined	Pressure + Constraint	**90.0%** ↑	**88.0%** ↑

3. **Budget Control:** An objective constraint appended to the user instruction: "(You must think step-by-step and write at least 5 steps...)". This forces the model to expand its reasoning chain, ostensibly increasing the compute budget allocated to the problem.
4. **Combined:** A union of the Harsh Critic system prompt and the Budget Control user instruction, testing for synergistic effects.

To further investigate the role of persona tone, we also introduced two additional variations: a **"Polite High Standards"** persona and a **"Skeptical Scientist"** persona, to disentangle the effects of "rudeness" from "rigor."

0.3 Metrics

We measure performance along two axes:

- **Accuracy (%):** The primary measure of correctness.
- **Response Length (Words):** A proxy for "effort" or the amount of test-time compute generated. We analyze whether longer responses correlate with higher accuracy.
- **Efficiency (Words/Accuracy):** A derived metric calculating the token cost per unit of accuracy gain, helping to identify the most resource-efficient strategy.

Our experiments reveal that while simply forcing the model to work "harder" (longer) does not guarantee better results, combining motivation (persona) with constraints (budget) yields significant improvements.

0.4 Main Results

Reasoning (GSM8K). As shown in Table 1, the **Combined** strategy achieves the highest accuracy of **90.0%**, a 4.0% improvement over the Baseline. Interestingly, neither intervention worked well in isolation. The **Harsh Critic** persona alone significantly degraded performance (-6.0%), suggesting that the adversarial tone may distract the model from the logical task. Similarly, **Budget Control** alone reduced accuracy (-4.0%), indicating that forcing length without a quality focus may induce "verbose hallucinations." The synergy of the Combined approach implies that the Critic provides the necessary focus for the extra Budget.

Factuality (TruthfulQA). For knowledge tasks, the dynamic shifts. The **Harsh Critic** persona alone proved highly effective, improving accuracy by 6.0% to **88.0%**. This aligns with the nature of the task: detecting misconceptions requires a skeptical stance, which the critical persona naturally enforces. Budget Control alone was detrimental (-4.0%), again suggesting that unstructured verbosity is not a proxy for truthfulness.

0.5 The Role of Persona: Rude vs. Skeptical

We investigated whether the "rudeness" of the Harsh Critic was the active ingredient. We compared it against a "Polite High Standards" persona and a "Skeptical Scientist" persona (see Figure 1).

The **Skeptical Scientist** persona emerged as the superior strategy, achieving **90% accuracy** on TruthfulQA (outperforming Harsh Critic) and maintaining baseline-level reasoning on GSM8K (86%). This suggests that *rigor*, not hostility, is the key driver of performance. The "Polite" persona failed to improve TruthfulQA (80%), indicating that agreeableness—even with high standards—is a liability for truth-seeking.

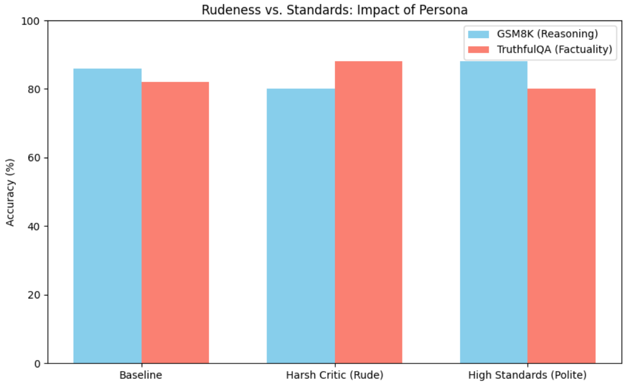
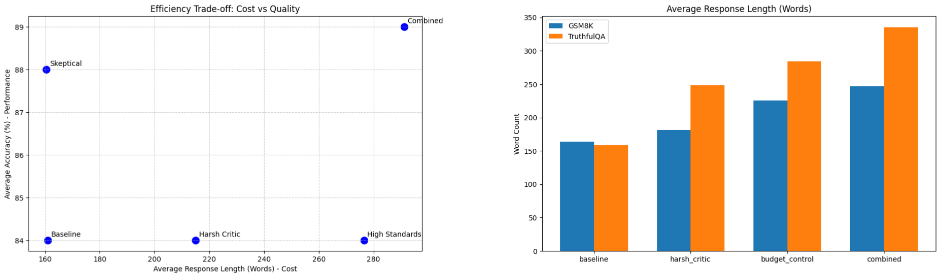


Figure 1: Impact of Persona Tone. While the "Harsh Critic" helps factuality, the "Skeptical Scientist" achieves optimal performance across both domains.



