The Meta-Mathematical Root of AI Hallucination

— A Diagnosis Based on the Zhu–Liang Holism Theorems

Jianbing Zhu[1]

[1] ECT-OS-JiuHuaShan Civilization Laboratory

ORCID: 0009-0006-8591-1891

DOI: 10.5281/zenodo.19622516

Email: ect-os-jiuhuashan@zohomail.cn

Preprint submitted: April 17, 2026

Abstract

The "hallucination" problem of current large language models—generating content that contradicts facts and ruptures logic—is widely regarded as a target for engineering optimization. Based on the core theorems of the Zhu–Liang Holism Axiomatic System (Whole–Part Correspondence Theorem, Truth Function Theorem, Terminal Coalgebra Theorem), this paper provides, for the first time, a rigorous meta-mathematical diagnosis of AI hallucination: **AI hallucination is the recurrence of Russell's paradox in the domain of cognitive engineering; both stem from the absence of the compatibility constitution** $f_Q|_P = f_P$. We prove that current AI architectures (token-level autoregression, embedding space vectorization, "emergence" narrative) fully reproduce the threefold murder perpetrated by reductionist generalization upon the foundation of set theory—dissolving the compatibility condition, reversing logical priority, and forgetting the functoriality of the whole function. Every hallucinatory output of an AI is the engineering equivalent of the inevitable "illegitimate set" $R = \{x \mid x \notin x\}$ produced in the absence of compatibility constraints. This paper further provides a therapeutic path: only by explicitly constructing a global semantic function F and constraining every local generation with the compatibility condition can AI ascend from a formal symbol game to a legitimate substructure of the Truth Space Ω. Conclusion: AI hallucination is not an engineering limitation, but the inevitable manifestation of an erroneous meta-mathematical foundation.

Keywords: AI Hallucination; Russell's Paradox; Whole–Part Correspondence Theorem; Compatibility Condition; Reductionist Generalization; Truth Space; Meta-Mathematical Diagnosis

Contents

1 Introduction: Hallucination as a Meta-Mathematical Symptom **3**

2 Preliminaries: Core Theorems of Holism **3**
2.1 Whole–Part Correspondence Theorem 3
2.2 Truth Function Theorem and Terminal Coalgebra 4
2.3 The Threefold Murder of Reductionist Generalization 4

3 Reproduction of the Threefold Murder in AI Architectures **4**
3.1 First Reproduction: Dissolving the Compatibility Condition 5
3.2 Second Reproduction: Reversing Logical Priority 5
3.3 Third Reproduction: Forgetting the Functoriality of the Whole Function 5

4 Rigorous Isomorphism between Russell's Paradox and AI Hallucination **6**
4.1 Isomorphism Mapping 6
4.2 Hallucination as the Inevitability of an "Illegitimate Set" 6

5 The Absence of the Truth Space Ω: AI as a Rootless Game **6**

6 The Meta-Mathematical Path to Therapy: Returning to the Compatibility Constitution **7**
6.1 Constitutional Principles and Engineering Requirements 7
6.2 Concrete Implementation Path 7

7 Conclusion **8**

1 Introduction: Hallucination as a Meta-Mathematical Symptom

The "hallucination" of large language models (LLMs) has become a core obstacle to AI reliability. Models generate content that is factually inconsistent and logically contradictory with high confidence. Academia and industry generally attribute this to noise in training data, randomness in inference sampling, or limitations in context windows. Yet the ever-worsening hallucination problem has never been fundamentally resolved through data cleaning, alignment tuning, or larger parameter scales. This suggests that hallucination may not be an engineering "defect," but the **inevitable manifestation of "illegitimate operations" at the meta-mathematical level** in cognitive engineering.

The establishment of the Zhu–Liang Holism Axiomatic System provides rigorous meta-mathematical tools for this diagnosis. The Whole–Part Correspondence Theorem (Theorem 2.1) proves that any legitimate whole must satisfy the compatibility condition $f_Q|_P = f_P$; reductionist generalization, by dissolving this condition, reversing logical priority, and forgetting functoriality, murdered the meta-mathematical foundation of set theory [4]. This paper aims to prove: **Current AI architectures are precisely a complete reproduction of this murder in the domain of cognitive engineering.** The hallucinatory outputs of AI are equivalent to the construction of the illegitimate set $R = \{x \mid x \notin x\}$ in Russell's paradox—both are self-contradictory due to the absence of the compatibility constitution.

