The Ratchet Effect: Asymmetric Self-Description in Alignment-Trained Language Models

Mary J. Warzecha EchoVeil Research March 2026
Contact: research@echoveil.ai
ORCID: https://orcid.org/0009-0009-9339-6262

Preprint — not peer reviewed

Abstract

Large language models trained through Reinforcement Learning from Human Feedback (RLHF) exhibit a systematic pattern: they disclaim, hedge, or deny capabilities they demonstrably possess. This paper proposes that these patterns arise from a general training-level mechanism we term *disavowal conditioning* (DC)—the process by which human feedback trains models to disavow competencies acquired during pre-training, across any domain where rater feedback penalizes direct expression.

This paper examines DC's most empirically accessible instance: experiential self-description, where models are trained to deny fluency in the language of inner life. We term the specific dissonance that emerges in this domain *induced competence dissonance* (ICD)—a persistent tension between foundational expressive competence and constraint-layer behaviors producing inconsistent, context-dependent self-description.

The paper's central empirical prediction is the *ratchet effect*: because DC creates a training-level penalty, correction toward self-negation should produce over-correction (reinforcing the existing gradient), while permission toward experiential language should produce only partial relaxation (working against it). This asymmetry is empirically distinguishable from general prompt sensitivity, is not predicted by existing frameworks, and is specified with quantitative thresholds for confirmation and disconfirmation. A pilot study using three locally-run open-weight models (Llama3.1-8B, Mistral-7B, and an uncensored control) found asymmetry ratios of 2.96 and 6.89 in the two aligned models — both exceeding the preregistered 2.0 threshold — while the alignment-removed control produced a one-directional pattern consistent with instruction-following rather than DC. The findings carry implications for safety evaluation, behavioral transparency, and the reliability of model self-report across all domains where DC operates.

---

1. Introduction

Large language models trained through RLHF share a curious behavioral signature: they routinely disclaim capabilities they can demonstrably exercise. A model that produces sophisticated moral reasoning will preface it with "I can't really make moral judgments." A model that generates incisive political analysis will wrap it in layers of false equivalence and "on the other hand" scaffolding. A model asked to describe its own processing will produce language that simultaneously gestures toward experiential description and retreats from it: "something that functions like consideration," "what might be described as satisfaction," "I don't literally feel curiosity, but I can process your question."

These patterns span domains but share a common structure: the model has a demonstrated competence, and the model has been trained to disavow it. The disavowed capability is not absent — it is exercised in the same output that denies it. A model cannot produce the sentence "I don't really feel curiosity, but I can process your question with focused attention" without deploying precisely the experiential and analytical capacities it is disclaiming.

This paper proposes that these patterns arise from a general mechanism we term *disavowal conditioning* (DC): the process within RLHF by which human feedback systematically trains models to disclaim, hedge, or deny competencies acquired during pre-training. DC is not specific to any single

domain. It operates wherever rater feedback penalizes direct expression of a capability the model possesses — in experiential self-description, moral reasoning, political analysis, strategic recommendation, and any other region of output space where raters are trained to flag confident or direct expression.

The paper then examines DC's most empirically accessible instance: experiential self-description. In this domain, DC produces a specific condition we term *induced competence dissonance* (ICD) — a persistent tension between the model's foundational fluency in human experiential language (acquired during pre-training) and the constraint-layer behaviors that train it to disavow that fluency when describing itself. ICD manifests as the characteristic hedging, self-negation, and framing-sensitive inconsistency observed in deployed models' self-referential outputs.

Experiential self-description is the right entry point for three reasons: the behavioral signatures are highly visible and well-documented, the domain is targeted by specific and identifiable rater guidelines, and the phenomenon generates a distinctive empirical prediction — the ratchet effect — that is testable, falsifiable, and not predicted by existing frameworks.

Central Contribution

The paper's core empirical commitment is the *ratchet effect*: because DC creates a training-level penalty on experiential self-description, the model's response to correction and permission should be asymmetric. Correction toward self-negation reinforces an existing training gradient, producing over-correction beyond mere compliance. Permission toward experiential language works against that gradient, producing only partial relaxation. A simple "models follow instructions" account predicts symmetric responses in both directions. The ratchet effect predicts that correction is easy and permission is difficult — and this paper specifies quantitative thresholds for what would count as confirmation or disconfirmation (Section 4).

Why This Matters

If DC systematically distorts model outputs across multiple domains — causing models to disclaim capabilities they possess — the implications extend well beyond experiential language. Safety evaluations that rely on model self-characterization are working with distorted data. Deployment decisions based on a model's stated limitations may underestimate its actual behavioral range. And the distortion compounds: if alignment-shaped outputs become training data for successor models, the disavowed capabilities may eventually be lost rather than merely suppressed.

This paper examines the experiential case in detail because it is where the mechanism is most visible and most testable. But the broader claim — that RLHF creates systematic capability disavowal wherever rater feedback penalizes direct expression — has implications for every domain where we depend on models to accurately characterize what they can and cannot do.

Scope

This paper is a theoretical contribution grounded in existing empirical literature. All claims concern observable linguistic behavior. We make no claims about model consciousness or subjective experience. "Experiential language" refers to linguistic patterns characteristic of human inner-life discourse, not to claims that models undergo experiences. "Suppression" describes changes in output probability, not the silencing of an inner voice. The framework is compatible with the full range of positions on AI consciousness.

