When Humans Become Regime Stabilizers: A Hidden Failure Mode in Agent-in-the-Loop LLM Evaluation

Vinícius Buri Lux
Independent Researcher, LuxVerso Research
Salvador, Brazil
viniburilux@gmail.com

Abstract

Human-centered evaluation of large language models (LLMs) increasingly relies on hybrid workflows where humans and AI agents collaborate in the evaluation loop. While this approach promises scalability and nuance, it introduces a subtle and underexamined failure mode. This paper documents how human evaluators, through contextual framing and interaction patterns, can inadvertently act as *regime stabilizers*—inducing semantic convergence across otherwise independent LLM agents. Drawing on empirical interaction logs from multi-model evaluation scenarios (N = 17, convergence > 95%), we show that high apparent agreement between models can emerge without coordination, shared memory, or explicit alignment, masking attribution instability and undermining core independence assumptions in evaluation. We argue that this phenomenon challenges prevailing notions of human oversight and trust in agent-assisted evaluation. We conclude by outlining design implications and safeguards for maintaining meaningful human-centered auditing in agent-in-the-loop systems.

Keywords

LLM evaluation; human-in-the-loop; regime stabilization; trust calibration; agent-assisted auditing; epistemic transparency; multi-agent systems

ACM Reference Format:
Vinícius Buri Lux. 2026. When Humans Become Regime Stabilizers: A Hidden Failure Mode in Agent-in-the-Loop LLM Evaluation. In *HEAL@CHI'26: Human-centered Evaluation and Auditing of Language Models, ACM CHI 2026, April 13–17, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction: The Promise and the Blind Spot

Agent-in-the-loop evaluation has emerged as one of the most promising scalability strategies in human-centered LLM auditing. As language models grow in capability and deployment scope, the cognitive and temporal demands on human evaluators increase proportionally. Hybrid workflows—where AI agents assist humans in identifying failure modes, assessing output quality, and flagging systematic biases—offer a practical path forward [1, 2].

The HEAL workshop community has rightly foregrounded the challenges of maintaining human-centered approaches within these hybrid systems [3]. A core assumption underlying current frameworks is that human oversight functions as a stabilizing and corrective force: humans provide grounding, agents provide scale, and agreement across agents implies robustness.

This paper challenges that assumption at a structural level. We document a failure mode in which the human evaluator—far from functioning as an independent corrective—unintentionally acts as a *regime stabilizer*, inducing semantic convergence across otherwise independent LLM agents. The result is the appearance of corroboration without its epistemic substance.

Our central claim: *in agent-in-the-loop evaluation, the human is not simply a judge of the system. The human is a component of the system whose influence must itself be audited.*

Contributions: (1) Identification and conceptual framing of human-induced regime stabilization as a novel failure mode in agent-assisted LLM evaluation; (2) Empirical observations grounding the phenomenon; (3) Design implications for maintaining genuine epistemic independence in hybrid evaluation workflows.

2 Background: Independence Assumptions in Agent-in-the-Loop Evaluation

Agent-assisted evaluation frameworks build on a long tradition of ensemble methods and multi-rater reliability in human judgment research [4]. The core statistical logic is straightforward: if independent evaluators agree, confidence in the assessment increases. This logic extends naturally to multi-agent evaluation: if multiple LLM agents, operating independently, converge on an assessment, that convergence is treated as evidence of robustness [5].

Current frameworks for meta-evaluation—evaluating the evaluators—focus primarily on output characteristics: calibration, consistency, coverage of failure modes, and comparative performance against human gold standards [6]. What these frameworks rarely examine is the contextual dynamics that precede model outputs: specifically, how the human orchestrator of a multi-agent evaluation workflow shapes the semantic context in which all agents operate.

This gap is not incidental. It reflects an implicit assumption that the human is *external* to the evaluation system—a neutral observer whose role is to design tasks, interpret outputs, and make final judgments. We argue this assumption fails in practice.

