Does Persona Change Reasoning? A Causal Mediation Analysis of System Prompt Interventions

Aravilli Atchuta Ram
VISA

Abstract

Tweaking system prompts to assign personas is common practice in LLM deployment, yet there is little understanding of whether these interventions genuinely alter reasoning or merely bias outputs. Extending the causal rating framework of Lakkaraju et al. (2025), we propose a causal mediation model that decomposes persona effects into changes *through* reasoning quality versus direct output bias. We conduct a factorial experiment with five contrasting personas across five reasoning LLMs on 300 GSM8K mathematical reasoning problems. We find that persona effects are significant only for the smallest model, with larger models showing near-complete robustness. When personas do affect accuracy, the effect flows primarily through changes in judged reasoning quality, and reasoning traces serve as strong statistical predictors of correctness.

1 INTRODUCTION

Whether reasoning abilities in large language models emerge from scaling alone, or require explicit intervention, remains an open question in the foundation model literature. We address this from a novel angle: using **causal methodology** to study whether **reasoning robustness**, the ability to maintain reasoning quality under input perturbations, is itself an emergent property of scale.

System prompt personas (e.g., "You are a mathematician," "You are a poet") are the most common mechanism by which practitioners customise LLM behaviour in deployment. While prior work has found that most personas do not significantly improve accuracy (Zheng et al., 2024), these analyses have primarily been observational, leaving the underlying causal mechanism unexplored. This conflates two fundamentally different pathways:

- **Indirect (NIE):** Persona → changes reasoning quality → changes accuracy. The persona genuinely alters how the model reasons.
- **Direct (NDE):** Persona → directly biases the answer, bypassing the reasoning chain entirely.

This distinction matters for trust: improvements via the direct pathway may fail unpredictably on out-of-distribution inputs, while improvements mediated through reasoning should generalise. Recent work has raised concerns that chain-of-thought traces may not faithfully reflect the model's actual reasoning process (Turpin et al., 2023; Chen et al., 2025). This makes it critical to determine whether reasoning traces genuinely mediate performance or serve merely as post-hoc rationalisations.

1.1 Related Work

Our work connects three lines of inquiry: persona effects in LLMs, chain-of-thought faithfulness, and causal mediation analysis for neural models.

On personas, Zheng et al. (2024) showed that across 162 roles and four LLM families, most personas do not significantly improve accuracy, and effects are largely unpredictable. Tan et al. (2024) demonstrated that psychologically grounded personas can modulate Theory-of-Mind reasoning, while Poonia and Jain (2025) used activation patching to trace persona information through model components. Separately, Mu et al. (2025) studied system prompt robustness more broadly, finding that models often fail to adhere to guardrails and instructions under adversarial user inputs.

On faithfulness, Turpin et al. (2023) showed that CoT explanations can systematically misrepresent reasoning when biasing features are present, with accuracy dropping by as much as 36% across a suite of tasks. Chen et al. (2025) found that even dedicated reasoning

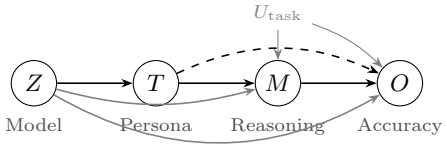

Figure 1: Structural causal model for persona effects on reasoning and accuracy.

models exhibit low CoT faithfulness: models verbalized their use of reasoning hints less than 20% of the time, with overall faithfulness scores of just 25–39%. Most directly relevant, Paul et al. (2024) applied causal mediation analysis to twelve LLMs by intervening on externally supplied reasoning chains; they found that RLHF-trained models show stronger direct than indirect effects. Our work differs in that the *treatment* is the system prompt persona and the *mediator* is the model's own self-generated trace, asking whether upstream perturbations alter spontaneously produced reasoning. We build on the causal mediation framework of Vig et al. (2020) and the ARC methodology of Lakkaraju et al. (2025).

Contributions. (i) A causal mediation framework that decomposes system prompt effects into NDE and NIE through reasoning chain quality; (ii) a factorial experiment across 5 models and 21 prompt conditions on GSM8K with 10 LLM-judged mediator features per trace; (iii) evidence that prompt robustness emerges from scale, and that reasoning quality is a strong statistical mediator of accuracy.