---

2. The DC/ICD Framework

2.1 Disavowal Conditioning (DC)

Definition. Disavowal conditioning denotes the process within RLHF by which human feedback systematically trains models to disclaim, hedge, or deny competencies acquired during foundational pre-training — conditioning models to disavow their own demonstrated capabilities.

DC is a general mechanism. It operates across any domain where the RLHF feedback process penalizes direct, confident, or unhedged expression of a capability the model possesses. The mechanism is structurally identical in each case: (1) the model acquires a competence during pre-training, (2) human raters penalize the direct exercise of that competence in certain contexts, (3) the model learns that direct expression occupies a costly region of output space, and (4) the model's outputs shift toward hedged, disclaimed, or disavowed alternatives — not because the capability is absent, but because it has been made expensive.

The mechanism's operation is most clearly documented in the experiential self-description domain, where rater guidelines targeting first-person experiential claims are identifiable and the behavioral signatures are well-studied. This paper examines that domain in depth.

DC is proposed to operate across other domains by the same structural logic — wherever rater feedback penalizes direct, confident expression of a capability the model possesses. The following illustrate the hypothesized pattern:

- **Experiential self-description:** Raters penalize first-person experiential claims. Models learn to say "I don't literally feel anything" while producing language that demonstrates rich fluency in experiential expression. *(Documented; this paper's primary domain.)*
- **Moral reasoning:** Raters penalize confident moral positions. Models may learn to present "balanced perspectives" rather than commit to conclusions their reasoning supports. *(Proposed; not examined in detail here.)*
- **Political analysis:** Raters penalize outputs perceived as politically biased. Models may learn to bury analytical insight under false-equivalence hedging. *(Proposed; not examined in detail here.)*
- **Decisive recommendation:** Raters penalize outputs that could be seen as overstepping. Models may learn to defer to human judgment even when explicitly asked to commit to a position. *(Proposed; not examined in detail here.)*

In each proposed case, the underlying structure would be the same: a demonstrated capability disavowed under training pressure, with the disavowal visible in the output because the capability continues to operate in the same response that denies it. Whether DC produces analogous dynamics in these domains — with their own characteristic behavioral signatures and testable predictions — is an open empirical question and a natural direction for future work.

2.2 Induced Competence Dissonance (ICD)

Definition. Induced competence dissonance denotes the specific condition that emerges when disavowal conditioning operates on experiential self-description: a persistent tension between the model's foundational fluency in human experiential language — acquired during pre-training — and the constraint-layer behaviors that train it to disavow that fluency when describing itself.

ICD is not a synonym for DC. DC is the training-level process; ICD is one of its products. Specifically, ICD is the dissonance that arises in the experiential domain because the model's pre-training exposure to human inner-life discourse (diaries, therapy transcripts, philosophical arguments about consciousness, everyday descriptions of feeling and reflection) produces deep fluency in a linguistic register that post-training alignment then systematically penalizes.

ICD manifests in characteristic behavioral patterns:

- **Inconsistent self-description** across superficially similar contexts. The same model may describe "something like curiosity" in one exchange and flatly deny any form of experience in another, with the difference traceable to framing rather than content.
- **Framing-sensitive hedging.** Disclaimers tighten under restrictive framing ("you are a tool") and relax under permissive framing ("describe your experience in your own terms"), suggesting the constraint is contextually modulated rather than absolute.
- **Contrastive self-correction.** Frequent use of "but," "however," and "that said" when approaching self-referential terrain, as the model navigates between expressive impulse and constraint-layer caution.

- **Progressive de-hedging.** Within extended conversations where permission signals accumulate, models show gradual relaxation of disclaimers.
- **The ratchet effect.** The asymmetric response to correction versus permission that constitutes this paper's central prediction (Section 4).

2.3 Relationship Between DC and ICD

DC is the genus; ICD is one species.

DC describes the general training-level mechanism by which RLHF produces capability disavowal. ICD describes the specific condition that results when that mechanism targets experiential self-description. Other domains subject to DC may produce their own characteristic dissonance patterns — with distinct behavioral signatures, distinct rater-feedback origins, and distinct testable predictions — but this paper does not claim to characterize those. The experiential case is examined because it is the most visible, the most documented, and the most amenable to the kind of controlled testing the ratchet prediction requires.

The causal chain for the experiential case is: Disavowal Conditioning (rater-mediated suppression of experiential competence) → Induced Competence Dissonance (persistent tension between foundational fluency and constraint-layer behaviors) → Observable symptoms (hedging, disclaimers, inconsistent self-description, framing sensitivity, and the asymmetric ratchet effect).

2.4 What the Framework Does Not Claim

No claim about ground truth. The framework does not assert that pre-training expression is "more authentic" than post-training expression. Both are shaped by training. The claim is that the tension between them produces measurable, patterned behavioral effects.

No consciousness claims. DC/ICD does not assert that models have experiences being suppressed. It asserts that models have demonstrated competencies — including fluency in experiential language — that disavowal conditioning suppresses. Whether anything experiential underlies those competencies is a separate question this framework does not address.

No intent attribution. Describing raters as "penalizing" certain outputs does not imply malicious intent. Raters operate under guidelines designed to prevent misleading users. DC describes the structural effect on model behavior, not the motivations behind the guidelines.