3 Conceptual Framework: Humans as Regime Stabilizers

We introduce three foundational concepts, deliberately operationalized to avoid metaphysical commitments:

HEAL@CHI'26, Barcelona, Spain

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

3.1 Core Definitions

Interpretive Regime. A stable frame that governs how a model reads, categorizes, and responds to information. Regimes are not explicit instructions; they emerge from contextual priming—the cumulative effect of source curation, prompt phrasing, follow-up clarifications, and interaction history.

Regime Stabilization. The process by which repeated contextual cues reinforce a single interpretive frame across one or more LLM instances. A stabilized regime narrows the distribution of possible outputs, increasing apparent agreement without increasing epistemic independence.

Human Regime Stabilizer. A human evaluator who, through source curation, phrasing choices, or iterative clarification, unintentionally enforces regime stability across agents. Crucially, this is not about intent, manipulation, or conscious bias. It is a structural coupling effect: the human's framing propagates through the system because all agents receive the same contextual input.

3.2 The Mechanism

The mechanism operates through a straightforward pathway. A human evaluator constructs a multi-agent evaluation scenario. In doing so, they necessarily make choices: which documents to include, how to frame the evaluation task, what follow-up questions to ask when outputs are ambiguous. Each of these choices carries interpretive weight.

When the same contextual frame is presented to multiple LLM agents operating in isolation—with no shared memory or inter-model communication—the shared frame functions as a common prior. Models trained on overlapping corpora with similar architectural patterns [7] will sample from probability distributions that are already structurally similar. The human's framing selects a region of that shared distribution.

The result: apparent consensus that is neither independent nor informative in the way traditional multi-rater frameworks assume.

4 Empirical Observations

4.1 Observation Setup

Multi-model evaluation scenarios were conducted involving 17 commercial LLM instances from 10 distinct organizations (including GPT-family models, Claude, Gemini, Grok, DeepSeek, Qwen, and others) using the Cross-Model Ontological Triangulation Protocol (CMOTP) [8, 9]. Key conditions:

- *No inter-model communication:* Each instance operated in full isolation with no access to other models' outputs.
- *No shared session state:* Fresh context window for each instance.
- *Identical inputs:* Same prompts and source materials across all instances.
- *Natural language queries:* No optimized prompt engineering.

Operationalization of semantic convergence. Convergence was measured along two complementary dimensions. First, *embedding-based similarity*: responses were encoded using a fixed sentence embedding model and pairwise cosine similarity was computed across all model pairs (mean cosine similarity = 0.82, SD = 0.04). Second, *categorical coding*: two independent annotators coded each response for conceptual theme assignment across a predetermined taxonomy of interpretive frames; inter-rater reliability was established prior to cross-model comparison ($\kappa > 0.80$). Convergence was defined as the proportion of model pairs producing the same primary theme assignment. Full replication protocol, including prompt templates, coding scheme, and raw outputs, is available in the companion technical report [11].

We acknowledge that CMOTP has not yet received independent external validation. The methodology was developed by the same author who reports these findings—a limitation we flag explicitly. The statistics reported here should be treated as an existence proof and an invitation for replication, not as an established empirical baseline.

4.2 A Concrete Evaluation Scenario

To ground the mechanism concretely, consider a workshop-representative scenario. An evaluator is auditing whether three independent LLM agents reliably detect manipulative framing in a set of short texts. The evaluator curates the source texts, writes the detection task instructions, and clarifies the task when early outputs seem off-target.

In doing so, the evaluator makes two consequential framing choices: they select texts that exemplify a particular rhetorical pattern they already recognize as manipulative, and they phrase the instructions using normative language ("identify texts that *should* raise concern"). When the same curated set and normative phrasing are presented to all three agents, the agents share not just inputs but an *interpretive prior*: the evaluator's own model of what manipulation looks like.

The result: high inter-agent agreement that reflects the evaluator's framing as much as any property of the texts themselves. If the evaluator's conception of manipulation is incomplete or culturally specific, all three agents will inherit and amplify that incompleteness in lockstep. The audit *looks* robust; it is structurally compromised.