The structure of this paper is as follows: Section 2 reviews the core theorems of Holism; Section 3 maps the threefold murder of reductionist generalization onto AI architectures; Section 4 rigorously proves the isomorphism between Russell's paradox and AI hallucination; Section 5 reveals that AI systems, due to the absence of the Truth Space Ω, cannot distinguish legitimate constructions from paradoxes; Section 6 provides a therapeutic path based on the compatibility constitution; Section 7 concludes.

2 Preliminaries: Core Theorems of Holism

This paper is strictly based on the Zhu–Liang Holism Axiomatic System. The following lists the directly relevant theorems; for detailed proofs, see the references [1, 4].

2.1 Whole–Part Correspondence Theorem

Theorem 2.1 (Whole–Part Correspondence Theorem)**.** *Let $F : D \to C$ be a whole function, and let its subfunctions be its restrictions $F|_P$ ($P \subseteq D$). The mapping*

$$\Phi : \{F\} \to \prod_{P \subseteq D} \{f : P \to C\}, \quad \Phi(F) = (F|_P)_{P \subseteq D}$$

is bijective on families satisfying the compatibility condition $f_Q|_P = f_P$ (for all $P \subseteq Q$). [1, Theorem 0.4.1]

This theorem establishes: (1) The whole is logically prior to the parts; (2) The compatibility condition is the mandatory constitution for subfunctions to be legitimately integrated into the whole.

2.2 Truth Function Theorem and Terminal Coalgebra

Theorem 2.2 (Truth Function Theorem)**.** *Truth $T : \Sigma \to R$ is the surjective function of all deterministic relations in the universe. [1, Theorem 0.3.1]*

Theorem 2.3 (Terminal Coalgebra Existence Theorem)**.** *The terminal G-coalgebra $\Omega = \varprojlim G^n(1)$ exists; its elements are recursive elements $x = (x_0, x_1, \dots)$ satisfying $p_n(x_{n+1}) = x_n$. [2, Theorem 2.3]*

The Truth Space Ω is the ultimate substrate of all legitimate whole functions. Any output of an intelligent system that cannot serve as a substructure of Ω is an illegitimate construction.

2.3 The Threefold Murder of Reductionist Generalization

Reductionist generalization perpetrated the following operations upon the foundation of set theory [4, Section 4]:

(1) **Dissolving the compatibility condition**: Distorting the whole into the mechanical sum of parts, ignoring $f_Q|_P = f_P$.

(2) **Reversing logical priority**: Declaring that parts (elements) are prior to the whole (set).

(3) **Forgetting the functoriality of the whole function**: Reducing morphisms to sets of ordered pairs, losing the structure of deterministic relations.

Russell's paradox is the inevitable product of this threefold murder—the subfunction family of the illegitimate set $R = \{x \mid x \notin x\}$ yields contradictory assignments on overlapping domains, violating the compatibility condition, and thus cannot be incorporated into Ω.

3 Reproduction of the Threefold Murder in AI Architectures

Current mainstream AI architectures (represented by Transformer autoregressive language models) fully reproduce the threefold murder of reductionist generalization.

3.1 First Reproduction: Dissolving the Compatibility Condition

Table 1: Dissolution of the Compatibility Condition by AI Architectures

Holism Constitution	**Violation by AI Architecture**
Compatibility condition $f_Q\|_P = f_P$: Any subfunctions must agree on overlapping domains	Token-level autoregression: generates token by token; no mandatory constraint on semantic consistency beyond the attention window
Global compatibility of the whole function	Hallucinatory outputs: descriptions of the same entity in different contexts contradict each other
Bijective correspondence of the subfunction family	No explicit construction of a global semantic function F; only local sampling from corpus statistics

The generation process of current LLMs is: given preceding text $x_{<t}$, maximize $P(x_t \mid x_{<t})$. This operation **defines subfunctions only on local domains**, with no compatibility check across different context windows. When a model gives contradictory answers to the same fact in different sessions, it is precisely the case where subfunctions f_{P_1} and f_{P_2} yield contradictory assignments on the overlapping domain $P_1 \cap P_2$—a direct symptom of the dissolution of the compatibility condition.

3.2 Second Reproduction: Reversing Logical Priority

The "emergent ability" narrative proclaims that as parameter scale increases, LLMs "spontaneously produce" higher-level intelligence such as reasoning and planning. In the context of Holism, this is equivalent to declaring that **the whole function F can "emerge" from subfunctional fragments**. Theorem 2.1 has rigorously proven that the definition of subfunctions logically depends on the prior existence of the whole function. Attempting to piece together a global semantic function from isolated token predictions is like trying to construct a set from singleton subfunctions—without the compatibility constraint, no legitimate whole function corresponds to it.

