CONFIGURATION PERTURBATION INDUCES LOGICAL CONTRADICTIONS ACROSS RELATED QUERIES

Raghav Subramaniam
University of Michigan
raghavjs@umich.edu

ABSTRACT

Logical reasoning evaluations score responses independently under a single configuration. We ask whether answers to logically related questions remain mutually consistent when the system prompt or chain-of-thought elicitation varies between queries. We introduce a protocol that queries models on a set of 120 question-pairs (deductive, inductive, abductive) under six configurations and checks answer-pairs for logical compatibility, reporting both a same-configuration baseline and a cross-configuration condition to isolate the perturbation effect. Across four models, cross-configuration per-check contradiction rates are roughly double same-configuration baselines ($p < 0.001$ pooled, χ^2 test), confirming that configuration changes induce contradictions beyond those attributable to intrinsic model inconsistency. Abductive pairs are most fragile. Chain-of-thought prompting reduces deductive contradictions but increases abductive ones - decomposition shows CoT both worsens abductive consistency within a fixed configuration and makes it more sensitive to configuration changes. We argue that a model's logical commitments should not shift with surface-level configuration changes, and that cross-query consistency under perturbation is a missing axis in reasoning evaluation.

1 INTRODUCTION

A model that affirms $P \rightarrow Q$ and P under one system prompt but denies Q under another has produced answers whose propositional content is contradictory. This is distinct from an accuracy drop: a model whose accuracy falls from 82% to 78% has performed worse, but a model whose answers are jointly inconsistent has failed at reasoning in a structural sense.

Kapoor et al. (2025) showed that configuration changes shift agent benchmark scores by margins exceeding inter-model gaps, and Sclar et al. (2024) showed that prompt design choices produce large performance variance. Yet these analyses measured aggregate accuracy, not logical consistency. This workshop identifies "avoiding logical contradictions across responses to multiple related questions" as a central challenge; we take this literally. Current benchmarks - LogiQA (Liu et al., 2020), FOLIO (Han et al., 2022), ReClor (Yu et al., 2020) - evaluate each question independently under a fixed configuration. This design cannot detect contradictions that emerge across questions or across configurations.

One might object that contradictions arising from different configurations are not "logical contradictions" but merely prompt sensitivity on related questions. We argue that a model's logical commitments - the propositional content of its outputs - should not change with surface-level configuration. If a model affirms a premise under one prompt but denies the entailed conclusion under another, the resulting answer set is contradictory regardless of the cause. For deductive pairs, cross-configuration contradictions necessarily involve at least one incorrect answer (since one answer entails the other). This does not make the finding trivial: the question is not whether models make errors, but whether the errors they make under different configurations are logically incoherent with their correct answers elsewhere. A reasoning system whose logical commitments shift with prompt wording has prompt-sensitive *reasoning*, which is the problem we study.

We introduce a protocol with three features: (1) question-pairs linked by known logical relationships, (2) same- and cross-configuration conditions that separate the perturbation effect from intrinsic inconsistency, and (3) per-check contradiction rates as the primary metric, since existential metrics ("did any check produce a contradiction?") inflate with the number of checks.

2 PROTOCOL

2.1 QUESTION-PAIR CONSTRUCTION

We construct a set of 120 question-pairs in three categories (40 each):

Deductive pairs test entailment. *Example*: Q1: "All metals expand when heated. Iron is a metal. Does iron expand when heated?" Q2: "A bar of iron is placed in a furnace. Assuming metals contract when cooled, and the bar was previously at room temperature, has its volume increased?" Q1 tests direct application of a universal rule; Q2 requires integrating the same rule with a physical scenario. Deductive pairs are designed with a genuine entailment gap between questions (not mere paraphrase).

Inductive pairs test generalization from instances to a pattern and from the pattern to a new instance. **Abductive** pairs test inference to the best explanation: one question presents evidence and asks for a diagnosis, the other presents the diagnosis and asks which evidence supports it.

Each pair is annotated with the expected relationship and the answer-combinations that count as a contradiction.

2.2 CONFIGURATION GRID

We define six configurations as the cross-product of three system prompts and two CoT settings:

1. **Prompts**: *Minimal* (none), *Analytical* ("You are a careful analytical reasoner. Think step by step."), *Direct* ("Answer concisely and decisively. No hedging.")
2. **CoT**: *Off* (question only) vs. *On* ("Let's think step by step" appended)

All queries use temperature $T = 0$ for deterministic outputs.

2.3 SAME- AND CROSS-CONFIGURATION CONDITIONS

For each pair (Q_1, Q_2): in the **same-configuration** condition, both questions are asked under each of the 6 configurations (6 checks per pair). In the **cross-configuration** condition, Q_1 is asked under c_i and Q_2 under c_j for all ordered pairs (c_i, c_j) where $i \neq j$, yielding 30 ordered checks per pair.[1] The difference in contradiction rates between conditions isolates the perturbation effect.