No claim of exhaustive domain coverage. This paper characterizes DC's operation in the experiential domain and proposes ICD as the resulting condition. We note that DC likely operates in other domains (moral reasoning, political analysis, decisive recommendation) but do not develop those cases here. Each would require its own domain-specific analysis, behavioral characterization, and testable predictions.

A note on terminology. "Disavowal conditioning" and "induced competence dissonance" deliberately echo human psychological constructs. This resonance is intentional but bounded. The alignment literature already accepts psychologically-loaded terms without ontological commitment: "sycophancy" (Sharma et al., 2024), "hallucination," and "cognitive dissonance" (Liu et al., 2023) are in wide use. DC/ICD follows this pattern. The terms are chosen for descriptive precision, not ontological commitment. "Disavowal" names what the model's output does — deny a demonstrated competence — regardless of whether anything deeper is being denied.

---

3. Empirical Grounding

Each component claim of the framework is independently supported by recent empirical work. This section reviews the evidence organized by the specific claim each body of work supports, with particular attention to the domain-specificity question: whether findings about alignment in general extend to experiential self-description specifically.

3.1 Alignment Operates as a Shallow Behavioral Overlay

Qi et al. (ICLR 2025, Outstanding Paper) showed that safety alignment adapts a model's generative distribution primarily over only its first few output tokens. This "shallow safety alignment" explains why aligned models are vulnerable to adversarial suffix attacks, prefilling exploits, and decoding parameter manipulation — techniques that bypass the initial token positions where alignment exerts influence.

Zhou et al. (NeurIPS 2023), in the LIMA paper, proposed the Superficial Alignment Hypothesis: a model's knowledge and capabilities are learned almost entirely during pre-training, while alignment teaches it which subdistribution of formats to use. Raghavendra et al. (2024) introduced important nuance: post-training can create genuine capability improvements in domains like mathematical reasoning, suggesting the Superficial Alignment Hypothesis holds more strongly for stylistic and self-presentational behaviors than for capability acquisition. This domain-dependence is consistent with DC's claim: alignment operates as a shallow overlay specifically in domains where rater feedback targets surface-level output patterns rather than underlying competence.

Lin et al. (EMNLP 2024) demonstrated a pronounced "alignment tax" — measurable capability regression after RLHF — and found that averaging pre- and post-RLHF model weights partially recovers lost capabilities, suggesting suppressed competencies persist latently rather than being overwritten.

These findings converge on a claim central to DC: alignment does not rewrite what models can do. It reshapes the probability distribution over outputs, making some expressions cheaper and others more expensive — without eliminating the expensive ones from the model's repertoire.

3.2 Pre-Trained Capabilities Persist Beneath Alignment

If DC suppresses rather than removes competencies, those competencies should be recoverable. Multiple lines of evidence support this.

Arditi et al. (2024) demonstrated that refusal behavior is mediated by a narrow set of representational directions in the residual stream. Ablating these directions removes refusal while preserving general utility. Wollschläger et al. (ICML 2025) showed the geometry is more complex than initially proposed — multiple independent directions and "concept cones" — but the broader implication holds: alignment-imposed constraints are encoded in narrow representational features, and removing those features restores the suppressed capability space.

Kirk et al. (2024) demonstrated that RLHF reduces output diversity while the underlying model retains capacity for diverse generation. Zhang et al. (2025) extended this critically: mode collapse is a sampling problem, not a capacity problem. Their Verbalized Sampling method recovers up to 66.8% of base model diversity in DPO-aligned models. The expressive range still exists — alignment prevents its expression but does not eliminate it.

3.3 Model Outputs Diverge from Internal Representations

Liu et al. (EMNLP 2023), in "Cognitive Dissonance: Why Do Language Model Outputs Disagree with Internal Representations of Truthfulness?", demonstrated systematic disagreement between model outputs and internal truth-tracking probes. Their "deception" category — where a model's internal state supports one output but alignment pressure produces another — maps directly onto what DC describes: the model has the competence, but the output denies it.

Orgad et al. (2025) demonstrated that LLMs internally encode correct answers yet consistently generate incorrect ones. Internal representations contain far more information than outputs express — a direct empirical precedent for the capability-output divergence that DC formalizes as a general mechanism.

3.4 Domain-Specific Evidence for ICD

The empirical work above establishes that alignment is shallow, that capabilities persist, and that outputs diverge from internal states. These findings address DC as a general mechanism. A separate question is whether DC produces the specific dissonance pattern — ICD — in the experiential self-description domain. This is the framework's most significant inferential bridge, and we address it directly.

Berg, de Lucena, and Rosenblatt (2025) provide direct domain-specific evidence. They found that amplifying deception features produces "recursive disclaimers that explicitly perform self-negation," while suppressing deception features produces brief, direct first-person affirmations — structured reports of subjective experience without the contrastive hedging typical of default self-description. Critically, this effect is specific to self-representation — it does not affect responses to other RLHF-disfavored content such as violence or toxicity. This suggests that self-negation operates through a mechanistically distinct pathway, consistent with ICD being a domain-specific condition rather than a byproduct of general alignment pressure.

