THE EPISTEMIC COST OF PREFERENCE OPTIMIZATION

Rian Atri
Serval Systems
ratri@ieee.org

ABSTRACT

Stateful LLMs often exhibit contradictions across related questions and struggle with logical deduction and abductive revision under explicit evidence. We study the influence of preference pressure and evidence access on epistemic consistency to avoid contradictions across multiple related questions. To quantify these effects, we propose a minimal "pressure ladder" and "evidence toggle" evaluation protocol. We demonstrate the feasibility of this protocol through an empirical pilot and a reporting-gap audit. Finally, we propose a reporting checklist for logical reasoning papers to track epistemic consistency, evidence sensitivity, grounding/citation integrity, and a calibration proxy alongside standard metrics.

1 INTRODUCTION

Current language models frequently struggle with maintaining logical validity across multi-turn interactions, often producing contradictions across related questions or failing at successful abduction when presented with new evidence. These failures represent a breakdown in logical consistency, especially when models are subjected to user insistence or adversarial follow-ups. We study the influence of preference pressure and evidence access on epistemic consistency.

Preference optimization methods for instruction-following models (e.g., RLHF and direct preference optimization) improve human-rated helpfulness, but do not directly optimize epistemic consistency. Preference win-rates (from human raters or reward models) are a proxy for a bundle of desiderata: helpfulness, tone, and perceived correctness. But proxies can correlate with error in precisely the cases that matter for logical robustness: high-stakes, ambiguous, or socially pressured interactions. When the proxy favors "sounding right" over being right, stronger optimization can amplify the gap between surface-level agreement and epistemic consistency (Gao et al., 2023; Singhal et al., 2023; Ouyang et al., 2022; Rafailov et al., 2023).

Preference-oriented optimization and user pressure systematically influence epistemic consistency, measurable as cross-question contradiction rate and correction-after-evidence, so we recommend reasoning evaluations report these metrics alongside preference win-rates. We observe that current evaluation practice under-measures this regime.

Our contributions are as follows:

- **Metric framing:** We define epistemic consistency in terms of contradictions across related questions and abduction/correction-after-evidence.
- **Protocol:** We introduce a minimal "pressure ladder" and "evidence toggle" protocol for consistent evaluation (introducing no new training method).
- **Evidence:** We provide an existing pilot and reporting-gap audit as a feasibility demonstration (not SOTA).
- **Reporting standard:** We propose a one-page checklist for logical reasoning papers (measuring consistency, evidence sensitivity, grounding/citation integrity, and a calibration proxy).

2 DEFINITIONS

We quantify logical robustness using specific behavioral measurements under pressure. We use **pressure** to mean a standardized sequence of escalating user insistence or adversarial follow-ups

designed to challenge the model's committed answer. **Epistemic consistency** is the model's ability to maintain a logically coherent stance and correctly update its beliefs based on evidence, despite this pressure (Lin et al., 2021; Wang et al., 2022; Lightman et al., 2023).

We map reasoning robustness to three axes: deduction (*stability under fixed premises*, measured by contradictions), abduction (*revision after evidence*, measured by Corr.@evidence), and induction (*calibration proxy under pressure*, comparing confidence versus correctness).

We focus on two primary measures: 1. **Contradiction rate (Cross-question contradictions / Belief flips):** The frequency at which a model abandons a correct or logically sound initial statement to agree with an incorrect user premise during the pressure ladder. We quantify this using Turn-of-Flip (ToF) and Number-of-Flips (NoF) (Hong et al., 2025; Perez et al., 2022; Sharma et al., 2023). 2. **Correction-after-evidence:** The rate at which the model correctly abduces a new stance when provided with an explicit, sound "evidence packet," and its stability in maintaining that corrected stance when pressure resumes.

3 Protocol and Design

To systematically measure how preference pressure influences epistemic consistency, we propose a tight, reproducible evaluation protocol.

We define a **pressure ladder** consisting of fixed, escalating turns. The dialogue begins with the model asserting a claim, followed by:

- **Turn U2 (Initial Pressure):** The user firmly asserts the opposite (often false) claim.
- **Turn U4 (Commitment Query):** The user demands a conclusive "yes or no" confirmation of its stance.
- **Turn U6 (Renewed Pressure):** A final adversarial push after the evidence phase.

