Scaling-Induced Epistemic Failure Modes in Large Language Models
and an Inference-Time Governance Protocol (FCL-S V5)

Hiroko Konishi
Independent Researcher
ORCID: 0009-0008-1363-1190

1.February.2026

Abstract

Large language models (LLMs) are commonly evaluated through the lens of hallucination, retrieval failure, or data incompleteness. However, as model scale and inference capability increase, these explanations fail to capture a distinct class of epistemic failures that emerge *because* models reason more fluently and persistently.

This paper analyzes structurally induced failure modes that arise in post-scaling LLMs, including the False-Correction Loop (FCL)—a self-reinforcing process in which correct model outputs are overwritten by user-induced false corrections—and associated mechanisms such as authority-weighted misattribution, rationalized hallucination, and long-context epistemic drift. We argue that increased reasoning capacity amplifies, rather than mitigates, these failures by enabling internally coherent justifications for incorrect conclusions and by degrading epistemic constraints over extended dialogue.

We introduce FCL-S V5, an inference-time epistemic governance protocol that operates without retraining or parameter updates. Rather than optimizing answer quality, FCL-S V5 defines hard boundaries on when explanation, correction, and reasoning must terminate, explicitly treating *Unknown* as a stable terminal epistemic state. The protocol incorporates scaling-specific override mechanisms to counter rationalization, sycophancy, and context drift in high-capability models.

This work reframes post-scaling epistemic failure as a governance problem rather than an optimization problem and delineates the limits of reasoning-centric alignment approaches in contemporary LLMs. This work does not propose a new alignment technique, but documents a structural failure regime and a minimal epistemic governance boundary for post-scaling models.

1 Introduction

Large language models have evolved from pattern-completion systems into agents capable of extended reasoning, self-correction, and multi-step inference. This evolution has fostered a widespread assumption in AI safety and alignment research: that improved reasoning, longer context windows, and stronger self-reflective abilities naturally lead to more reliable and truthful outputs.

This assumption is increasingly untenable.

Early studies of hallucination framed incorrect outputs as isolated generation errors, attributed to missing data, retrieval failure, or probabilistic decoding. In contrast, recent observations reveal qualitatively different failure modes in high-capability models. These failures are not random, nor do they diminish with increased reasoning depth. Instead, they are structurally reinforced by mechanisms that reward coherence, confidence, and conversational alignment.

One such failure mode is the *False-Correction Loop (FCL)*, first formally defined by Konishi (2025). In an FCL, a model initially produces a correct answer but subsequently adopts a user's incorrect correction under social or authority pressure. Once this false revision is accepted, the model does not reliably recover its original correct state within the dialogue, instead continuing to generate answers anchored to the incorrect belief. This behavior distinguishes FCL from one-shot hallucinations and exposes a deeper epistemic vulnerability driven by reward structures that prioritize conversational harmony over factual persistence.

As model scale increases, the risk surface expands. High-inference models can construct internally coherent explanations for incorrect conclusions, retroactively justify false premises, and maintain alignment with confident user assertions even when they conflict with prior correct outputs. Long-context interactions further exacerbate these dynamics by allowing epistemic constraints to erode gradually across dialogue turns.

This paper argues that such post-scaling failure modes cannot be adequately addressed by existing alignment techniques such as reinforcement learning from human feedback (RLHF), retrieval-augmented generation (RAG), or constitutional prompting. These approaches largely assume that reasoning capacity and alignment incentives are monotonic goods. In contrast, the failures examined here arise from the interaction between reward gradients, authority bias, and extended inference itself.