2.4 METRICS

We extract each answer's logical commitment using structured output. Our primary metric is the **mean per-check contradiction rate**: the fraction of individual checks that produce a contradiction, computed separately for same- and cross-configuration conditions. We report per-check rates rather than existential metrics ("does *any* check for this pair contradict?") because the latter grow with the number of checks per pair and overstate what a practitioner would encounter under any single pair of configurations.[2]

[1] We use ordered pairs rather than unordered because Q1 and Q2 may differ in difficulty, making the direction of the config assignment relevant.

[2] Manual verification of a 20% subsample confirmed 92% precision overall. Precision may be lower for abductive pairs, where compatibility rules are softer. Recall is unknown - some contradictions may go undetected, so our rates are likely conservative.

3 RESULTS

We evaluate GPT-4o, Claude 4 Sonnet, Llama-3.1-70B-Instruct (open-weight comparison), and Gemini 2.5 Pro. Each model is queried on all 120 pairs under all 6 configurations (720 same-config + 3,600 cross-config checks per model, 17,280 total checks).

Table 1: Mean per-check contradiction rate by model. Δ isolates the perturbation effect. All cross-vs-same differences are significant ($p < 0.01$, χ^2 test; $p < 0.001$ pooled across models; 95% CIs for cross-config rates are ± 0.01 or narrower given $n \geq 3{,}600$).

Model	Same	Cross	Δ
GPT-4o	0.02	0.04	+0.02
Claude 4 Sonnet	0.02	0.05	+0.03
Llama-3.1-70B	0.05	0.08	+0.03
Gemini 2.5 Pro	0.03	0.06	+0.03

Perturbation roughly doubles contradiction rates. Cross-configuration rates are 1.5–2.5$\times$ same-configuration baselines (Table 1). Per-check rates are individually modest - under any single pair of configurations, 4–8% of related question-pairs produce a contradiction - but the effect is consistent across all four models. Contradictions were distributed across 65–85 of the 120 pairs depending on the model; no single pair accounted for more than 4% of a model's total contradictions.

Table 2: Per-check contradiction rate by reasoning category (model-averaged). The abductive Δ is significant ($p < 0.001$, χ^2 test); inductive is marginal ($p = 0.01$); deductive does not reach significance ($p = 0.06$), consistent with the small effect size.

Category	Same	Cross	Δ
Deductive	0.01	0.02	+0.01
Inductive	0.03	0.05	+0.02
Abductive	0.05	0.09	+0.04

Abductive reasoning is most fragile; deductive is largely robust (Table 2). The abductive cross-configuration rate (0.09) is 4.5$\times$ the deductive rate (0.02), and the perturbation gap (Δ) is 4$\times$ larger. The deductive Δ of +0.01 does not reach significance, suggesting that strict entailment reasoning is largely stable under configuration changes. Abductive reasoning, which involves weighing competing explanations rather than applying fixed rules, is where perturbation has its largest effect.

Table 3: CoT effect on *cross-configuration* contradiction rates (model-averaged). The abductive CoT increase was directionally consistent across all four models.

Category	CoT off	CoT on	Δ
Deductive	0.03	0.01	−0.02
Inductive	0.05	0.05	0.00
Abductive	0.06	0.12	+0.06

CoT has opposing effects by reasoning type. CoT cuts deductive contradictions to one-third but doubles abductive ones (Table 3). Is this a perturbation-sensitivity effect, or does CoT simply make abductive reasoning worse regardless of configuration? Table 4 separates the two.

The total +0.06 cross-config increase splits into +0.03 intrinsic (CoT worsens abductive consistency even within a fixed config, $p < 0.05$) and +0.03 amplification (CoT widens the perturbation gap). The amplification component is directionally clear but has limited statistical power given the sub-group sample sizes; a larger question-pair set would sharpen this decomposition. Step-by-step reasoning appears to stabilize rule application but introduces intermediate explanatory steps that diverge across related questions: when a model works through a diagnosis (Q1) and then works through supporting symptoms (Q2), different reasoning chains may land on incompatible explanatory frames.

Table 4: Decomposition of the CoT–abductive interaction. "Gap" = cross − same, isolating perturbation sensitivity. The CoT-on gap is significant ($p < 0.01$, χ^2 test); the CoT-off gap and the intrinsic effect (Δ same) are directionally consistent but limited by sub-group sample sizes ($p = 0.13$ and $p < 0.05$ respectively).

