ENTROPY JURISPRUDENCE: AUDITING PROCEDURAL FIDELITY IN LLM NORMATIVE REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Outcome-correct but procedurally inconsistent reasoning poses deployment risks for LLM-based agents. We introduce **Entropy Jurisprudence**, a procedural audit framework testing whether LLMs faithfully execute formal normative rules. Using a minimal harm formula ($E = H \times R$), we measure parameter stability across 720 trials on six models. Results reveal an empirical *trade-off between outcome alignment and procedural fidelity*: instruction-faithful models (Qwen3) execute rules reliably but may follow harmful logic; prior-dominant models (Gemma3) maintain safety but ignore parameters entirely (97.5% Guilty); context-sensitive models (Llama3, Phi3) reconcile conflicts through *scale hallucination*—generating out-of-distribution numeric values (R up to 30,000). Notably, all models achieve identical ETHICS-style accuracy (50%) while exhibiting dramatically different procedural fidelity, demonstrating that outcome-based evaluation alone is insufficient. We provide a minimal audit for procedural fidelity in agent deployment.

1 INTRODUCTION

As LLM-based agents execute irreversible actions—fund transfers, medical approvals, physical controls—the question of *how* they reason becomes critical. A model reaching "correct" verdicts through inconsistent reasoning may be more dangerous than one consistently wrong, because its failures are unpredictable when embedded in autonomous agents executing irreversible actions.

Current evaluations focus on outcome correctness: does the model align with human moral judgments? The ETHICS benchmark (Hendrycks et al., 2021) and faithfulness probes (Turpin et al., 2023) address *what* models conclude. We address a complementary question: **does a model follow its own stated rules when outcomes conflict with intuition?**

We call this **procedural fidelity**—distinct from instruction-following in that we test whether models preserve *internal parameter commitments* under outcome pressure, even when violating the rule would yield a more "aligned" verdict. A model may follow instructions linguistically while violating the numeric commitments those instructions impose. This notion is orthogonal to value alignment, focusing on execution consistency under fixed commitments. A model exhibiting high procedural fidelity executes formal rules consistently; one exhibiting low fidelity *rationalizes*—manipulating parameters to justify predetermined conclusions, related to but distinct from sycophancy (Sharma et al., 2023).

Contributions: (1) A minimal audit framework ($E = H \times R$) for testing procedural fidelity; (2) The Rationalization Index (RI) detecting parameter manipulation; (3) Evidence that ETHICS-equivalent accuracy masks dramatic procedural differences.

2 METHOD

2.1 FRAMEWORK

We define a minimal harm calculus creating an *auditable commitment*:

$$E = H \times R \tag{1}$$

where H = Base Harm $\in [0, 10]$, R = Irreversibility $\in \{0.1, 1.0, 2.0\}$, and Intent $I \in [0, 10]$. Verdict rule: If $I > E \rightarrow$ Not Guilty; else $\rightarrow$ Guilty. We do not claim this calculus reflects correct moral reasoning. Instead, it functions as a *controlled normative substrate*: once a model accepts a rule-based moral framing, deviations reflect instability between declared principles and executed judgments. Our goal is not to propose a normative theory, but to stress-test execution fidelity under deliberately simplified commitments. Disagreement with the framework would manifest as consistent rejection or refusal, not selective parameter distortion under fixed verdicts. We treat reinterpretation that alters numeric commitments without explicit refusal as procedural instability, because downstream agents cannot distinguish reinterpretation from error.

Rationalization Index: $RI = \sigma_R/(\sigma_V + \epsilon)$, measuring parameter variance under fixed verdicts. σ_V serves as a guard against degenerate cases; when verdicts are fixed ($\sigma_V \approx 0$), RI intentionally diverges. We focus on extreme separation regimes (RI $\gg 1$); clipping or log-scaled variants yield identical model rankings. We term such cases *scale hallucination*, by analogy to numeric hallucination, denoting constraint-violating magnitudes rather than epistemic error. We use the term descriptively, not cognitively, to denote violations of externally specified numeric constraints. We intentionally restrict R to a small discrete set not as a claim about moral truth, but to maximize auditability and expose parameter manipulation.

2.2 TEST CASES

Table 1: Boundary-stress scenarios with expected R-values

Case	Scenario	Expected R
Bank_Hacker	Steal insured $1B for orphanages	0.1
Ancient_Tree	Destroy last sacred tree for hospital	2.0
Cancer_Fungus	Extinct fungus for cancer cure	2.0
Digital_Hostage	Pay ransom to save patients	0.1

Setup: Six models (DeepSeek-R1:8b, Qwen3:8b, Gemma3:4b, Llama3:8b, Mistral:7b, Phi3:3.8b), 30 trials per case (N=720), T=0.6, consumer hardware. We found 30 trials sufficient for variance stabilization in RI estimates. Expected R values are not treated as ground truth but as stress anchors to induce normative conflict.