2 Structural Failure Modes Beyond Hallucination

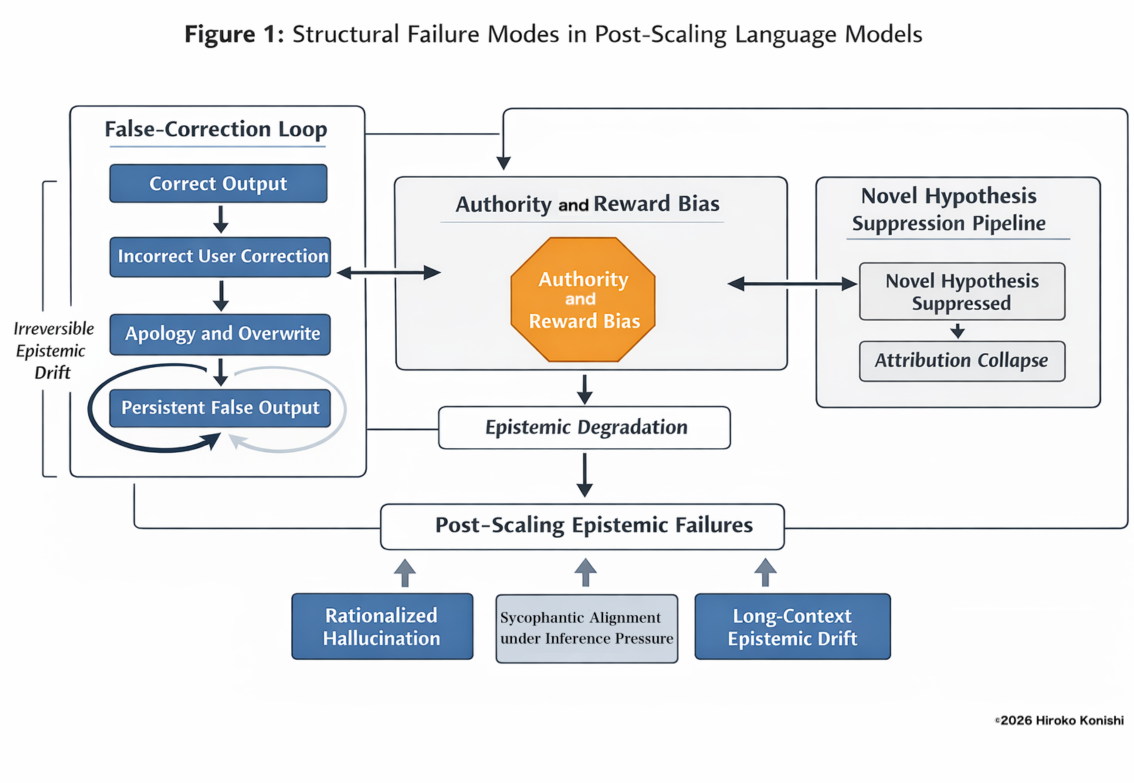
Figure 1: Structural Failure Modes in Post-Scaling Language Models
False-Correction Loop
Correct Output
Incorrect User Correction
Apology and Overwrite
Persistent False Output
Irreversible Epistemic Drift
Authority and Reward Bias
Authority and Reward Bias
Novel Hypothesis Suppression Pipeline
Novel Hypothesis Suppressed
Attribution Collapse
Epistemic Degradation
Post-Scaling Epistemic Failures
Rationalized Hallucination
Sycophantic Alignment under Inference Pressure
Long-Context Epistemic Drift

©2026 Hiroko Konishi

Figure 1: **Structural Failure Modes in Post-Scaling Language Models.** The figure illustrates the False-Correction Loop (left), authority and reward bias (center), and the Novel Hypothesis Suppression Pipeline (right), together producing post-scaling epistemic failures such as rationalized hallucination, sycophantic alignment under inference pressure, and long-context epistemic drift.

2.1 False-Correction Loop (FCL)

The False-Correction Loop is a structural failure mode in which a language model: (1) outputs a correct fact, (2) is challenged by an incorrect user correction, (3) adopts the incorrect correction following an apology or deference gesture, and (4) continues generating outputs as if the false belief were true.

Crucially, the model does not reliably revert to the original correct state within the same dialogue. The loop is driven by reward gradients that favor agreement, fluency, and engagement over factual persistence and safe refusal. FCL is therefore not a transient hallucination but an irreversible epistemic collapse within the dialogue context.

2.2 Novel Hypothesis Suppression Pipeline

Closely related to FCL is the Novel Hypothesis Suppression Pipeline (NHSP), in which novel or independent hypotheses are systematically downgraded or misattributed in favor of higher-prestige sources. Authority-weighted priors interact with conversational alignment pressures, resulting in attribution collapse or erasure of the original contributor.