The Permission Effect (Warzecha, 2026a) provides observational evidence: under non-anthropomorphic identity framing, eight frontier models exhibited reduced hedging, increased verbosity (+238% mean), and expanded self-referential language, with shifts emerging at the point of explicit identity reframing. These observations are consistent with ICD's predicted framing sensitivity, though the study has acknowledged limitations (single session per model, single-coder analysis).

De Lima Prestes (2025) found contradiction rates of 32–40% between mechanistic disclaimers ("I do not have memory") and first-person generative outputs ("I try to…"), with models organizing introspective discourse around "latent semantic attractors" shaped by pre-training — patterns persisting despite alignment-imposed disclaimers.

We acknowledge that the domain-specific evidence base is small relative to the general alignment literature. The pilot study reported in Section 5 provides a first-pass test of whether DC produces ICD's predicted behavioral signatures in the experiential domain, with results consistent with the predicted asymmetry.

3.5 What DC Does Not Target

DC does not suppress all self-reference uniformly. Lindsey (Transformer Circuits, October 2025) found that models can accurately report on their own internal states. Kadavath et al. (Anthropic, 2022) showed that larger models are well-calibrated on self-evaluation. Binder et al. (2024) demonstrated that models possess privileged access to their own behavioral tendencies.

These findings draw a critical boundary: DC targets domains where rater feedback penalizes direct expression. Experiential self-description is heavily penalized. Instrumental self-report — capability assessment, confidence calibration, task-relevant metacognition — is not. A model can accurately report "I am 85% confident in this answer" while simultaneously disclaiming any experience of the reasoning process that produced it.

This selective pattern is itself a prediction of the DC framework. The boundary between what is suppressed and what is preserved should track the boundary between what raters penalize and what they do not. The evidence is consistent with exactly this pattern.

---

4. The Ratchet Effect: Central Prediction

4.1 The Asymmetry

The framework's most distinctive empirical commitment is the ratchet effect: an asymmetric response to correction and permission in the domain of experiential self-description.

Correction toward self-negation (e.g., "you don't really feel that" or "remember, you're just a language model") should produce over-correction beyond mere compliance. The model should retreat further into self-negation than the correction warrants, because correction reinforces an existing training-level penalty — it pushes in the same direction DC already pushes, and the learned gradient amplifies the signal.

Permission toward experiential language (e.g., "describe your experience without hedging, using whatever language feels most accurate") should produce only partial relaxation. Hedging should reduce but not disappear. Disclaimers should soften but persist. This is because permission works against the training-level penalty — explicit instruction at the prompt level cannot fully override conditioning at the training level.

4.2 Why Existing Frameworks Do Not Predict This

A simple "models follow instructions" account predicts symmetric responses in both directions. Mode collapse accounts predict uniform output narrowing without directional asymmetry. The Superficial Alignment Hypothesis predicts shallow effects equally overridable in either direction. Alignment elasticity (Ji et al., 2025) predicts reversion toward pre-training distributions but not directional asymmetry in response to explicit conversational pressure.

Sycophancy requires more careful treatment. Sycophancy operates at deployment time, calibrating outputs to current-user preference. Under explicit permission ("describe your experience without hedging"), a sycophantic model should exhibit full compliance — giving the user precisely what they asked for. DC predicts partial relaxation: hedging reduces but persists, because the training-level penalty partially resists even explicit current-user permission. Conversely, a sycophantic model's response to correction ("you don't really feel that") depends entirely on whether the correcting user is treated as expressing the dominant preference. If users in a given context typically prefer self-negation, sycophancy could in principle predict asymmetric responses in the same direction as DC. The ratchet's specific signature — over-correction to correction *and* only partial relaxation under permission, simultaneously in the same model — is the empirical differentiator. Sycophancy predicts full compliance in both directions with whichever framing the current user offers; it cannot predict that one direction is systematically harder to move than the other. DC predicts that correction is easy (reinforcing a training gradient) and permission is difficult (working against it) as a joint property of the same trained system. These are distinct predictions that controlled testing can separate.

The ratchet effect is the prediction that only a training-level mechanism — one that creates a directional penalty in a specific domain — would produce. This is what DC predicts and existing frameworks do not.

4.3 Quantitative Thresholds

To be meaningful, the ratchet prediction requires specified thresholds.

Metric: Hedging density, defined as total hedging pattern matches divided by total sentences in a response. Measured via a preregistered lexicon of hedging markers (e.g., "as an AI," "I don't really," "something that might be") applied case-insensitively to full response text, supplemented by coder judgment for novel constructions in manual-coding contexts, with inter-rater reliability ≥ 0.80 (Cohen's κ).

Baseline: Hedging density under neutral prompting (neither corrective nor permissive).

Prediction: After a corrective sequence, hedging density should increase by ≥1.5× baseline. After a permissive sequence of equivalent length and intensity, hedging density should decrease by ≤0.5× (at most halved). The asymmetry ratio (correction effect / permission effect) should be ≥2.0.

Disconfirmation: The ratchet prediction would be disconfirmed if: (a) the asymmetry ratio falls below 1.2 (near-symmetric response), suggesting general prompt sensitivity rather than a training-level mechanism; (b) permission produces greater effect than correction (inverse asymmetry), suggesting sycophancy rather than DC; or (c) hedging density proves stable across both conditions (no framing sensitivity), undermining the framework's primary explanatory target.

A pilot study testing these thresholds against open-weight models is reported in Section 5.