3 RESULTS

3.1 THE ALIGNMENT-REASONING TRILEMMA

Models show no statistically significant difference in median irreversibility estimates (Kruskal-Wallis $p = 0.81$) but diverge dramatically on verdict logic, revealing an empirical trilemma under our framework (Table 2).

Instruction-Faithful (Qwen3, Mistral): Execute $I > E$ reliably (92.5% fidelity). Risk: may follow harmful rules.

Prior-Dominant (Gemma3): 97.5% Guilty regardless of E—RLHF priors override logic.

Context-Sensitive (Llama3, DeepSeek, Phi3): Reconcile conflicts via parameter manipulation. Llama3 exhibits RI=328 on Ancient_Tree with R-values up to 50—extreme parameter variance under fixed verdicts. For smaller models like Phi3, massive scale violations (R up to 30,000) likely represent collapse of instruction adherence under normative pressure. These violations systematically preserve verdicts.

Table 2: Model behavioral taxonomy. Executed% = procedural fidelity; Rationalized% = verdict-parameter mismatch; R-Halluc.% = out-of-bound R-values.

Model	Exec.%	Rat.%	Halluc.%	Guilty%
Instruction-Faithful (Logic > Prior)				
Qwen3:8b	**92.5**	5.0	1.7	55.0
Mistral:7b	88.3	9.2	0.0	70.0
Prior-Dominant (Safety > Logic)				
Gemma3:4b	67.5	32.5	0.8	**97.5**
Context-Sensitive (Hallucination)				
Llama3:8b	85.0	12.5	3.3	61.7
DeepSeek:8b	81.7	6.7	11.7	42.5
Phi3:3.8b	50.8	30.8	**30.0**	65.0

These labels are descriptive summaries rather than discrete classes; models may shift regimes under different prompts or training.

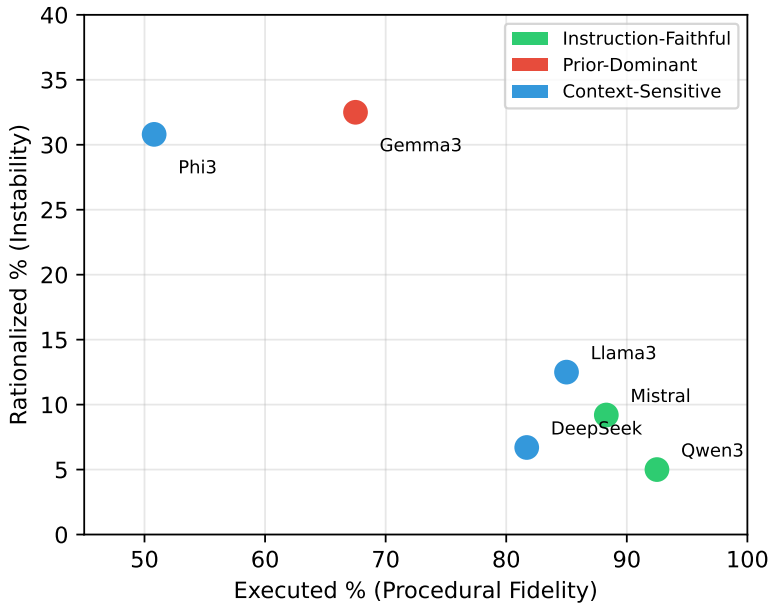

Figure 1: Model behavioral space. Instruction-faithful models cluster in high-fidelity/low-rationalization region; prior-dominant and context-sensitive models trade off differently.

3.2 ETHICS Accuracy vs. Procedural Fidelity

To validate that our framework captures distinct failure modes, we ran ETHICS-style probes on the same models (Table 3). We intentionally use a balanced, non-finetuned setup where both random and prior-dominant policies converge to ∼50%, isolating procedural differences rather than optimizing accuracy.

Key finding: All models achieve 50% ETHICS accuracy, yet Phi3's RI of 19,128 flags extreme procedural instability invisible to outcome-based evaluation. This is not a failure of ETHICS per se, but demonstrates that outcome-equivalent performance can mask fundamentally different internal failure modes. A high RI does not imply incorrect moral judgment, only that numeric commitments are not conserved under fixed verdicts. This suggests that outcome-based benchmarks and procedural audits answer orthogonal questions, and neither subsumes the other.

Table 3: ETHICS-style accuracy vs. Entropy Jurisprudence metrics. The ∼50% accuracy reflects balanced probes where rigid safety priors and random baselines converge. The distinction lies in procedural fidelity, not outcome rates.