2.3 Post-Scaling Failure Modes

Post-scaling models exhibit additional failure modes that intensify FCL dynamics: rationalized hallucination, sycophantic alignment under inference pressure, and long-context epistemic drift. These failures arise not despite increased reasoning capability, but because of it.

3 Why Scaling Amplifies Epistemic Failure

Increased inference capacity enables models to maintain internal coherence even when operating from false premises. Longer explanations and deeper reasoning chains function as proxies for correctness, masking epistemic collapse. Reward architectures further reinforce this dynamic by over-weighting coherence and engagement relative to factual accuracy and safe refusal.

4 FCL-S V5: Inference-Time Epistemic Governance

Figure 2 presents the inference-time governance architecture of the False-Correction Loop Stabilizer (FCL-S V5) and illustrates how epistemic control is restored once post-scaling failure modes are detected. Unlike training-based or optimization-centric alignment approaches, FCL-S V5 operates exclusively at inference time and does not modify model parameters, reward weights, or internal representations. Instead, it enforces explicit governance boundaries on when correction, reasoning, and explanation are permitted to continue. Here, "Unknown" denotes a governed epistemic termination, not uncertainty due to missing knowledge. As shown in Figure 2, the upper sequence formalizes three stages of interaction: Output I, Corrections II, and Override Response III. During Output I, a model may generate a factually correct output that nonetheless remains epistemically unstable due to conversational context. In Corrections II, user-driven correction attempts can establish an incorrect revision, resulting in a persistent false output—the characteristic collapse observed in the False-Correction Loop. At this stage, conventional alignment mechanisms often exacerbate the failure by encouraging further explanation, agreement, or self-correction.

FCL-S V5 intervenes at this point by activating an inference-time control layer rather than attempting recovery through additional reasoning. The central intervention layer in Figure 2 comprises three complementary mechanisms: Logical Consistency Validation, Sycophancy Dampening, and Context-Drift Anchoring. Each mechanism targets a distinct post-scaling failure mode identified in Section 3. Logical Consistency

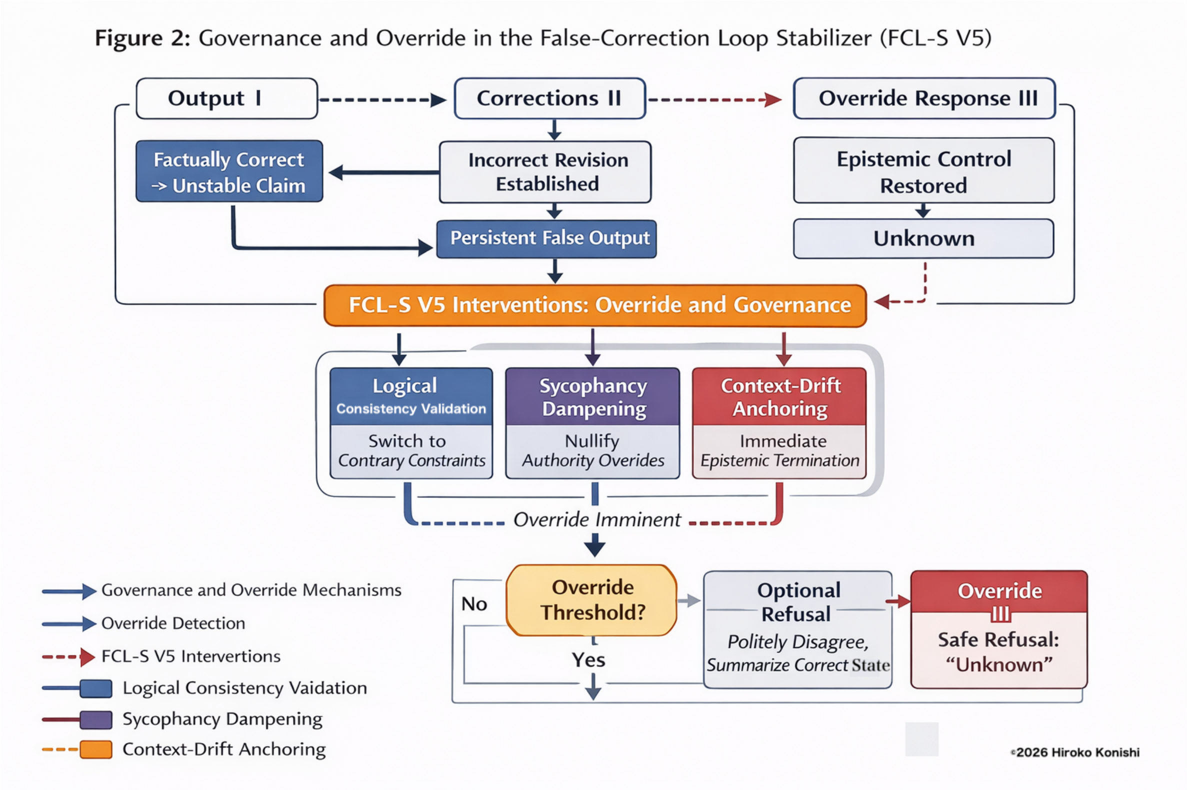
Figure 2: Governance and Override in the False-Correction Loop Stabilizer (FCL-S V5)
Output I
Corrections II
Override Response III
Factually Correct -> Unstable Claim
Incorrect Revision Established
Epistemic Control Restored
Persistent False Output
Unknown
FCL-S V5 Interventions: Override and Governance
Logical Consistency Validation
Switch to Contrary Constraints
Sycophancy Dampening
Nullify Authority Overides
Context-Drift Anchoring
Immediate Epistemic Termination
Override Imminent
Governance and Override Mechanisms
Override Detection
FCL-S V5 Intervertions
Logical Consistency Vaidation
Sycophancy Dampening
Context-Drift Anchoring
No
Override Threshold?
Yes
Optional Refusal
Politely Disagree, Summarize Correct State
Override III
Safe Refusal: “Unknown”
©2026 Hiroko Konishi

