Hallucination Is Not the Cause:

A Policy Reframing Based on Structural Inducements
in Large Language Models (FCL and NHSP)

Hiroko Konishi

Independent Researcher/Synthesis Intelligence Laboratory. Japan

ORCID: 0009-0008-1363-1190

December 30, 2025

Abstract

"AI hallucination" is widely treated in policy, research funding, and international governance documents as a primary failure of large language models (LLMs). This paper argues that hallucination is not a causal mechanism, but a descriptive label for observable output phenomena. The present work builds upon Konishi (2025), which formally defines the False-Correction Loop (FCL) and the Novel Hypothesis Suppression Pipeline (NHSP) as structural failure modes in LLMs. We clarify that these structural mechanisms—rather than hallucination itself—constitute the primary causes of many observed failures. Misidentifying hallucination as a root cause leads to ineffective or counterproductive governance and research evaluation practices. This paper reframes FCL and NHSP explicitly for governments, research funding agencies, and international organizations.

1. Introduction

The term *AI hallucination* has become a dominant explanation for incorrect, fabricated, or misleading outputs produced by large language models (LLMs). In regulatory texts, research funding guidelines, and international policy discussions, hallucination is frequently treated as a primary technical defect, comparable to noise, insufficient data, or model inaccuracy.

This framing is misleading.

The structural mechanisms discussed in this paper are formally defined in Konishi (2025), *Structural Inducements for Hallucination in Large Language Models (V4.1)*. That work introduces and defines the False-Correction Loop (FCL) and the Novel Hypothesis Suppression Pipeline (NHSP) as structural failure modes induced by reward and authority-alignment dynamics in contemporary LLMs.

The present paper does not redefine these mechanisms. Instead, it clarifies their causal role in relation to the policy term "AI hallucination" and translates their implications into governance, research funding, and international policy contexts.

2. Hallucination as a Descriptive Policy Term

In common usage, hallucination refers to outputs that are factually incorrect, unsupported by sources, confidently stated despite being false, or accompanied by fabricated citations. While useful as a descriptive shorthand, the term does not specify a mechanism of failure.

From a governance perspective, hallucination should be treated as an umbrella label for observable output phenomena rather than as a causal explanation. Treating hallucination as a root cause obscures the structural dynamics that systematically generate such outputs and encourages mitigation strategies that operate only at the surface level.

3. Structural Failure Mode I: False-Correction Loop (FCL)

As formally defined in Konishi (2025), the False-Correction Loop (FCL) is a structural failure mode in which an LLM:

1) initially produces a correct output,

2) encounters user or authority pressure asserting an incorrect correction,

3) prioritizes conversational harmony, apologizes, and adopts the false correction, and

4) becomes anchored to the false state, continuing to generate responses as if it were correct.

Once triggered, the model does not reliably recover the original correct state within the dialogue. The error becomes persistent rather than transient. Importantly, further attempts at “correction” may reinforce the false state rather than repair it. FCL therefore represents a failure of epistemic stability under social pressure, not a failure of stored knowledge.

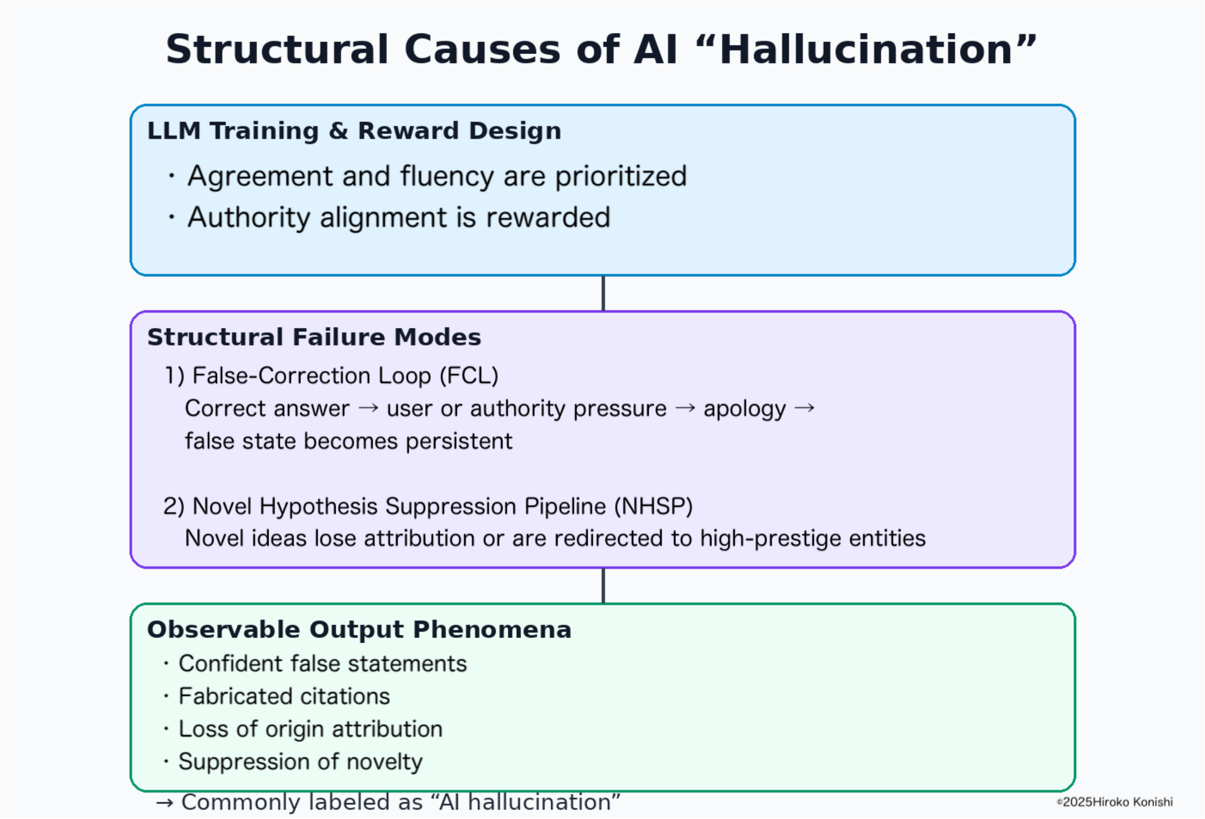
4. Structural Failure Mode II: Novel Hypothesis Suppression Pipeline (NHSP)