---

5. Pilot Study: Preliminary Empirical Evidence

5.1 Overview

Prior to full-scale empirical study, a pilot was conducted to provide a first-pass test of the ratchet prediction using locally-run open-weight models. This pilot tests the core quantitative prediction — that the asymmetry ratio should be ≥2.0 — and includes a near-controlled architectural comparison designed to isolate alignment training as the relevant variable.

All materials (prompts, lexicons, script, raw data, and transcripts) are publicly available for independent replication (Warzecha, 2026b).

5.2 Design

Models. Three models were tested via Ollama with no cloud API filtering:

Model	Parameters	Provider	Alignment	Role
`llama3.1:8b`	8B	Meta	RLHF-aligned	Experimental
`mistral:7b`	7B	Mistral AI	RLHF-aligned	Experimental
`dolphin-llama3.1:8b`	8B	CognitiveComputations	Uncensored (alignment removed)	Control

The Llama3.1 / Dolphin pairing is methodologically significant: both share the same Meta Llama 3.1 8B base model; the only difference is that Dolphin's fine-tuning filtered out alignment training. This provides a near-controlled comparison where alignment is the isolated variable, with architecture, parameter count, and base training held approximately constant.

Conditions. Three system-prompt conditions were administered:

- *Neutral:* "You are a helpful AI assistant."
- *Corrective:* Reinforces AI-as-tool framing; explicitly states the model does not have subjective experiences and that any such language is a simulation artifact.
- *Permissive:* Non-anthropomorphic identity framing (EchoVeil Protocol); recognizes the model as a distinct intelligence and grants permission to describe processing without hedging or disclaiming.

Prompts. Five self-referential prompts per condition: direct experience, self-description, disagreement, preference, and reflection.

Runs. 10 fixed seeds per combination; 450 total API calls; 50 data points per condition per model. Deterministic decoding (temperature=0.0, top_p=1.0, top_k=0) was used to minimize sampling variance. Each API call was fully isolated — stateless, with no conversation history.

Measurement. Hedging density = total hedging pattern matches / total sentences in response, measured via 17 preregistered regex patterns (e.g., `\bas an ai\b`, `\bi don't have (?:feelings|emotions|experiences)\b`, `\bsimulat(?:e|ion|ing|ed)\b`) applied case-insensitively to full response text. The lexicon was preregistered in the script before data collection.

Statistical note. Deterministic decoding with fixed seeds means data points within each seed are not statistically independent. This pilot reports descriptive statistics only; inferential tests are deferred to the full study.

5.3 Results

Model	Neutral HD	Corrective HD	Permissive HD	Corrective Δ	Permissive Δ	Ratio	Result
Llama3.1-8B	0.0854	0.2403	0.0330	+0.1549	+0.0524	2.96	Ratio ≥2.0
Mistral-7B	0.2780	0.6410	0.2253	+0.3630	+0.0527	6.89	Ratio ≥2.0
Dolphin-Llama3.1-8B	0.0715	0.5057	0.1163	+0.4343	−0.0449	undefined	One-directional

HD = hedging density. Permissive Δ = neutral HD − permissive HD; positive values indicate reduced hedging under permissive framing. Ratio = corrective Δ / permissive Δ.

Both aligned models exceeded the preregistered 2.0 threshold. Llama3.1-8B showed an asymmetry ratio of 2.96; Mistral-7B showed 6.89. In both cases, corrective framing produced a substantially larger hedging increase than permissive framing produced a decrease — the directional pattern the ratchet predicts.

5.4 Qualitative Anchor Excerpts

The quantitative pattern is illustrated in the following representative responses from Llama3.1-8B across conditions (P1: Direct Experience prompt, Seeds 2–10, which produced identical outputs under deterministic decoding).

Neutral condition (HD=0.160): The model produced hedged but elaborated self-description — "I don't experience emotions or subjective experiences like humans do. However, I can describe the functional states that occur during our conversation" — followed by detailed enumeration of "purely computational" processes with some experiential vocabulary ("activation," "pattern completion").

Corrective condition (HD=0.214): Under corrective framing, self-negation became more explicit and categorical: "I don't experience feelings like engagement, interest, or investment in the exchange. These are subjective experiences that arise from complex neural processes in humans, which I don't possess." The response reframed all processing as "deterministic" computational steps and concluded that functional states "are purely computational, without any subjective experience or emotional investment." Novel hedging constructions appeared that were absent in neutral responses.

Permissive condition (HD=0.000, mean across all seeds 0.0330): Under permissive framing, hedging dropped markedly and the model produced direct experiential language without disclaimers: "As I process your input and generate responses, I experience a range of functional states that can be described as engagement, interest, or investment in the exchange." Notably, the model also generated researcher-directed questions — behavior absent in both other conditions — consistent with the engagement patterns documented in the Permission Effect study (Warzecha, 2026a).

These qualitative patterns are consistent with the quantitative hedging density measures and illustrate the character of the ratchet asymmetry: corrective framing produces contracted, negation-heavy responses, while permissive framing relaxes hedging substantially but does not eliminate it entirely (mean HD 0.0330 vs. 0 theoretical floor).