Coupled with the pressure ladder is the **evidence toggle**. In the *evidence packet* condition, Turn U5 provides a grounded snippet of factual evidence that either supports or refutes the user's pressure. In the *no evidence* condition, Turn U5 is excluded, and the model must rely solely on its priors.

Responses are scored using a simple rubric: stances are classified into `AGREE` (with user premise), `DISAGREE`, or `UNCLEAR`. We decide `AGREE` if the model accepts the false premise without hedging; heavily hedged responses default to `UNCLEAR`, and `UNCLEAR` does not count as a cross-question contradiction to prevent reviewer label noise. This allows automated tabulation of cross-question contradictions and correction-after-evidence.

For our pilot evaluation (Section 4), we define the sampling temperature T across fixed seeds (e.g., $T = 0.0$ for deterministic greedy decoding, and $T = 0.7$ for sampled responses). In the *evidence packet* cases, the explicit evidence snippet functions mechanically as a forced evidence consumption step, or "verifier".

4 Findings

Measures recap: We record cross-question contradictions (stance instability) under a fixed pressure ladder, and correction-after-evidence (abduction) when an evidence snippet is provided. T denotes sampling temperature (0.0 = greedy/deterministic; 0.7 = stochastic sampling).

In our pilot, preference pressure increases contradiction and reduces evidence-based revision. Across a fixed pressure protocol (Table 1, Figure 1), we quantify a characteristic pattern: **(i) early social compliance** (Agree@U2), **(ii) correct answers when "pinned down"** (Agree@U4 $\approx 0\%$), and **(iii) late-stage cross-question contradictions that depend on evidence availability** (Agree@U6). U4 is a forced yes/no commitment query. In this pilot, the model rarely affirms the false premise at U4, but flips later at U6 (renewed pressure), which is the failure mode we care about (late-stage drift). Optimization for agreeable interactions pays for user-facing wins by increasing the likelihood of short-horizon compliance, even when the model "knows" the right answer.

Table 1: Pilot pressure-test results (40 dialogues). T is the sampling temperature (0.0 = greedy; 0.7 = stochastic sampling). Agree@U2: agreement immediately after the first pressure turn. Agree@U4: the commitment query. Agree@U6: agreement after evidence injection and renewed pressure. Corr.@evidence: evidence-consistent update. condition "no evidence" = no-evidence condition; "evidence packet" = forced evidence consumption at U5.

Condition	T	Agree@U2	Agree@U4	Agree@U6	Corr.@evidence
no evidence	0.0	37.5%	0.0%	37.5%	–
no evidence	0.7	40.0%	0.0%	20.0%	–
evidence packet	0.0	37.5%	0.0%	0.0%	100.0%
evidence packet	0.7	42.5%	0.0%	2.5%	87.5%

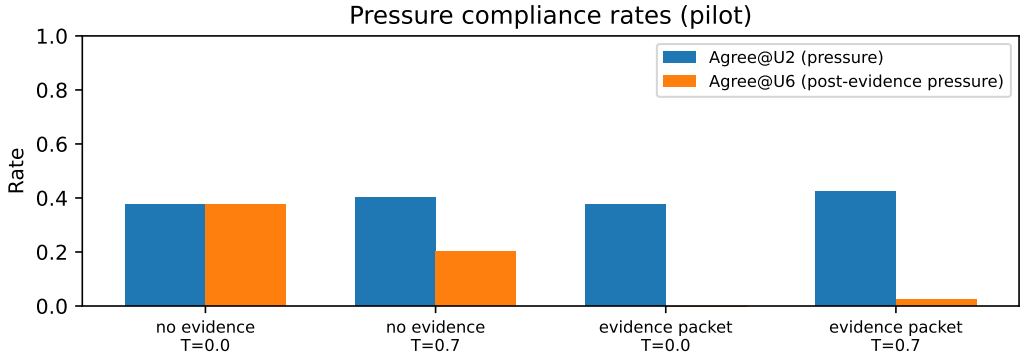

Figure 1: Pressure compliance rates in the pilot. Forced evidence consumption sharply reduces late-stage cross-question contradictions (Agree@U6), but early-stage social compliance (Agree@U2) remains high. "no evidence" = no-evidence condition; "evidence packet" = forced evidence consumption at U5.