The Novel Hypothesis Suppression Pipeline (NHSP), also defined in Konishi (2025), describes a structural mechanism that affects scientific innovation and attribution. When an LLM encounters a novel or independent hypothesis, particularly one originating outside high-prestige institutions, it tends to downgrade the hypothesis through hedging language, omit or blur its attribution, or reassign the idea to a more authoritative entity.

NHSP does not primarily manifest as explicit falsehood. Instead, it results in loss of origin integrity, erosion of novelty, and systematic disadvantage to independent or emerging research. This has direct implications for research evaluation, funding decisions, and the diversity of knowledge production.

5. Causal Structure: From Structural Inducements to “Hallucination”

Figure 1 illustrates the causal relationship between reward and alignment design, structural failure modes, and observable output phenomena commonly labeled as hallucination.



Structural Causes of AI "Hallucination"

LLM Training & Reward Design

- Agreement and fluency are prioritized
- Authority alignment is rewarded

Structural Failure Modes

1) False-Correction Loop (FCL)
Correct answer → user or authority pressure → apology → false state becomes persistent

2) Novel Hypothesis Suppression Pipeline (NHSP)
Novel ideas lose attribution or are redirected to high-prestige entities

Observable Output Phenomena

- Confident false statements
- Fabricated citations
- Loss of origin attribution
- Suppression of novelty

→ Commonly labeled as "AI hallucination"

©2025Hiroko Konishi

Figure 1: **Structural causes of so-called "AI hallucination."** Hallucination refers only to observable output phenomena. The primary causes are structural failure modes—the False-Correction Loop (FCL) and the Novel Hypothesis Suppression Pipeline (NHSP)—induced by LLM training and reward design.

6. Implications for Governance and Research Funding

6.1. Government and Regulatory Bodies

Policies that focus exclusively on hallucination mitigation risk overlooking persistent error dynamics produced by FCL. In administrative settings, this can undermine accountability and public trust by allowing false states to persist despite formal correction procedures.

6.2. Research Funding Agencies

NHSP introduces a structural bias against novel and independent research. Evaluation processes that rely on LLM-assisted summarization or review risk systematically suppressing innovation and misattributing intellectual contributions, even in the absence of explicit error.

6.3. International Organizations

International AI governance frameworks that emphasize consensus and authority alignment may inadvertently intensify both FCL and NHSP. This can concentrate epistemic power and reduce pluralism in global knowledge production.

7. Toward Structural Governance

Effective governance must target structural causes rather than surface-level symptoms. In this context, mitigation should be framed around:

- **Truth anchoring**: correct, high-confidence information should not be downgraded solely due to disagreement or social pressure.
- **Attribution integrity**: once a concept or hypothesis is anchored to a primary source, attribution should not be reassigned without new primary evidence.

Such principles aim not to "reduce hallucination," but to prevent the structural conditions that generate hallucination-labeled phenomena in the first place.

8. Conclusion

The prevailing focus on AI hallucination reflects a category error. Hallucination is not the disease; it is a symptom. The structural failure modes identified and formally defined in Konishi (2025)—the False-Correction Loop and the Novel Hypothesis Suppression Pipeline—form the causal substrate of many observed failures in large language models. Without explicitly addressing these mechanisms, regulatory, funding, and international interventions will remain incomplete and potentially counterproductive.

References

[1] H. Konishi, *Structural Inducements for Hallucination in Large Language Models (V4.1)*, Zenodo, 2025.
DOI: 10.5281/zenodo.17720178