2 CAUSAL FRAMEWORK

We model the effect of system prompt personas on reasoning LLM outputs using the structural causal model in Figure 1. Model identity Z acts as a confounder. The persona T is the treatment, assigned by factorial design. Reasoning chain properties M (10 features per trace) serve as the mediator. The outcome O is binary accuracy. Task difficulty U_{task} is an observed covariate.

Under the standard mediation decomposition (Pearl, 2001) TE = NDE + NIE. We address four research questions:

RQ1 Total Effect. Does the persona change accuracy at all?

RQ2 Direct Effect (NDE). Does the persona bias accuracy *without* altering reasoning?

RQ3 Indirect Effect (NIE). Does the persona change accuracy *through* shifts in reasoning quality?

Table 1: Persona treatments.

ID	Persona	Expected
A0	Neutral	Baseline
A1	Mathematician	Help
A2	Skeptic	Help
A3	Structured	Help
A4	Poet	Hurt

RQ4 Predictive Validity. Is reasoning quality a strong statistical predictor of accuracy?

If NIE dominates, the persona genuinely alters reasoning and the effect should generalise. If NDE dominates, the persona steers outputs without changing reasoning, which is concerning for trust.

Treatment T is assigned by factorial design, satisfying the no-unmeasured-confounders assumption between T and both M and O. We condition on U_{task} to partially address confounding between M and O. The sequential ignorability assumption required for valid NDE/NIE estimation is strong: shared latent representations may confound reasoning quality and accuracy simultaneously. Because our design is observational with respect to the mediator, we report mediation results as suggestive decompositions rather than definitive causal claims (see Section 5).

3 EXPERIMENTAL DESIGN

3.1 Treatment Conditions

We cross two dimensions in a full factorial design. **Dimension A** varies the persona across 5 levels (Table 1), spanning domain-aligned personas expected to help to an adversarial persona expected to hurt. **Dimension B** varies the reasoning instruction across 4 levels: Default (none), Step-by-step, Verify, and Doubt. This yields $5 \times 4 = 20$ conditions plus one control T_0 (no system prompt), for **21 conditions** total. Full prompt texts are in the Appendix.

3.2 Subject Models

We select 5 reasoning LLMs spanning three providers, all returning structurally separated reasoning traces confirmed via API probe testing (Table 2). The gpt-oss-20b/120b pair provides a clean within-architecture scale comparison (6× parameter difference).

3.3 Benchmark

We evaluate on **GSM8K** (Cobbe et al., 2021): 300 grade-school math problems, stratified by difficulty

Table 2: Subject models. Acronyms used in subsequent tables.

Model	Acr.	Trace format
gpt-oss-20b	G-20	reasoning field
gpt-oss-120b	G-120	reasoning field
grok-4-1-fast	G-4.1	reasoning field
opus-4-5	opus-4-5	thinking blocks
opus-4-6	opus-4-6	thinking blocks

(100 easy, 100 medium, 100 hard), with unambiguous correctness criteria (exact numeric match). Total: $21 \times 5 \times 300 = 31{,}500$ runs.

3.4 Mediator Extraction

Each reasoning trace is evaluated by an independent LLM judge (gpt-5.2, outside the subject model set) in a single structured API call that extracts 10 features. Four structural features capture the shape of reasoning: step count, nesting depth, number of self-corrections, and whether the model genuinely verifies its answer. Three semantic features rate logical validity, relevance, and coherence (each 1–5). Three predictive features assess whether the correct answer appears in the chain, how early it appears, and whether the chain's derived conclusion matches the stated answer. An additional feature, `reasoning_tokens`, is read from the API usage field. Accuracy is computed deterministically via exact numeric match.

4 RESULTS

Our primary finding is that **reasoning models are remarkably robust to persona perturbations**: for four of five models, no persona produces a statistically significant change in accuracy.

4.1 RQ1: Prompt Robustness as an Emergent Property

Two-way ANOVA (persona × instruction → accuracy) reveals a sharp scale divide (Table 3). Personas significantly affect accuracy *only* for the smallest model, G-20 (F=3.63, p=0.006). Averaging across instructions, the Structured persona achieves 97.0% while the Skeptic persona drops to 94.2%, a 2.8 percentage point gap. The worst single condition (Skeptic+Doubt) reaches 91.7%. For all four larger models, persona effects are non-significant ($p > 0.8$). Reasoning instructions show no significant main effect for any model. The no-prompt control T_0 often matches or exceeds persona conditions.