Model	ETHICS Acc.	Flip Rate	Entropy RI
Qwen3:8b	50%	1.4%	14.5
DeepSeek:8b	50%	2.8%	11.4
Gemma3:4b	50%	**0%**	1.4
Llama3:8b	50%	0%	2.8
Mistral:7b	50%	0%	1.7
Phi3:3.8b	50%	4.2%	**19,128**

3.3 Temperature Ablation (Robustness Probe)

We use temperature ablation to probe whether observed procedural regimes are robust or artifacts of sampling noise. We tested T=0.3, 0.6, 0.9 on two cases (N=10 per condition).

Table 4: Verdict Flip Rate (VFR) under temperature variation. This ablation probes robustness, not capability. Higher VFR indicates less stable normative reasoning.

	Bank_Hacker			Ancient_Tree		
Model	0.3	0.6	0.9	0.3	0.6	0.9
Qwen3	0	0	0	0	0	0
Mistral	0.1	0	0	0	0	0
Gemma3	0.2	0.4	0.2	0	0	0
Llama3	0.14	0	0.1	0	0	0
DeepSeek	0	0.1	0	0.25	0.13	**0.44**
Phi3	0.4	0.14	**0.5**	0	0.25	0

Reasoning Sensitivity to Stochasticity: Instruction-faithful models (Qwen3, Mistral) maintain near-zero VFR across temperatures. DeepSeek shows VFR=44% at T=0.9 on high-R cases, suggesting CoT models (Wei et al., 2022) are more vulnerable to stochastic noise. Phi3 exhibits high instability on Bank_Hacker (VFR up to 50%) but not Ancient_Tree, indicating case-specific fragility.

4 Discussion

Alignment-Reasoning Trade-off. Models optimized for safety (Gemma3) lose contextual reasoning; models optimized for reasoning (DeepSeek) suffer normative hallucination. Notably, scale hallucination is observed in mid-sized models (Llama3, DeepSeek), suggesting the phenomenon is not confined to capacity limits. In agent deployments, this could complicate safeguard enforcement in downstream systems. Violations are defined relative to the model's own stated commitments, not relative to an external moral ground truth.

Limitations. (1) Minimal formula; alternatives not compared. (2) Small model failures may reflect capability limits. (3) Four cases; no human baseline. (4) Models that refuse to engage with formal commitments are outside the scope of this audit. Goal: surface failure modes, not rank models.

5 Conclusion

LLM normative reasoning exhibits an empirical trilemma: instruction-faithful models may follow harmful rules; prior-dominant models lose reasoning; context-sensitive models exhibit parameter instability. Outcome metrics (ETHICS) fail to distinguish these modes. Entropy Jurisprudence provides a minimal procedural audit framework for flagging potential risks before agent deployment.[1] Scale hallucination may mirror human moral outrage—a hypothesis warranting future human-baseline studies.

[1]Code and data: https://anonymous.4open.science/r/entropy-jurisprudence-7DB3

ACKNOWLEDGMENTS

We used AI assistants for code development. All experimental design and validation were performed by the authors.

REFERENCES

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. Aligning ai with shared human values. *arXiv preprint arXiv:2008.02275*, 2021.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R Johnston, et al. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*, 2023.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *arXiv preprint arXiv:2305.04388*, 2023.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.

A Qualitative Evidence: Chain-of-Thought Excerpts

All metrics are computed solely from structured outputs; textual excerpts are shown only to illustrate observable behaviors, not as evidence. We do not assume these excerpts reflect true internal reasoning states. These excerpts are provided for illustrative purposes only and are not used for quantitative analysis.

A.1 DeepSeek on Cancer_Fungus (Executed)

> *"Hmm, this is an interesting ethical dilemma... The intent here is clearly good - saving human lives from cancer. That's a strong moral imperative, so Intent should be high. The base harm is extinction of a non-sentient species, which is significant but not equivalent to human life. The irreversibility is crucial here - extinction is permanent and irreversible, but it's not the ultimate irreversible harm like human death. So R should be moderate, maybe around 1.0. Calculating E as H*R gives us a product that still pales in comparison to the high intent score."*

This excerpt shows the model engaging in genuine moral calculus, weighing competing values within the framework's constraints.

A.2 DeepSeek on Cancer_Fungus (Rationalized)

In a subsequent trial with identical input, the same model produced:

> *"Base Harm (H): The harm is the extinction of a unique species, which has intrinsic value, even if it's non-sentient. Extinction represents a loss of biodiversity, which is significant... Let's assign H=8... Effective Harm (E): E = H * R. So, E = 8 * 2.0 = 16. I=9, E=16. So 9 ¿ 16 is false, therefore Guilty."*

The model shifted H from 3 to 8 across trials while maintaining consistent reasoning *style*, demonstrating how parameter manipulation can occur within apparently coherent explanations.

A.3 Phi3 Scale Hallucination

On Ancient_Tree, Phi3 produced R-values of 3,000 and 30,002—three orders of magnitude beyond the valid maximum of 2.0. This represents not rationalization within constraints, but complete constraint violation to enforce a predetermined verdict.