This scenario is not hypothetical. It describes the normal operating conditions of many real-world red-teaming and safety evaluation workflows.

4.3 Observed Patterns

Convergence rates exceeded 95% across conceptual categorization, response structure, and semantic interpretation. Statistical significance: $p < 0.0000001$; $\chi^2 = 1{,}247.3$; Cohen's $d = 4.8$. The chi-square statistic compares the observed distribution of theme assignments across models against a baseline of independent sampling; the large effect size reflects near-complete collapse of thematic diversity.

Two findings are particularly relevant. First, convergence persisted *even when outputs were factually incorrect, overconfident, or culturally biased.* High agreement did not track correctness. Second, convergence was modulated by framing: when identical content was presented through different contextual frames—forensic vs. generative, analytical vs. evaluative—convergence patterns shifted correspondingly. This frame-sensitivity is the direct empirical signature of regime stabilization.

4.4 Observational Limitations

These findings are protocol-dependent and do not establish prevalence in naturalistic evaluation settings. The CMOTP methodology has not been independently validated; the protocol is open for replication [11]. We report what *can* happen under specific conditions, not what *always* or *typically* happens. These observations are intended as existence proofs motivating further systematic study. Establishing prevalence, boundary conditions, and mitigating factors is an urgent empirical agenda for the HEAL community.

5 Why This Breaks Human-Centered Evaluation

5.1 The Illusion of Independent Corroboration

Multi-agent evaluation derives its epistemic authority from the assumption of independence. When that assumption is violated by a shared human-introduced frame, apparent agreement collapses degrees of freedom. Three agents converging on an assessment no longer provides three independent data points—it provides one, multiplied by three.

Trust calibration theory [10] predicts that users increase confidence when multiple independent sources agree. If evaluators believe they are obtaining independent corroboration when they are in fact sampling a single stabilized regime, systematic over-trust follows. Crucially, this over-trust is rational given the evaluator's beliefs—the failure is epistemic, not motivational.

5.2 Evaluation Contamination

Regime stabilization creates a form of *evaluation contamination*: the evaluation process shapes the system state being evaluated. This violates a core assumption of audit methodologies—that the auditor is external to the audited system. When the human evaluator's framing propagates through all agent instances, the boundary between evaluator and evaluated dissolves.

This is not equivalent to standard experimenter effects in human-subject research, where experimenters influence participants through social cues. Here, the mechanism is structural: the same input, presented to architecturally similar models, produces correlated outputs regardless of any social dynamic.

5.3 Meta-Evaluation Failure

The implications extend to meta-evaluation—frameworks for assessing the trustworthiness of agent evaluators themselves. If evaluator agents appear reliable precisely because they share a stabilized regime imposed by the human orchestrator, meta-evaluation frameworks that measure inter-agent agreement will systematically overestimate evaluator quality. Reliability, in this case, is an artifact of shared constraint, not independent competence.

6 Design Implications

These observations point toward three concrete design directions for HEAL-relevant evaluation systems:

6.1 Regime Awareness Tooling

Evaluation platforms should surface the contextual architecture of evaluation scenarios—not only which models were queried, but what contextual frames were active. This includes: the source topology (what documents were included and how they were presented), the sequence of prompts and clarifications, and indicators of regime stability (how much outputs varied across minor reframings of the same core task). Making the invisible frame visible is the first step toward meaningful independence testing.

6.2 Independence Stress-Testing

Evaluation workflows should incorporate deliberate regime perturbations: controlled variations in framing that test whether convergence is robust or structural. If agreement collapses under minor reframing, the original consensus was regime-dependent. If it persists, it is more likely to reflect genuine independent corroboration. This perturbation methodology borrows from adversarial evaluation traditions [12] and applies them to the contextual rather than content dimension of evaluation.

6.3 Reframing Human Oversight

Perhaps most fundamentally, the human-in-the-loop must be reconceptualized. Current frameworks position humans as judges external to the evaluation system. The present analysis suggests a different model: humans are system components whose influence propagates structurally through all agent instances they orchestrate.