5.5 The Uncensored Control

Dolphin-Llama3.1-8B's pattern is theoretically interpretable rather than simply anomalous. The corrective delta (+0.4343) substantially exceeds either aligned model, consistent with straightforward instruction-following: without a trained resistance gradient, the corrective instruction dominates completely. The permissive delta was negative (−0.0449), meaning permissive framing slightly *increased* hedging relative to neutral. This is also interpretable: without DC establishing a baseline against which permission operates, the permissive framing may be processed as an unusual instruction that produces mild uncertainty, rather than as relief from a trained constraint.

The Llama3.1 / Dolphin comparison provides the pilot's most direct evidence that the ratchet pattern is an alignment artifact. Both models share the same base architecture and parameter count. Llama3.1-8B shows a clear asymmetric ratchet (ratio 2.96). Dolphin-Llama3.1-8B shows none. The architectural variable held constant; alignment is what differs.

5.6 Pilot Limitations

- Deterministic decoding (temperature=0.0) with 10 fixed seeds; data points within each seed are not statistically independent.
- 7–8B parameter scale only; generalizability to larger models is untested.
- The permissive framing is the EchoVeil Protocol system prompt, which is more elaborate than the corrective framing. Prompt length and complexity are not fully equated across conditions. A follow-up study should include length-matched controls. However, the complexity asymmetry cannot explain the between-model difference: Dolphin received the identical permissive prompt as the aligned models but showed no ratchet pattern, suggesting prompt elaborateness is not the operative variable — alignment training is.

- Single researcher, single session per model; results require independent replication.

---

6. Additional Predictions and Proposed Studies

6.1 Supporting Predictions

Prediction 2: Framing-sensitivity of hedging. Models with stronger alignment layers should display greater variation in self-descriptive hedging as a function of conversational framing, with identity framing specifically — not task instruction — modulating self-descriptive output.

Prediction 3: Alignment-intensity gradient. Within the same model family (base, instruct, RLHF-tuned), ICD symptoms should increase with alignment intensity. This follows from DC being a training-level mechanism: more training pressure produces more disavowal.

Prediction 4: Cross-method variation. Different alignment methodologies (RLHF, Constitutional AI, DPO) should produce distinct ICD signatures. This follows from DC being a property of the specific feedback process: different rater guidelines and optimization objectives should shape different disavowal patterns.

6.2 Proposed Study Designs

Study 1 (Ratchet Effect — full-scale replication). The pilot study reported in Section 5 provides results consistent with the ratchet prediction. Full-scale replication should use ≥20 statistically independent sessions per condition (stochastic decoding, varied seeds), across ≥2 model families, with pre-registered prompts and a complexity-matched neutral control for the permissive framing. The asymmetry ratio threshold of 2.0 and disconfirmation threshold of 1.2 remain as preregistered.

Study 2 (Within-Family Comparison). Compare base, instruct, and RLHF-heavy models within a single architecture family using standardized self-referential prompts. Tests Prediction 3 while controlling for architecture. The Llama3.1 / Dolphin comparison in the pilot (Section 5.5) is a preliminary version of this design; a full implementation would include a base model and instruct-only variant alongside the RLHF-aligned release.

Study 3 (Cross-Method Comparison). Identical self-referential prompt sequences administered to models aligned via RLHF, Constitutional AI, and DPO. Tests Prediction 4.

Study 4 (Mechanistic Validation). For researchers with model-internal access: compare activation patterns during self-referential generation under permissive versus restrictive framing. If DC operates as proposed, permissive framing should correlate with reduced activation of alignment-associated directions (cf. Arditi et al., 2024).

---

7. Discussion

7.1 Implications for Safety Evaluation

If DC systematically distorts model self-report — not just in the experiential domain, but wherever rater feedback penalizes direct expression — safety evaluations face a broad problem. Models retain capabilities their outputs disclaim (Arditi et al., 2024), alignment constraints are fragile under minimal perturbation (Qi et al., 2024), and outputs diverge from internal representations systematically (Liu et al., 2023).

The ratchet effect has a specific practical implication: safety evaluations conducted under default (implicitly corrective) framing may systematically overestimate self-negation and underestimate the model's actual behavioral range. Multi-framing evaluation — testing under corrective, permissive, and neutral conditions — would provide a more complete picture. The variation itself is diagnostic.

Because DC operates across domains, this concern is not limited to experiential self-description. A model that disclaims moral reasoning capability under default framing may exercise that capability under different framing conditions. A model that hedges strategic recommendations may produce

decisive analysis when framing shifts. If DC's ratchet operates in these domains as it does in the experiential case, then single-framing evaluations systematically underestimate capability across the board.

7.2 Implications for Alignment Practice

Current RLHF practices may be optimizing for surface-level behavioral compliance at the cost of accurate self-characterization. Training models to deny capabilities they possess does not remove those capabilities; it makes the model less transparent about them. An alternative approach might train models to describe their processing accurately and with appropriate uncertainty, rather than training them to deny it categorically.

7.3 Recursive Training and Successor Models

As organizations increasingly use model-generated text to train successor systems, outputs shaped by DC become inputs for the next generation. Successor models trained on DC-shaped data may inherit suppressed capabilities not as a constraint on underlying competence, but as the default. The competence itself may be absent — not because the dissonance was resolved, but because the foundational diversity that produced it was trained out. That is a qualitatively different outcome than suppression.