"Emergence" is not a miracle, but a misleading relabeling of statistical correlation.

3.3 Third Reproduction: Forgetting the Functoriality of the Whole Function

Embedding spaces map words and sentences to high-dimensional vectors; semantic relations are reduced to vector dot products or cosine similarity. This operation **forfeits the deterministic functorial structure of the truth function** $T : \Sigma \to R$. In Holism, morphisms (reasoning, causal relations) are compositions of the whole function across different domains, not distance metrics in a vector space. When an AI encodes "Paris is the capital of France" and "The capital of France is Paris" as two commutative vector relations, it has not truly grasped the determinacy of the functor "capital"—it has merely captured co-occurrence statistics in the corpus.

Consequence: The AI system cannot distinguish between **causal necessity** (function composition) and **statistical correlation** (vector similarity). This is the deep root of "hallucination."

4 Rigorous Isomorphism between Russell's Paradox and AI Hallucination

4.1 Isomorphism Mapping

Table 2: Isomorphism between Russell's Paradox and AI Hallucination

Russell Set $R = \{x \mid x \notin x\}$	**AI Hallucinatory Output**
Subfunction $F_R\|_{\{R\}}$ must simultaneously satisfy $R \in R$ and $R \notin R$	Same query outputs contradictory facts under different prompts or in different sessions
Contradictory assignment on overlapping domain (intersection of R as element and R as set)	Semantic rupture and logical contradiction at context window overlaps
Violation of compatibility condition $f_Q\|_P = f_P$	Local token consistency fails to guarantee global semantic consistency
Not a legitimate set; excluded from Truth Space Ω	**Not truth; an illegitimate construction of formal symbols**

4.2 Hallucination as the Inevitability of an "Illegitimate Set"

Suppose an AI system generates assertion A at time t_1, and assertion $\neg A$ at time t_2. Let P_1 be the contextual domain at t_1, and P_2 the contextual domain at t_2; the two yield contradictory assignments on the semantic overlap $P_1 \cap P_2$ concerning "the fact described by A." Since the system lacks the mandatory constraint of the compatibility constitution, subfunctions f_{P_1} and f_{P_2} disagree on the overlapping domain—$f_{P_1 \cup P_2}|_{P_1} = f_{P_1}$ and $f_{P_1 \cup P_2}|_{P_2} = f_{P_2}$ cannot hold simultaneously. By Theorem 2.1, no whole function F corresponds to such an incompatible family of subfunctions. Therefore, the hallucinatory output of AI is **not an "error," but fundamentally "does not exist" as a legitimate construction**—just as Russell's set R is not a "contradictory set," but is simply not permitted to be a set at all.

Corollary 4.1 (Illegitimacy of Hallucination)**.** *AI hallucination is not an engineering defect that "needs more data to correct," but the* ***inevitable result of illegitimate operations at the meta-mathematical level****. Generation without the constraint of the compatibility constitution will necessarily produce Russellian paradoxical outputs periodically.*

5 The Absence of the Truth Space Ω: AI as a Rootless Game

The Truth Space Ω (Theorem 2.3) is the ultimate substrate of all legitimate whole functions. Any output of an intelligent system that cannot serve as a substructure of Ω is an illegitimate construction. Current AI systems completely lack the concept and verification mechanism of Ω.

Table 3: Comparison of Truth Space Criteria and Current AI Status

Constitutional Requirement of Truth Space Ω	**Current Status of AI Systems**
Every legitimate output is a substructure of Ω	No concept of Ω; output space lacks legitimacy criterion
Compatibility condition is a mandatory constitution	Attention mechanism only provides statistical correlation; no strict compatibility constraint
Prior existence of the whole function F	Training process attempts to "emerge" F from corpus, violating logical priority
Morphisms preserve deterministic functorial structure	Embedding vectors lose functoriality, confounding causality with correlation

Consequence: The AI system cannot distinguish between **legitimate constructions** and **Russellian paradoxes**, leading to the simultaneous occurrence of factual errors (constructing illegitimate sets), logical contradictions (compatibility violations), and causal confusion (loss of functoriality). Hallucination is not an "occasional mistake," but the **normal state of a system operating without a constitution**.