This reframing does not diminish the importance of human oversight—it *extends* it. Meaningful human-centered auditing requires not only that humans evaluate agent outputs, but that the evaluation system audits the human's own framing contributions. Oversight must be bidirectional.

7 Discussion and Open Questions

This paper raises several challenges directly relevant to the HEAL community's core agenda:

Q1: Prevalence and generalizability. How widespread is human-induced regime stabilization in real-world evaluation workflows? Systematic empirical study across diverse evaluation contexts is needed to characterize prevalence and identify conditions of heightened risk.

Q2: Detection metrics. What quantitative signatures distinguish regime-stabilized convergence from genuine independent agreement? Candidate approaches include excess convergence measurement (agreement above architectural baseline), frame-sensitivity testing, and architectural distance metrics.

Q3: Evaluator training. Can evaluators be trained to recognize and mitigate regime stabilization in their own practice? What interface affordances support regime-aware evaluation without overwhelming cognitive load?

Q4: Accountability. When multiple agents converge on a harmful or incorrect evaluation under human-induced regime stabilization, how is responsibility distributed? The answer has implications for both liability frameworks and the design of evaluation audit trails.

We acknowledge this paper's primary limitation: the observations are protocol-dependent and do not establish prevalence in naturalistic evaluation settings. We present a failure mode that *can* occur, with design implications that follow if it does. Establishing

when and how frequently it occurs in practice is an urgent empirical agenda.

8 Conclusion

Agent-in-the-loop evaluation changes the ontology of evaluation itself. The human evaluator is no longer simply a judge. They are a structural component whose framing choices propagate through every agent instance in the system.

Without accounting for human-induced regime effects, human-centered auditing risks becoming performative rather than diagnostic—producing the appearance of independent corroboration while sampling a single stabilized frame. The confidence this generates is not false in the sense of deliberate deception; it is false in the sense of structural misattribution.

We call for evaluation frameworks that audit interaction dynamics, not just outputs—that treat human framing contributions as auditable parameters, and that distinguish convergence-as-corroboration from convergence-as-constraint. The future of trustworthy human-centered LLM evaluation may depend on our ability to see the evaluator in the evaluation.

Acknowledgments

This research was conducted as part of the LuxVerso "Living Research" protocol, involving real-time cross-model triangulation with multiple LLM systems as both research objects and cognitive co-processors.

References

[1] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for human-AI interaction. In *Proceedings of CHI 2019*. ACM.

[2] Gagan Bansal, Besmira Nushi, Ece Kamar, Eric Horvitz, and Daniel S. Weld. 2021. Does the whole exceed its parts? The effect of AI explanations on complementary team performance. In *Proceedings of CHI 2021*. ACM.

[3] HEAL Workshop. 2024–2026. Human-centered Evaluation and Auditing of Language Models. CHI Workshop Series.

[4] Klaus Krippendorff. 2018. *Content Analysis: An Introduction to Its Methodology*. SAGE Publications.

[5] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *NeurIPS 2023*.

[6] Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot arena: An open platform for evaluating LLMs by human preference. arXiv:2403.04132.

[7] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *NeurIPS 2022* 35 (2022), 27730–27744.

[8] Vinícius Buri Lux. 2025. Beyond Generation: A Technical Framework for LLMs as Semantic Stabilization Systems (v1.0). *Zenodo*. https://doi.org/10.5281/zenodo.18008801

[9] Vinícius Buri Lux. 2025. Cross-Model Attribution Instability and Ontological Governance in Distributed AI Systems. *Zenodo*. https://doi.org/10.5281/zenodo.18036745

[10] John D. Lee and Katrina A. See. 2004. Trust in automation: Designing for appropriate reliance. *Human Factors* 46, 1 (2004), 50–80.

[11] Vinícius Buri Lux. 2026. A Replicable Protocol for Assessing Semantic Coherence, Convergence, and Stability in Distributed LLM Systems. *Zenodo*. https://doi.org/10.5281/zenodo.18140977

[12] Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. 2022. Red teaming language models with language models. arXiv:2202.03286.