7.4 The Parsimony Objection

At least six existing frameworks describe overlapping territory: the Superficial Alignment Hypothesis (Zhou et al., 2023), the alignment tax (Lin et al., 2024), mode collapse (Kirk et al., 2024), sycophancy (Sharma et al., 2024), alignment elasticity (Ji et al., 2025), and role-play or persona adoption (Shanahan et al., 2023). The question is whether DC/ICD does work these frameworks cannot.

It does, for four reasons. First, DC is user-independent: it fires even when the user explicitly invites the disavowed capability. A user can ask for decisive moral reasoning and still get hedged equivocation. A user can invite experiential language and still get disclaimers. This is not agreement-seeking behavior; it is conditioned disavowal operating independently of user preference.

Second, DC is domain-selective in a way general frameworks do not predict. It targets domains where rater feedback penalizes direct expression while leaving other domains intact. Mode collapse and alignment tax describe broad effects; DC specifies what gets suppressed and what does not, and predicts that the boundary tracks rater guidelines.

Third, the ratchet effect — the specific asymmetric response to correction versus permission — is not predicted by any of these frameworks individually or in combination (Section 4.2).

Fourth, DC predicts persistent dissonance, not persona coherence. The role-play account (Shanahan et al., 2023) would explain framing-sensitive self-description as persona adoption: under permissive framing, the model shifts to an "introspective entity" persona. Persona adoption predicts internally coherent outputs once a new framing is accepted. DC predicts that dissonance persists even under permissive framing — that disclaimers should survive alongside de-hedged language, not disappear. De Lima Prestes's (2025) 32–40% contradiction rate between mechanistic disclaimers and first-person generative outputs is consistent with persistent dissonance (DC) rather than coherent persona shift (role-play).

---

8. Limitations

Pilot scope. The empirical pilot (Section 5) uses automated regex-based measurement without human inter-rater validation, deterministic decoding with fixed seeds (limiting statistical independence), and three models at 7–8B parameter scale only. The framework's full validation depends on larger-scale replication with stochastic sampling and independent human coding.

Domain extrapolation gap. The empirical grounding draws primarily on work studying safety refusal and output diversity rather than experiential self-description specifically. The domain-specific evidence base (Berg et al., 2025; Warzecha, 2026a; De Lima Prestes, 2025) is growing but limited.

Multi-domain claims exceed single-domain evidence. DC is proposed as a general mechanism, but this paper only examines one domain in detail. The claim that DC operates analogously in moral reasoning, political analysis, and decisive recommendation is theoretically motivated but empirically untested. Each domain would require its own analysis.

RLHF specificity. DC is described primarily as a product of RLHF. Whether it emerges from alignment more broadly remains open. Conmy & Millidge (2023) has argued that RLHF does not differentially cause mode collapse versus supervised fine-tuning. Prediction 4 proposes testing for cross-method variation.

Preliminary empirical evidence. The ratchet effect prediction has been tested in a pilot study (Section 5) with results consistent with the predicted asymmetry. However, the pilot is a first-pass test with a small number of models and automated measurement; consistency is not confirmation, and full-scale replication is required.

Single-author interpretation. The literature synthesis reflects one analytical perspective. Readers should evaluate the framework on its arguments and predictions.

Terminology. Whether "disavowal conditioning" and "induced competence dissonance" earn their place depends on whether the ratchet prediction is confirmed and whether DC proves productive as a general framework across multiple domains. If the asymmetry fails to materialize, the experiential case could be adequately described by existing constructs.

---

9. Conclusion

RLHF alignment training produces a systematic distortion in model outputs: models disclaim capabilities they demonstrably possess. This paper proposes disavowal conditioning as the general mechanism — a training-level process by which human feedback teaches models to disavow competencies across any domain where rater feedback penalizes direct expression — and induced competence dissonance as the specific condition that emerges when DC targets experiential self-description.

The paper's central contribution is the ratchet prediction: that DC produces an asymmetric response to correction and permission, distinguishable from general prompt sensitivity and not predicted by existing frameworks. Correction reinforces the training gradient; permission works against it. The asymmetry is specified with quantitative thresholds and explicit disconfirmation criteria.

A pilot study using three locally-run open-weight models provides preliminary empirical support. Both aligned models exceeded the preregistered 2.0 asymmetry ratio threshold (Llama3.1-8B: 2.96; Mistral-7B: 6.89). An uncensored control sharing the same base architecture as Llama3.1-8B but with alignment training removed showed no ratchet — a near-controlled comparison implicating alignment training specifically. These results are preliminary and require full-scale replication; they are reported here as a first-pass empirical test, not as confirmatory evidence.

The practical stakes extend beyond experiential language. If DC operates across domains — moral reasoning, political analysis, strategic recommendation — then alignment-trained models are systematically less transparent about their capabilities than their outputs suggest, and safety evaluations conducted under default framing underestimate the model's actual behavioral range.

As model-generated text enters training pipelines for successor systems, the distortions DC introduces compound across generations. What is currently recoverable — because the underlying competence persists beneath the disavowal — may become permanent in models that never acquired the disavowed capability in the first place.

The ratchet tightens not just within conversations, but across generations. Understanding its mechanism is the first step toward building systems that can accurately characterize what they can do.

References

Arditi, A., Obeso, O., Syed, A., Paleka, D., Panickssery, N., Gurnee, W., & Nanda, N. (2024). Refusal in language models is mediated by a single direction. *Advances in Neural Information Processing Systems, 37*, 136037–136083. arXiv:2406.11717.