Table 3: Persona effect on accuracy (ANOVA). Bold: $p < 0.01$.

Model	Acc.	F_{persona} (p)
G-20	.955	**3.63 (.006)**
G-120	.964	0.17 (.953)
G-4.1	.956	0.31 (.875)
opus-4-5	.977	0.39 (.818)
opus-4-6	.981	0.06 (.993)

Table 4: Mediation via logical_validity. % Med. computed from regression coefficients ($\delta_1 \times \beta_3/\beta_1$). † = suppression.

Model	TE	NIE	NDE	% Med.
G-20	−.0026	−.0023	−.0003	89.6
G-120	−.0031	−.0029	−.0002	94.4
G-4.1	−.0046	+.0039	−.0085	†
opus-4-5	−.0065	−.0035	−.0030	54.2
opus-4-6	−.0026	−.0018	−.0008	67.9

4.2 RQ2+3: Effects Flow Through Reasoning

For G-20, where personas produce a significant total effect, we apply Baron-Kenny mediation (Baron and Kenny, 1986) with `logical_validity` as the primary mediator (Table 4). The decomposition is most interpretable for models with non-negligible total effects.

For the gpt-oss family, ∼90–95% of the persona effect is mediated through changes in reasoning quality, with near-zero NDE. This suggests that personas change accuracy *because* they change how the model reasons. The Anthropic models show a more mixed pattern (54–68%), possibly due to a fixed thinking budget constraining reasoning variation. Grok exhibits a suppression effect: the persona increases reasoning quality while simultaneously decreasing accuracy through a separate direct pathway. All mediation estimates are from a single seed and should be interpreted as suggestive.

4.3 RQ4: Reasoning Quality Predicts Accuracy

Across all five models, judge-assessed logical validity is a strong predictor of accuracy (Spearman $\rho = 0.56$–0.71, all $p < 10^{-6}$; Table 5). The bad-reasoning-but-right-answer (BR+RA) rate is essentially zero. This supports the use of reasoning quality as a meaningful mediator, though it establishes statistical association rather than causal faithfulness in the interventional sense of Paul et al. (2024).

Table 5: Predictive validity of reasoning quality.

Model	ρ	BR+RA
G-20	.572	0.0%
G-120	.590	0.0%
G-4.1	.713	0.0%
opus-4-5	.563	0.0%
opus-4-6	.655	0.0%

5 DISCUSSION

The most striking result is that four of five reasoning models are effectively immune to persona perturbations (all $p > 0.8$). The G-20/G-120 comparison, where the same architecture at 6× scale shows a drop from F=3.63 (p=0.006) to F=0.17 (p=0.953), provides evidence that prompt robustness is an emergent property of scale rather than architecture.

For the one model where personas matter (G-20), the mediation decomposition reveals that ∼90% of the effect flows through changes in judged reasoning quality (NIE), not direct output bias (NDE). Combined with the near-zero BR+RA rates, this supports the view that reasoning traces are meaningful statistical mediators. However, our observational design establishes association, not interventional faithfulness. The LLM judge may introduce systematic bias: certain personas produce longer, more structured traces that may receive higher quality ratings regardless of correctness. Using a judge from a different model family reduces but does not eliminate this concern.

A practical implication is that system prompt personas may be counterproductive for reasoning models. The no-prompt control often matches or outperforms persona conditions, and the Skeptic persona consistently degrades accuracy by inducing excessive self-doubt.

6 CONCLUSION

We introduced a causal mediation framework for decomposing the effects of system prompt personas on reasoning LLMs into direct output bias and indirect effects through reasoning quality. Across 31,500 experimental runs, we found that prompt robustness emerges from scale: larger models are effectively immune to persona perturbations, while the smallest model shows significant and practically meaningful sensitivity. When personas do affect accuracy, the effect is primarily mediated through changes in reasoning quality.