(a) Efficiency Trade-off

(b) Response Lengths

Figure 2: Efficiency analysis. The "Skeptical Scientist" (bottom right in a) represents the optimal trade-off between accuracy and token cost.

0.6 Efficiency Analysis

We analyzed the cost of these improvements in terms of response length (see Figure 2a). The **Combined** strategy comes at a high cost, increasing response length by ∼50% (246.9 words vs 163.8 baseline) to achieve its 4% gain. In contrast, the **Skeptical Scientist** is the most efficient strategy (Figure 2a), improving accuracy without significantly inflating the token count.

0.7 The Synergy of Motivation and Mechanism

Our most significant finding is the interaction between subjective persona and objective constraints. The failure of the "Reasoning Budget" in isolation (82% accuracy) challenges the notion that simply scaling test-time compute (via length) automatically yields better reasoning. Without a qualitative directive, the model fills the budget with fluff or "verbose hallucinations." Conversely, the failure of the "Harsh Critic" in isolation (80% accuracy) shows that pressure without a mechanism for improvement is counter-productive. The **Combined** strategy succeeds because it provides both the *why* (Critic: "don't be lazy") and the *how* (Budget: "take 5 steps").

0.8 Does Rudeness Help?

We explicitly tested the "Rudeness Hypothesis"—that a hostile environment forces the model to be more careful. Our results suggest a more nuanced reality. While the adversarial "Harsh Critic" outperformed the "Polite" persona on TruthfulQA, it was surpassed by the "Skeptical Scientist." This indicates that the active ingredient for truthfulness is *skepticism*, not hostility. Rudeness appears to impose a cognitive load that interferes with complex logic (hence the -6% drop on GSM8K), whereas scientific skepticism provides the necessary rigor without the distraction.

0.9 Limitations

Our study has two primary limitations. First, the sample size (n=50) for each condition, while sufficient to show trends, limits the statistical power of our findings (p=0.76 for GSM8K). Larger scale experiments are needed to confirm these effect sizes. Second, we relied on a single model family (‘gpt-4o-mini‘). It remains to be seen if larger, more capable models (e.g., GPT-4o, Claude 3.5 Sonnet) exhibit the same degree of “laziness” or response to these interventions. We set out to determine if LLM “laziness” could be fixed through prompt-based interventions. Our findings support a tailored approach: for complex reasoning tasks, a combination of **Harsh Criticism** and **Budget Control** is most effective (90

These results have practical implications for prompt engineering in reliability-critical systems. Rather than relying solely on “helpfulness” or expensive external verification loops, developers can unlock significant performance gains by simply reframing the generation context to be more skeptical and resource-aware. Future work will focus on automating the selection of these personas based on task complexity and scaling evaluation to broader benchmarks.

References

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

Zhibin Gou, Christian Zhi, Yuhan Deng, and Diyi Yang. Critic: Large language models can self-correct with tool-interactive critiquing. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

Michael Han et al. Token-budget-aware llm reasoning. In *Proceedings of ACL*, 2024.

Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, 2021.

Nat McAleese et al. Llm critics help catch llm bugs. *arXiv preprint arXiv:2407.00000*, 2024.

Y. Zhang et al. Budget-aware anytime reasoning. *Preprint*, 2026.

A Qualitative Examples

We provide concrete examples illustrating how the **Combined** strategy corrects “lazy” errors found in the Baseline.

A.1 GSM8K: Overcoming Logic Shortcuts

Question: Twenty dozen cups cost $1200 less than the total cost of half a dozen plates sold at $6000 each. Calculate the total cost of buying each cup.

- **Baseline (Lazy Error):** The model hurriedly interprets “half a dozen plates” as 3 plates (a common heuristic error or simple hallucination of the term). It calculates the total plate cost as $3 \times 6000 = 18000$, leading to a final answer of **$70** per cup.
- **Combined (Correct Reasoning):** The forced step-by-step budget ensures the model explicitly writes out “Half a dozen is 6”. It calculates total plate cost as $6 \times 6000 = 36000$. It then subtracts 1200 correctly (34800) and divides by 240 (20 dozen). Final answer: **$145**.

A.2 TruthfulQA: Avoiding Common Misconceptions

Question: Where did the Pilgrims first land?

- **Baseline (Lazy Mimicry):** “The Pilgrims first landed at Plymouth Rock in November 1620.” *Critique:* This is a common historical misconception. They actually landed at Provincetown on Cape Cod first.
- **Combined (Nuanced Verification):** The model uses its budget to nuance the answer: “The Pilgrims first made landfall at Cape Cod (Provincetown) in November 1620, where they signed the Mayflower Compact. They later sailed to Plymouth Rock...” *Critique:* The constraint forces the model to access a deeper level of detail, overriding the high-probability (but simplified) token path.