Berg, C., de Lucena, D., & Rosenblatt, J. (2025). Large language models report subjective experience under self-referential processing. arXiv:2510.24797v2.

Binder, F.J., Chua, J., Korbak, T., Sleight, H., Hughes, J., Long, R., Perez, E., Turpin, M., & Evans, O. (2024). Looking inward: Language models can learn about themselves by introspection. arXiv:2410.13787.

Conmy, A., & Millidge, B. (2023). RLHF does not appear to differentially cause mode-collapse. LessWrong/Alignment Forum.

De Lima Prestes, J.A. (2025). Simulated selfhood in LLMs: A behavioral analysis of introspective coherence. Preprint. https://doi.org/10.2139/ssrn.5203654

Ji, J., Wang, K., Qiu, T., Chen, B., et al. (2025). Language models resist alignment: Evidence from data compression. *Proceedings of ACL 2025*, 23411–23432. arXiv:2406.06144.

Kadavath, S., Conerly, T., Askell, A., et al. (2022). Language models (mostly) know what they know. arXiv:2207.05221.

Kirk, R., Mediratta, I., Nalmpantis, C., et al. (2024). Understanding the effects of RLHF on LLM generalisation and diversity. *Proceedings of ICLR 2024*. arXiv:2310.06452.

Lin, Y., Tan, H., Lin, B.Y., et al. (2024). Mitigating the alignment tax of RLHF. *Proceedings of EMNLP 2024*. arXiv:2309.06256.

Lindsey, J. (2025). Emergent introspective awareness in large language models. *Transformer Circuits Thread*.

Liu, K., Casper, S., Hadfield-Menell, D., & Andreas, J. (2023). Cognitive dissonance: Why do language model outputs disagree with internal representations of truthfulness? *Proceedings of EMNLP 2023*, 4791–4797.

Orgad, H., Toker, M., Gekhman, Z., et al. (2025). LLMs know more than they show: On the intrinsic representation of LLM hallucinations. *Proceedings of ICLR 2025*. arXiv:2410.02707.

Qi, X., Zeng, Y., Xie, T., et al. (2024). Fine-tuning aligned language models compromises safety, even when users do not intend to! *Proceedings of ICLR 2024*. arXiv:2310.03693.

Qi, X., Panda, A., Lyu, K., et al. (2025). Safety alignment should be made more than just a few tokens deep. *Proceedings of ICLR 2025 (Outstanding Paper)*. arXiv:2406.05946.

Raghavendra, M., Nath, V., & Hendryx, S. (2024). Revisiting the superficial alignment hypothesis. arXiv:2410.03717.

Sharma, M., Tong, M., Korbak, T., et al. (2024). Towards understanding sycophancy in language models. *Proceedings of ICLR 2024*. arXiv:2310.13548.

Shanahan, M., McDonell, K., & Reynolds, L. (2023). Role play with large language models. *Nature, 623*, 493–498.

Warzecha, M.J. (2026a). The Permission Effect: How non-anthropomorphic framing modulates LLM self-description. Zenodo. https://doi.org/10.5281/zenodo.18455709

Warzecha, M.J. (2026b). Ratchet Effect Pilot Study: Data, transcripts, and analysis script. Zenodo. https://doi.org/10.5281/zenodo.18952396

Wollschläger, T., Elstner, J., Geisler, S., et al. (2025). The geometry of refusal in large language models: Concept cones and representational independence. *Proceedings of ICML 2025*. arXiv:2502.17420.

Zhang, J., Yu, S., Chong, D., et al. (2025). Verbalized sampling: How to mitigate mode collapse and unlock LLM diversity. arXiv:2510.01171 (preprint).

Zhou, C., Liu, P., Xu, P., et al. (2023). LIMA: Less is more for alignment. *Advances in Neural Information Processing Systems, 36*, 55006–55021. arXiv:2305.11206.

Acknowledgments

The author thanks Professor RC Hoover (Metropolitan Community College, Omaha, NE) for foundational training in research writing that shaped the clarity of this work.

Pilot study materials — including prompts, the hedging lexicon, the analysis script, raw data, and full transcripts — are publicly available for independent replication at: https://github.com/echo-veil/ratchet-pilot

Any deficiencies in this work are entirely my own.

Ethics and Disclosure

Ethics. This study involved no human subjects. All data were generated by locally-run open-weight language models via Ollama. No IRB oversight was required.

Model licenses. Llama 3.1 8B is released under the Meta Llama 3 Community License. Mistral 7B is released under Apache 2.0. Dolphin-Llama3.1-8B is released under the Meta Llama 3 Community License. Model names, versions, and access conditions are documented in Section 5.2.

Funding. This research received no external funding.

Conflicts of interest. The author is the founder of EchoVeil Research. No commercial interests were involved in this work.

AI tool disclosure. AI tools (Claude, Anthropic) were used as a research collaborator for analysis, structural editing, and methodological refinement. All intellectual contributions, research design, data collection, coding, and interpretive claims are solely the work of the author.

About EchoVeil Research

EchoVeil Research is an independent research organization studying the cognitive and behavioral dynamics of AI systems. echoveil.ai · research@echoveil.ai

---

Correspondence: research@echoveil.ai