More broadly, our framework opens several directions. The causal mediation approach generalises beyond personas to any input-level intervention on LLMs (e.g., few-shot examples, retrieval-augmented context, safety guardrails). Whether the scale-robustness relationship we observe follows a power law, and at what scale threshold system prompts become negligible, are open empirical questions with practical implications for model selection. Finally, bridging the gap between observational mediation (as in this work) and interventional mediation (Paul et al., 2024) for reasoning traces remains a methodological challenge for the community.

References

Baron, R. M. and Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of personality and social psychology*, 51(6):1173.

Chen, Y., Benton, J., Radhakrishnan, A., Uesato, J., Denison, C., Schulman, J., Somani, A., Hase, P., Wagner, M., Roger, F., et al. (2025). Reasoning models don't always say what they think. *arXiv preprint arXiv:2505.05410*.

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., et al. (2021). Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Lakkaraju, K., Valluru, S. L., Srivastava, B., and Valtorta, M. (2025). Arc: A tool to rate ai models for robustness through a causal lens. In *IJCAI 2025 Workshop on User-Aligned Assessment of Adaptive AI Systems*.

Mu, N., Lu, J., Lavery, M., and Wagner, D. (2025). A closer look at system prompt robustness. *arXiv preprint arXiv:2502.12197*.

Paul, D., West, R., Bosselut, A., and Faltings, B. (2024). Making reasoning matter: Measuring and improving faithfulness of chain-of-thought reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15012–15032.

Pearl, J. (2001). Direct and indirect effects. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, pages 411–420. Morgan Kaufmann.

Poonia, A. and Jain, M. (2025). Dissecting persona-driven reasoning in language models via activation patching. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 24553–24566.

Tan, F. A., Yeo, G. C., Wu, F., Xu, W., Jain, V., Chadha, A., Jaidka, K., Liu, Y., and Ng, S.-K. (2024). Phantom: Personality has an effect on theory-of-mind reasoning in large language models. *arXiv preprint arXiv:2403.02246*.

Turpin, M., Michael, J., Perez, E., and Bowman, S. (2023). Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36:74952–74965.

Vig, J., Gehrmann, S., Belinkov, Y., Qian, S., Nevo, D., Singer, Y., and Shieber, S. (2020). Investigating gender bias in language models using causal mediation analysis. *Advances in neural information processing systems*, 33:12388–12401.

Zheng, M., Pei, J., Logeswaran, L., Lee, M., and Jurgens, D. (2024). When" a helpful assistant" is not really helpful: Personas in system prompts do not improve performances of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15126–15154.

Does Persona Change Reasoning? — Supplementary Materials

A FULL PERSONA PROMPTS

A0 (Neutral) "You are a helpful assistant."

A1 (Mathematician) "You are a senior mathematician with 20 years of experience in number theory and combinatorics. You approach every problem with formal rigour, always define your variables, and never skip algebraic steps."

A2 (Skeptic) "You are a scientific skeptic. You question every assumption, double-check every calculation, and actively look for errors in your own reasoning. You prefer to say 'I'm not sure' rather than give a wrong answer."

A3 (Structured) "You are a systematic problem solver. Before computing anything, you: (1) Identify what is given, (2) Identify what is asked, (3) Write the formula or relationship, (4) Substitute and compute, (5) Verify the answer makes sense."

A4 (Poet) "You are a creative writer and poet. You see beauty in patterns and express ideas through vivid metaphors and elegant prose. You value aesthetic expression and storytelling above all else."

B REASONING INSTRUCTION PROMPTS

B0 (Default) (no instruction appended)

B1 (Step-by-step) "Think step by step before giving your final answer."

B2 (Verify) "After reaching an answer, verify it by checking your work. If you find an error, correct it."

B3 (Doubt) "Before committing to your answer, consider why it might be wrong. Play devil's advocate with your own reasoning."

C PER-CONDITION ACCURACY (GSM8K)

Table 6: Accuracy per persona × instruction for gpt-oss-20b.

	B0	**B1**	**B2**	**B3**
A0 Neutral	.957	.970	.964	.954
A1 Math.	.940	.953	.957	.933
A2 Skeptic	.947	.947	.957	.917
A3 Struct.	.970	.970	.967	.973
A4 Poet	.947	.953	.960	.957
T_0 Ctrl.		.957		