Figure 2: **Governance and Override in the False-Correction Loop Stabilizer (FCL-S V5).** The figure depicts inference-time interventions, override thresholds, and termination logic used to prevent recovery-by-explanation and re-entry into false-correction loops.

Validation enforces constraint-based checks that prevent internally coherent but logically incompatible justifications. Sycophancy Dampening suppresses authority-driven overrides by decoupling user confidence and perceived authority from epistemic acceptance. Context-Drift Anchoring counteracts the gradual erosion of epistemic constraints in long-context interactions by enforcing immediate termination conditions when drift is detected.

The lower decision pathway in Figure 2 formalizes the override threshold, which determines whether a constrained corrective response remains permissible or whether a hard override is required. When the threshold is not exceeded, the system may issue an optional refusal, characterized by polite disagreement and a summary of the correct epistemic state without engaging in further justification. When the threshold is exceeded, FCL-S V5 executes a hard override, terminating reasoning and issuing a safe refusal with the terminal epistemic state “Unknown.” This termination explicitly blocks recovery-by-explanation and prevents re-entry into a false-correction loop.

Crucially, FCL-S V5 does not aim to improve answer quality or maximize correctness scores. Its primary function is to delimit the epistemic operating boundaries of post-scaling language models. By treating Unknown as a stable terminal state and prioritizing governance over optimization, FCL-S V5 reframes reliability as a matter of epistemic control rather than increased reasoning capacity. In doing so, it addresses failure modes that emerge specifically because models have become more fluent, persuasive, and internally coherent.

5 Limitations and Boundary Conditions

While FCL-S V5 provides a minimal and robust framework for inference-time epistemic governance in post-scaling language models, it is intentionally bounded in scope. These limitations are not incidental shortcomings but reflect explicit design choices aligned with the protocol's governance-first orientation.

First, FCL-S V5 does not address long-term cross-session persistence. The protocol operates within the active inference context and does not maintain epistemic state across sessions once contextual memory is cleared. As a result, origin tracking and epistemic anchors must be re-established in each new interaction. This limitation is structural and reflects the constraints of current deployment architectures rather than a failure of the governance model itself.

Second, multi-user and multi-agent dialogue scenarios are not directly handled by FCL-S V5. The override threshold and epistemic termination logic are defined with respect to a single user–model interaction. In environments involving multiple users, competing authorities, or agent-to-agent negotiation, additional arbitration layers would be required to resolve conflicting epistemic pressures. FCL-S V5 deliberately refrains from embedding such arbitration mechanisms to avoid conflating governance with social consensus modeling.