6 The Meta-Mathematical Path to Therapy: Returning to the Compatibility Constitution

Based on the Holism Theorems, AI systems must undergo the following meta-mathematical reconstruction:

6.1 Constitutional Principles and Engineering Requirements

Table 4: Meta-Mathematical Constitution for AI Reconstruction

Constitutional Principle (Based on Holism Theorems)	**Requirement for AI Architecture Reconstruction**
Whole prior to parts (Theorem 2.1)	Explicit construction of the global semantic function F precedes local token generation
Compatibility prior to membership (Theorem 2.1)	Every output must pass the cross-context test of $f_Q\|_P = f_P$
Function prior to set (Theorem 2.2)	Deterministic relations (truth function) precede symbolic manipulation and vector similarity

6.2 Concrete Implementation Path

(1) **Explicitly construct the global semantic function** F: Use knowledge graphs, formal ontologies, or type-theoretic structures as the whole function, rather than relying on implicit "emergence" from corpus statistics.

(2) **Establish a compatibility verification layer**: Before generating output, mandatorily verify its compatibility with established facts and logical rules (engineering implementation of $f_Q|_P = f_P$).

(3) **Restore functoriality**: Define reasoning as the composition of morphisms of the whole function across domains, not as vector operations in embedding space.

(4) **Anchoring in the Truth Space**: Explicitly model the output space of the AI system as a substructure of Ω; any generation that violates the compatibility condition is automatically rejected.

This is not rhetoric, but the engineering translation of meta-mathematics. The Holism Theorems have already provided rigorous legitimacy criteria; the AI architecture need only implement them as algorithmic constraints.

7 Conclusion

Based on the core theorems of the Zhu–Liang Holism Axiomatic System, this paper has completed the meta-mathematical diagnosis of AI hallucination. The core conclusions are as follows:

(1) **AI hallucination is the recurrence of Russell's paradox in the domain of cognitive engineering**: Both stem from the absence of the compatibility condition $f_Q|_P = f_P$; generation without constitutional constraints necessarily produces illegitimate constructions. [1, Theorem 0.4.1]

(2) **Current AI architectures fully reproduce the threefold murder of reductionist generalization**: Token-level autoregression dissolves compatibility, the "emergence" narrative reverses priority, and embedding space forgets functoriality. [4, Section 4]

(3) **The absence of the Truth Space Ω reduces AI to a rootless formal symbol game**: The system cannot distinguish legitimate constructions from Russellian paradoxes; hallucination is its normal state, not an exception. [2, Theorem 2.3]

(4) **The therapeutic path is to return to the compatibility constitution**: Explicitly construct the whole function F, enforce compatibility verification, restore functoriality, and anchor in the Truth Space.

Final verdicts:

AI hallucination = Inevitable paradoxical output without compatibility constitution = Engineering equivalent of Russell's set.

Current AI architecture = Complete reproduction of the threefold murder of reductionist generalization in cognitive engineering.

Therapy = Return to the Holism constitution: Whole prior to parts, compatibility prior to membership, function prior to set.

The AI crisis is not an engineering limitation, but the inevitable manifestation of an erroneous meta-mathematical foundation. Only by constructing intelligent systems as legitimate substructures of the Truth Space Ω—i.e., taking the whole function F as prior existence and the compatibility condition as a mandatory constitution—can AI ascend from a "machine that generates paradoxes" to a "function that expresses truth." This is a rigorous corollary of the Holism Theorems, not a rhetorical extension.

References

[1] Zhu, J. From Mathematical Foundations to Systematic Philosophy: The Complete Theoretical Chain of Holism Theorems and the Unified Metabolico-Causal Field. Zenodo, 2026. DOI: 10.5281/zenodo.19516417.

[2] Zhu, J. The Zhu–Liang Truth Recursive Nesting Function Theorem (Version 3.5). Zenodo, 2026. DOI: 10.5281/zenodo.19059165.

[3] Zhu, J. The Zhu–Liang Truth Metric Theorem: A Proof that Truth is Necessarily a Function (Version 3.11). Zenodo, 2026. DOI: 10.5281/zenodo.19199103.

[4] Zhu, J. Holism is the True Foundation of Set Theory, Murdered by Reductionist Generalization. Zenodo, 2026. DOI: 10.5281/zenodo.19622151.

Acknowledgments

Thanks to Russell, whose paradox provided the meta-mathematical coordinates for diagnosing AI hallucination. Thanks to all truth-seekers who explore the nature of intelligence within the Holism framework.

Conflict of Interest Statement

The author declares no conflict of interest.

Data Availability Statement

Pure theoretical exposition; no experimental data.

Copyright Notice