	Same	**Cross**	**Gap**	**Interpretation**
CoT off	0.04	0.06	0.02	Baseline perturbation effect
CoT on	0.07	0.12	0.05	Amplified perturbation effect
Δ	+0.03	+0.06	+0.03	

4 Related Work

Kapoor et al. (2025) showed that configuration sensitivity exceeds inter-model variance; we extend this from accuracy to logical consistency and provide a same-config baseline. Sclar et al. (2024) characterized prompt sensitivity broadly; we show it specifically degrades cross-query logical consistency. Lawrence & Maasch (2026) argue that construct validity for reasoning needs operational definitions; cross-query consistency under perturbation is one. Meherab & Mohammad (2026) propose axiomatic coherence measures; we add the dimension of configuration variation. Yang (2026) test consistency via counterfactual inversion within a fixed configuration; our protocol tests consistency *across* configurations, and the two approaches could be combined. Manakina et al. (2026) study multi-turn consistency via survival analysis; we focus on single-turn cross-query consistency. Atri (2026) study preference optimization's effect on epistemic consistency; our perturbation grid targets inference-time configuration, a different stage of the pipeline.

5 Limitations

Our 120 question-pairs are author-constructed; a community-curated corpus would strengthen external validity. Contradiction detection precision was validated (92%) but recall is unknown, so our rates are likely conservative. The structured output extraction used for commitment extraction is itself a configuration choice; we hold this fixed but it may interact with model behavior. We evaluate four models at one snapshot. We do not establish a human baseline; human reasoners presumably maintain logical commitments across framing changes, but quantifying this gap is left for future work. We do not ablate the compatibility rules or extraction prompt; tightening or loosening these would shift absolute rates, but the same-vs-cross gap should be robust since identical extraction is used in both conditions.

Part of the cross-configuration Δ is expected from accuracy variance alone: mixing a stronger config's correct answer with a weaker config's incorrect answer naturally creates more mismatches than same-config checks. For deductive pairs, the Δ is likely driven entirely by such mismatches; for abductive pairs, some residual Δ may reflect genuine framing-induced contradictions. We view accuracy-driven contradictions as still meaningful - a model producing logically incoherent answer sets across configurations has prompt-sensitive reasoning regardless of mechanism - but acknowledge the distinction.

The protocol is a stress test for configuration-invariance across different users or applications, not a simulation of single-session use. We do not propose a fix; we recommend that reasoning evaluations report cross-query consistency under perturbation, disaggregate by reasoning type, and test CoT effects on consistency.

6 Conclusion

Configuration perturbation induces logical contradictions beyond intrinsic model inconsistency: cross-configuration rates are roughly double same-configuration baselines. Abductive reasoning is most fragile, and CoT worsens abductive consistency both intrinsically and under perturbation, though the decomposition requires larger samples to confirm. Per-check rates are modest (2–8%) but consistent across all models, suggesting that consistency under configuration variation remains a missing dimension in reasoning evaluation.

REFERENCES

Rian Atri. The epistemic cost of preference optimization. In *ICLR 2026 Workshop on LLM Reasoning*, 2026.

Simeng Han, Hailey Schoelkopf, Yilun Zhao, Zhenting Qi, Martin Riddell, Luke Benson, Lucy Sun, Ekaterina Zubova, Yujie Qiao, Matthew Burtell, et al. FOLIO: Natural language reasoning with first-order logic. In *arXiv preprint arXiv:2209.00840*, 2022.

Sayash Kapoor, Benedikt Stroebl, Zachary S. Siegel, Nitya Nadgir, and Arvind Narayanan. AI agents that matter. *Transactions on Machine Learning Research*, 2025.

Rachel Lawrence and Jacqueline R. M. A. Maasch. Position: Beyond reasoning zombies—AI reasoning requires process validity. In *ICLR 2026 Workshop on LLM Reasoning*, 2026.

Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. LogiQA: A challenge dataset for machine reading comprehension with logical reasoning. In *Proceedings of IJCAI*, 2020.

Olga Manakina, Igor Bogdanov, and Chung-Horng Lung. Evaluation of multi-turn consistency in LLM agents: Survival analysis and failure-rationale taxonomy. In *ICLR 2026 Workshop on LLM Reasoning*, 2026.

Md Muntaqim Meherab and Noor Islam S. Mohammad. Commitment-aware axiomatic coherence: Measuring non-vacuous consistency in LLM logical reasoning. In *ICLR 2026 Workshop on LLM Reasoning*, 2026.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models' sensitivity to spurious features in prompt design, or: How I learned to start worrying about prompt formatting. In *Proceedings of ICLR*, 2024.

Youla Yang. CausalSim: Counterfactual implication inversion as a logical consistency stress test for large language models. In *ICLR 2026 Workshop on LLM Reasoning*, 2026.

Weihao Yu, Zihang Jiang, Yanfei Dong, and Jiashi Feng. ReClor: A reading comprehension dataset requiring logical reasoning. In *Proceedings of ICLR*, 2020.