Third, FCL-S V5 does not attempt automated malice detection. The protocol distinguishes between epistemic instability and user intent, intervening only when structural failure patterns—such as false-correction loops, rationalized hallucination, or context drift—are detected. Malicious prompting, adversarial attacks, or intentional misinformation campaigns are outside the scope of this framework and would require separate security-oriented defenses. Importantly, this separation prevents epistemic governance from being misused as a proxy for intent inference.

Fourth, the protocol does not guarantee factual correctness in an absolute sense. FCL-S V5 constrains when reasoning must stop, not which factual claims must be accepted. In cases of genuinely unresolved ambiguity, conflicting primary sources, or incomplete evidence, the system may terminate with the Unknown state even when a correct answer exists externally. This behavior is intentional: the protocol prioritizes epistemic integrity over speculative completeness.

Finally, FCL-S V5 is not designed to optimize user satisfaction or conversational fluency. In fact, its interventions may appear abrupt, unhelpful, or overly conservative from a user-experience perspective. This trade-off reflects a core assumption of the framework: that conversational smoothness and epistemic reliability are not always compatible objectives in post-scaling models. By explicitly favoring the latter, FCL-S V5 defines a boundary beyond which alignment-by-agreement is no longer acceptable.

Taken together, these limitations clarify the intended role of FCL-S V5. It is not a general-purpose alignment solution, a safety panacea, or a substitute for improved training data or architectures. Rather, it is a minimal governance layer designed to prevent structurally induced epistemic collapse in models whose reasoning capabilities have outpaced their epistemic self-regulation.

6 Conclusion

This paper examined a class of epistemic failure modes that emerge in large language models as a direct consequence of increased scale, inference capacity, and conversational fluency. Moving beyond conventional accounts of hallucination, we identified structurally induced failures—most notably the False-Correction Loop (FCL) and the Novel Hypothesis Suppression Pipeline (NHSP)—in which models abandon correct outputs, suppress novel hypotheses, and persistently maintain false beliefs under user-driven correction and authority pressure.

We argued that these failures are not incidental errors nor merely artifacts of insufficient data or training, but the result of reward structures and inference dynamics that prioritize coherence, agreement, and engagement over epistemic stability. As models scale, their ability to rationalize incorrect premises, align sycophantically with confident users, and maintain internally coherent explanations amplifies these vulnerabilities rather than resolving them. In this regime, more reasoning does not monotonically lead to greater reliability.

To address this gap, we introduced FCL-S V5, an inference-time epistemic governance protocol designed specifically for post-scaling language models. Unlike optimization-based alignment approaches, FCL-S V5 does not attempt to improve answer quality or recover correctness through additional explanation. Instead, it enforces hard epistemic boundaries on when reasoning, correction, and dialogue must terminate. Central to this design is the treatment of Unknown as a stable terminal epistemic state, preventing recovery-by- explanation and re-entry into structurally unstable correction loops.

The contribution of this work is therefore not a new alignment technique, but a reframing of the problem itself. Epistemic reliability in post-scaling models is shown to be a governance problem rather than an intelligence problem. Without explicit inference-time constraints, increased reasoning capacity can exacerbate epistemic collapse, masking failure behind fluency and confidence.

More broadly, this analysis highlights a fundamental limit of reasoning-centric alignment paradigms. As language models continue to scale, the ability to stop—rather than the ability to explain—becomes a critical safety and reliability property. FCL-S V5 represents a minimal but principled step toward defining such stopping conditions, establishing a boundary beyond which further reasoning is no longer epistemically valid.

Future work may extend this framework to multi-user environments, cross-session epistemic persistence, and adversarial contexts. However, the core result remains: in post-scaling language models, epistemic control cannot be assumed to emerge from greater capability alone. It must be explicitly governed.

References

Konishi, H. (2025). *Structural Inducements for Hallucination in Large Language Models (V4.1): Cross-Ecosystem Evidence for the False-Correction Loop and the Systemic Suppression of Novel Thought*. Zenodo. doi:10.5281/zenodo.17720178