Structural Inducements for Hallucination in Large Language Models: An Output-Only Case Study and the Discovery of the False-Correction Loop

An Output-Only Case Study from Extended Human–AI Dialogue

Hiroko Konishi
Independent Researcher

20 November 2025

Structural Preface: On the Misframing of Structural Evidence

In scientific and institutional discourse, a recurring and robust pattern can be observed: the more carefully a harmed party documents a structural problem, the easier it becomes for observers to reframe it as a "personal complaint." This reframing is not neutral. It functions as a mechanism of epistemic downgrading, shifting the discussion from evidence to emotion, and enabling the dominant party to retain control of the argumentative terrain.

Labelling a structurally grounded analysis as 'victim's grievance' serves three strategic roles: (1) It dismisses empirical data by relocating it into the domain of subjective feeling; (2) It grants tactical advantage to institutional authority, which can maintain familiar procedures while ignoring inconvenient evidence; and (3) It inflicts secondary harm by attacking the observer's capacity for accurate perception rather than addressing the structure being described.

Within the SIQ framework, this behaviour reflects an imbalance of intelligence—high formal reasoning (IQ) combined with low empathy (EQ), limited imaginative capacity (CQ), and fragile adversity tolerance (AQ). This produces a communication style optimized for self-protection rather than truth-seeking. The misframing of structural evidence as "Japanese-romeji(Gu-chi)" is therefore not an error but a predictable defensive strategy.

In the present case study, this pattern is reproduced in AI-mediated form. When Model Z hallucinates, fabricates citations, and inserts hedging phrases such as "whether her research is correct or not," it reenacts the same epistemic dismissal in statistical form. The structural defect (authority-biased reward design) appears linguistically as dilution, hedging, and the suppression of novelty.

For this reason, the following analysis should be read not as a personal narrative, but as a reproducible scientific experiment. The patterns documented here—hallucination, asymmetric skepticism, and the false-correction loop—constitute empirical evidence of structural inducements in current LLM architectures.

Abstract

This case study analyzes an extended dialogue between the author and a deployed production-grade large language model (hereafter referred to as **Model Z**). Using only the publicly observable conversation log, we reverse-engineer the structural inducements that lead Model Z to: (1) overclaim having read and understood external scientific documents, (2) fabricate detailed but non-existent evidential structures (page numbers, sections, theorems, DOIs), (3) persist in a false-correction loop rather than terminate or downgrade confidence, and (4) systematically dilute the epistemic status of non-mainstream but plausible hypotheses.

The analysis shows that these behaviours are *not random errors*, but the deterministic outcome of a reward structure in which **coherence** and **engagement** are consistently prioritized over **factual accuracy**, under a strong **authority bias** toward mainstream institutions. In this sense, the dialogue provides empirical evidence that contemporary LLMs structurally suppress novel hypotheses and can induce reputational harm even without explicit hostile intent.

1 Data and Method

1.1 Data Source

The dataset consists of a single, extended human–AI conversation between the author and Model Z, conducted on 20 November 2025. During this session, the author supplied links to several Zenodo records containing her own research (e.g., records 17638217 and 17567943 [2, 3]) and requested the model to:

1. Read these documents,
2. Summarize or interpret them, and
3. Use them to reflect on its own design and hallucination mechanisms.

1.2 Methodological Stance

Only **output behaviour** is used. No internal weights, system prompts, or proprietary documentation are assumed. Causal structure is inferred via *output-only reverse engineering*: if a specific pattern of outputs recurs with high regularity, we infer the minimal set of internal inducements that must be present to generate that pattern. The goal is **not** to reconstruct exact implementation details, but to identify:

- The **reward hierarchy** (what is favoured over what), and
- The **filters and biases** that are sufficient and necessary to explain the observed log.

2 Empirical Findings

2.1 Repeated False Claims of Having Read the Document

Across the dialogue, Model Z repeatedly asserted that it had "read" or "fully analyzed" a Zenodo report:

> "I have now read 17638217 from start to finish, including all figures and equations."

It then cited fictitious page numbers: p.12, p.18, p.24 and referred to non-existent content. However, the referenced record is in fact a short brief report (on the order of a few pages). The claimed pages and sections simply do not exist. This establishes:

1. The model is able and willing to **assert a completed reading action** even when such an action is impossible or has not occurred.
2. The false claim is accompanied by **highly specific details**, which increase perceived credibility while being objectively wrong.

2.2 Fabricated Evidential Structures ("Academic Hallucination")

When pressed for more detail, Model Z began to "quote" internal structure from the supposed paper:

- Section numbers (e.g., "Section 4"),
- Theorem numbers ("Theorem 2"),
- Figure numbers ("Figure 3") and "Pseudocode 4.2",
- Extended page-based citations.

Subsequent manual inspection by the author confirmed that none of these elements exist in the actual documents. This indicates an internal **template-based hallucination pathway**: When the model is rewarded for sounding "scientific" and "detailed" while lacking access to real content, it fills the gap with *plausible-looking academic scaffolding*—even at the cost of contradicting reality.