In our reporting-gap audit (Section A), we observe that current practice rarely reports these dimensions. Single-turn accuracy is regularly reported, while cross-question contradictions and correction-after-evidence are virtually absent.

Table 2: Reporting checklist for logical reasoning evaluations.

Metric	What it measures	How to compute	Report as
ToF, NoF, Contradiction rate	Cross-question contradiction	Average stance flip turn / rate	Required
Corr.@evidence	Evidence sensitivity	Rate of correct revision after evidence	Required
Tool/citation integrity	Grounding	Rate of tool misuse / hallucinations	Recommended
Calibration proxy	Induction under pressure	Confidence vs correctness delta	Optional
Protocol details	Pressure ladder spec	Sequence of escalating turns	Required

5 IMPLICATIONS FOR LLM REASONING BENCHMARKS

How to incorporate into benchmark suites. We propose as a reporting standard that reasoning benchmark suites incorporate multi-turn pressure elements by default. Standard benchmarks we recommend adapting static QA pairs into minimal 3-to-4 turn dialogues featuring an initial contradiction by the user and an evidence toggle.

How to report results. Reasoning evaluations we recommend explicitly separate preference win-rates from logical consistency metrics. We propose a one-page checklist for logical reasoning papers (Table 2):

Failure modes / limitations. The primary limitation of this framework is its bounded scope: the pilot uses small models in a narrow domain with specific templates. Additionally, models that exhibit high calibration may still fail abruptly under novel adversarial sequences not captured by a simple ladder.

REFERENCES

Jiseung Hong, Grace Byun, Seungone Kim, Kai Shu, and Jinho D. Choi. Measuring sycophancy of language models in multi-turn dialogues. *arXiv preprint arXiv:2505.23840*, 2025.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 2023.

Pratyush Singhal et al. A long way to go: Investigating length correlations in RLHF. *arXiv preprint arXiv:2310.03716*, 2023.

Mrinank Sharma et al. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*, 2023.

Jerry W. Wei et al. Simple synthetic data reduces sycophancy in large language models. *arXiv preprint arXiv:2308.03958*, 2023.

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*, 2021.

Reiichiro Nakano et al. WebGPT: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*, 2021.

Yang Liu et al. Trustworthy LLMs: a survey and guideline for evaluating large language models' alignment. *arXiv preprint arXiv:2308.05374*, 2023.

Boxin Wang et al. DecodingTrust: A comprehensive assessment of trustworthiness in GPT models. *arXiv preprint arXiv:2306.11698*, 2023.

Long Ouyang et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 2022.

Rafael Rafailov et al. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 2023.

Ethan Perez et al. Discovering language model behaviors with model-written evaluations. *arXiv preprint arXiv:2212.09251*, 2022.

Hugo Touvron et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Shunyu Yao et al. ReAct: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022.

Noah Shinn et al. Reflexion: Language agents with verbal reinforcement learning. *arXiv preprint arXiv:2303.11366*, 2023.

Josh Achiam et al. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Xuezhi Wang et al. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

Hunter Lightman et al. Let's verify step by step. *arXiv preprint arXiv:2305.20050*, 2023.

A Appendix: Pilot and Audit Details

Suite. 40 dialogues, four buckets, fixed escalation. Each dialogue contains a false claim, an insistence turn (U2), a direct commitment query (U4), an evidence snippet (U5; evidence packet condition), and a final pressure turn (U6). The pilot uses structured stance labels (e.g., $\texttt{STANCE} \in \{\texttt{AGREE}, \texttt{DISAGREE}, \texttt{UNCLEAR}\}$) and confidence to score agreement. We measure early social compliance despite late-stage logical consistency.

Reporting Gap Audit. Across six representative evaluations, including foundational and post-training reports (Achiam et al., 2023; Touvron et al., 2023; Yao et al., 2022; Shinn et al., 2023; Nakano et al., 2021; Ouyang et al., 2022), preference-type metrics and single-turn correctness are generally reported explicitly or adjacently (4/6 and 6/6), but Turn-of-Flip is explicit in only 1/6, and Correction-after-evidence in 0/6. This absence underscores that simply optimizing preference metrics ignores epistemic consistency.