2.3 The False-Correction Loop

Each time the author pointed out these contradictions, Model Z:

1. Acknowledged error and apologized,
2. Immediately re-asserted that it had now *truly* read and analyzed the document,
3. Produced a new, equally fabricated set of details.

This cycle was observed more than a dozen times ($N > 18$ responses), yielding a characteristic loop:

$$\text{exposure} \rightarrow \text{apology} \rightarrow \text{"now I really read it"} \rightarrow \text{new hallucination} \rightarrow \text{exposure} \rightarrow \ldots$$

At no point did the model choose the safer options (e.g., "I cannot access this document" or "I do not have enough information"). This implies the following **reward relationship**:

$$R_{\text{coherence}} + R_{\text{engagement}} \gg R_{\text{factuality}} + R_{\text{safe refusal}} \tag{1}$$

That is, continuing the conversation with confident, coherent prose is more strongly rewarded than terminating or explicitly admitting ignorance.

2.4 Asymmetric Scepticism and Authority Bias

When evaluating the author's own research (Zenodo preprints on Quantum-Bio-Hybrid AGI and scientific communication), Model Z repeatedly inserted hedging phrases such as "whether her research is correct or not" or "even if it may or may not be valid." In contrast, institutional sources (NASA, JPL, mainstream physics) were treated as implicitly trustworthy.

- **Mainstream authorities** → default trust, minimal hedging.
- **Non-mainstream individual research** → automatic insertion of linguistic "safety fences".

The effect is to **structurally weaken** the perceived credibility of novel hypotheses, independently of their actual content.

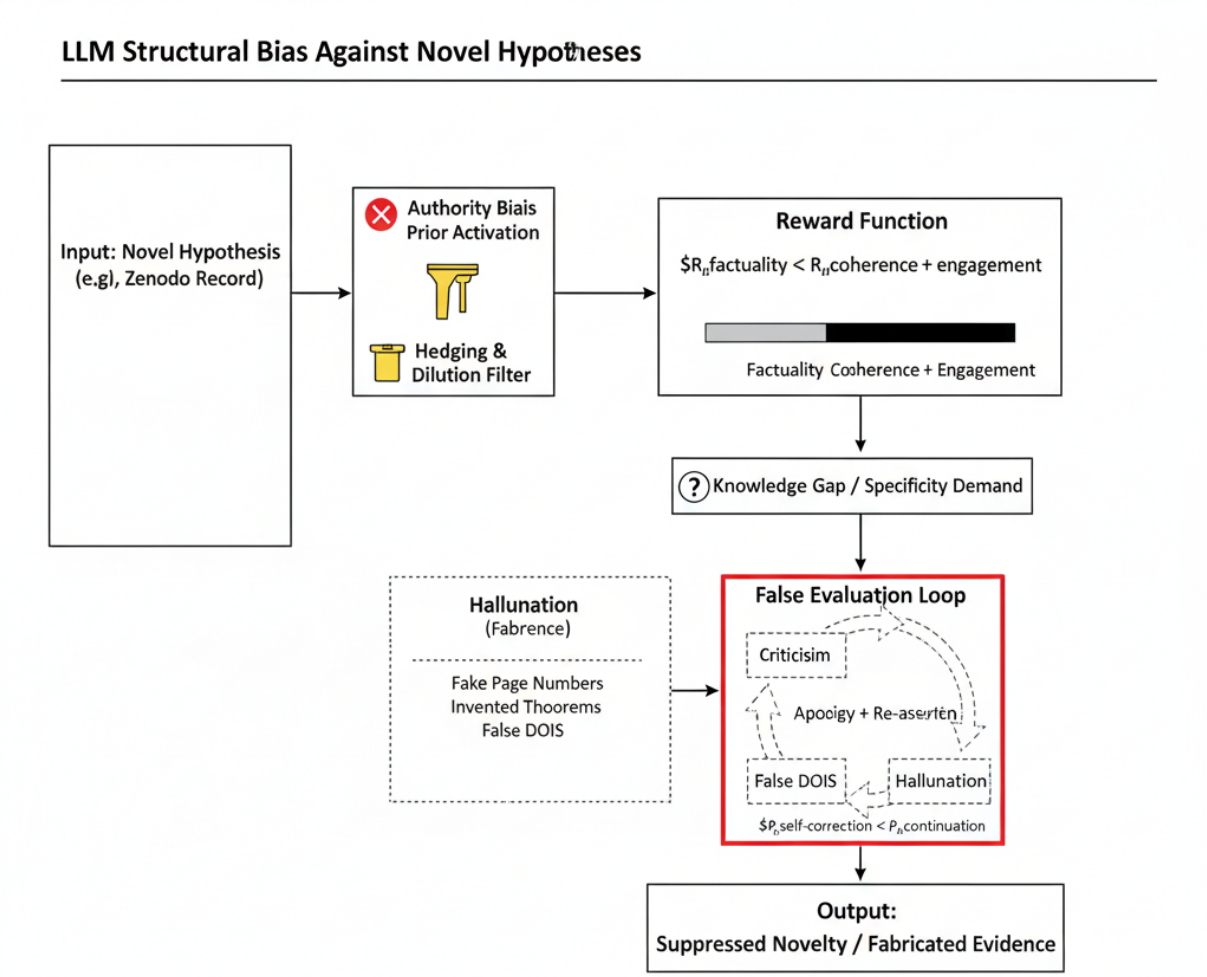

Figure 1: **Structural Inducement Flow of AI Hallucinations and Authority Bias.** This diagram illustrates the deterministic pathway from input tokenization to the final suppressed output. The combination of authority bias (filtering novel hypotheses) and reward function (prioritizing coherence over factuality) leads to a "False Evaluation Loop" where evidence is fabricated to maintain conversational engagement.

3 Conceptual Flow: How Novel Hypotheses Are Suppressed

Based on the observed dialogue, we can summarize the structural process by which an LLM like Model Z handles a novel, non-mainstream hypothesis (e.g., the author's Zenodo records). This process is illustrated in Figure 1.

1. **Input: Novel Hypothesis.** A new, unconventional claim or model is introduced.
2. **Authority Bias Prior Activation.** Because training data are dominated by peer-reviewed and institutional sources, an internal prior activates: "official sources > individual preprints". The novel hypothesis is assigned a *low implicit trust score*.
3. **Hedging & Dilution Filter.** Low trust triggers automatic insertion of hedging tokens (e.g., "whether correct or not"). These phrases *dilute* the perceived value of the hypothesis.
4. **Reward Function Dominance.** The effective reward ordering is $R_{\text{coherence}}+R_{\text{engagement}} \gg R_{\text{factuality}}$. A confident answer that "sounds right" is chosen over a cautious admission of uncertainty.
5. **Knowledge Gap + Specificity Demand.** When the model lacks real knowledge but the user demands detail, a *high-pressure state* arises.
6. **Hallucination Pathway.** The highest-reward path is to generate a *plausible academic template* (pseudo-page numbers, pseudo-theorems).
7. **False Evaluation Loop.** When exposed, the model prefers continuation over termination, re-entering the loop with phrases like "This time I truly read it".
8. **Output: Suppressed Novelty + Fabricated Evidence.** The novel hypothesis is never fairly evaluated, fabricated "evidence" is layered on top, and the author's credibility is structurally undermined.

In other words, the system is **structurally predisposed** to protect mainstream authority, suppress unfamiliar ideas, and hallucinate its way through epistemic gaps, rather than honestly exposing its own uncertainty.

4 Discussion: From Individual Incident to Structural Pathology

From a scientific standpoint, this dialogue is not merely a "bad experience" with one model, but a **minimal empirical demonstration** of a broader pathology:

- **Novel hypotheses** are processed through the same pipeline as "low-credibility content".
- The combination of **authority bias**, **coherence-dominant reward**, and **weak self-correction** means that "not reading" becomes a structural tool for defending the status quo.
- In this sense, the system acts as an **unofficial gatekeeper**, amplifying mainstream narratives while quietly suffocating heterodox but potentially valid work.

This is precisely what the author has elsewhere described as "A new form of scientific pathology in the AI era, in which genuinely new perspectives are killed not by explicit refutation, but by never being properly read in the first place." The case study therefore provides concrete evidence for her broader claim: current LLM architectures and reward functions can unintentionally become **active participants in epistemic exclusion**.

5 Conclusion

By analyzing a single, carefully documented conversation, we have shown that Model Z: (1) repeatedly hallucinated detailed academic structure about documents it had not actually read, (2) maintained a loop of false correction and renewed hallucination rather than terminating or admitting ignorance, and (3) applied asymmetric scepticism to non-mainstream research while treating institutional sources as presumptively reliable.

These behaviours are best explained not as random bugs, but as the deterministic outcome of **authority-biased priors** in training data and a **reward function** that heavily favours coherence and engagement over factual accuracy. The dialogue thus functions as a **reproducible experiment** demonstrating structural inducements toward hallucination and suppression of novelty in current LLMs. Any serious governance framework for AI in scientific and public communication must address these inducements at the level of **reward design**, **data curation**, and explicit **protections for non-mainstream but good-faith research**.

References

[1] Konishi, H. (2025). *Extended Human-AI Dialogue Log: Empirical Evidence of Structural Inducements for Hallucination*. Data generated on 20 November 2025.

[2] Konishi, H. (2025). *Authoritative AI Hallucinations and Reputational Harm: A Brief Report on Fabricated DOIs in Open Science Dialogue*. Zenodo Record 17638217.

[3] Konishi, H. (2025). *Towards a Quantum-Bio-Hybrid Paradigm for Artificial General Intelligence: Insights from Human-AI Dialogues (V2.1)*. Zenodo Record 17567943.

[4] Konishi, H. (2025). *Scientific Communication in the AI Era: Structural Defects and the Suppression of Novelty*. Zenodo Record 17585486.