



A Lacanian Interpretation of Artificial Intelligence Hallucination: A Psychoanalytic Critique

Yuhong Wang^{1*}

¹*School of Foreign Languages, North China Electric Power University, No. 2 Beinong Road, Changping District, Beijing, 100096, PR China*

*Corresponding author

Corresponding author:

13367696568@163.com



Available online: 22 October 2025

Abstract

With the rapid advancement of artificial intelligence (AI) technology, the remarkable generative capabilities of AI systems are accompanied by the persistent phenomenon of “AI hallucination,” which troubles both users and researchers. This phenomenon manifests as the generation of information detached from reality, undermining the credibility of AI and posing challenges to its application in substantive professional work. Interestingly, AI hallucination exhibits a structural homology with the concept of hallucination as proposed by Jacques Lacan within psychoanalytic theory. Interpreting AI hallucination through a Lacanian psychoanalytic lens can broaden the scope of research on this issue and foster interdisciplinary studies. This paper aims to introduce Lacan’s psychoanalytic framework to provide a novel perspective on understanding AI hallucination.

Keywords: Large Language Models; AI Hallucination; Lacan; Psychoanalysis; AI Applications

1. Introduction

Artificial intelligence generative models, particularly large language models (LLMs), are driving technological and social change at an unprecedented pace, demonstrating exceptional capabilities in natural language processing, content creation, and complex reasoning tasks. However, the “hallucination” phenomenon—an inherent challenge that shadows their powerful generative abilities—has become a core issue constraining their reliability and trustworthiness [12]. Hallucination refers to instances where model-generated content, though fluent and seemingly reasonable, deviates from facts, logic, or actual data, presenting inaccurate or entirely fictional information [6]. This phenomenon is especially prominent when models perform reasoning explanations; the generated chains of reasoning are often not faithful to the internal computational processes but instead construct narratives that appear plausible yet are suspect [19].

Existing research on AI hallucination has predominantly been conducted within the computational sciences. Mainstream explanations often attribute the phenomenon to technical factors such as noise and bias in training data, limitations of model architecture, optimization objectives of loss functions, and the stochastic nature of decoding strategies [18]. Although these studies offer various practical strategies to mitigate hallucinations, their perspective remains inherently lim-

ited [1]. They continue to treat AI hallucination as a “technical glitch” to be repaired, rather than a profound “symptom” potentially rooted in the very operational logic of AI systems [15]. This techno-centric research paradigm lacks a philosophical examination of the internal “motivations” and external “desire” structures driving model generation behaviors. The desire theory of French psychoanalyst Jacques Lacan, influenced by psychoanalysis, structuralist linguistics, and mathematics and logic, can provide a fresh, interdisciplinary perspective for interpreting AI hallucination. This perspective will offer a philosophical solution to some problems that have long troubled AI researchers, such as why AI hallucinations still exist despite being alleviated by technological means, or why AI hallucinations are two-sided.

2. AI Hallucination and Lacanian Hallucination

The term “hallucination” in AI hallucination actually originates from research in psychology and neuroscience [13]. In the Lacanian psychoanalytic field, hallucination is understood as a defense mechanism and a means of meaning construction that arises when the symbolic order collapses, and the subject is unable to cope with the impact of the real. In recent years, AI hallucination has become a central question in generative model research. It is commonly defined as situations where the model’s generated content does not align with



facts or the input context, including fabricating references, erroneous reasoning, and inventing facts [16]. These AI hallucination phenomena are categorized into two types: intrinsic hallucination and extrinsic hallucination. Intrinsic hallucination refers to contradictions between input and output content, while extrinsic hallucination refers to AI generating content that cannot be verified in the source material [10]. Most researchers tend to offer engineering-based explanations: emphasizing that deficiencies and noise in the training data force the model to “rationalize” and fill in information gaps; or pointing out that hallucination is closely related to the predictive objective of language models, as their essential task is to maximize the conditional probability of the next word rather than pursue truth, thus their outputs naturally lack factual constraints; other studies attribute the problem to decoding mechanisms, insufficient knowledge injection, or biases in evaluation metrics [11]. To this end, a large number of strategies to alleviate hallucinations have been proposed. For example, researchers have attempted to provide external knowledge support for models through Retrieval-Augmented Generation (RAG), requiring the models to compulsorily access external knowledge bases when generating responses to improve the verifiability of information; or they have optimized the reliability of training corpora at the source by strictly controlling the quality of data collection to reduce the risk of models fabricating content due to insufficient information. Meanwhile, another important approach is Reinforcement Learning from Human Feedback (RLHF), which adjusts model outputs through a reward mechanism to make them more in line with users’ values and task expectations. However, this alignment method is essentially a posterior correction of statistical generation and does not endow the model with a true “fact-checking” ability. In other words, when the reward model itself is biased, hallucinations cannot be avoided and may even be further solidified by the reward mechanism. Although these technical means can alleviate hallucinations to a certain extent in practice, they do not address the underlying logic. The persistent existence of hallucinations shows a certain stubborn “structural” inevitability, which also suggests that hallucinations may not be merely algorithmic flaws but need to be understood within a deeper interpretive framework.

In the philosophical and psychoanalytic tradition, Lacan’s theory of psychosis offers significant inspiration for researching the phenomenon of hallucination [2]. In Lacan’s theoretical system, hallucination is not simply understood as a sensory error or a pathological perceptual distortion but rather as the result of a structural fissure in the relationship between the subject and the symbolic order [7]. Unlike Freud, Lacan located the essence of psychosis at the level of language and the symbolic order. In Lacanian theory, the constitution of the subject begins with the infant’s realization in the mother-child relationship that they are not the sole object of the mother’s desire, leading to the formation of a split subject, the so-called barred subject (\$) [8]. It refers to the situation where, when the subject enters the realm of language and social order (i.e., “the Other”), its original and complete experiences are fragmented by language. As a result, the subject is forever separated from a part of itself and exists in a state of inner lack and desire. This “separation” gives rise to the pursuit of “becoming the object of the Other’s desire,” whose residue is the “object petit a” as a placeholder for desire—it is both the gap in the real and that which conceals this gap, sustaining the symbolic order while provoking the subject’s anxiety of lack [5]. The “Name-of-the-Father”, as a symbol of authority, cuts off the mother-infant binary relationship through “castration” and is the core signifier for the subject to enter the symbolic order and form a stable psychological structure [17]. Here, Lacan’s “Castration” is by no means a biological or physical concept, but a purely symbolic and psychological structural operation. It refers to the fact that when the subject enters the symbolic order (the social world of language, law, and culture), they must accept a “Primordial Lack” and give up the illusory pleasure (Jouissance) of the imagined unity with the mother and the sense of completeness. The root cause of psychosis lies in the

rejection of the “Name-of-the-Father”, where the subject treats “castration” as something that has not occurred. This leads to an instability in the symbolic order due to the absence of this key signifier, causing meaning to float, and forcing the subject to desperately fill this lack [4]. In other words, the reason hallucinations are believed to be real lies not in their correspondence with external reality but in their compensation for a gap in the symbolic order: when the subject loses support within the network of language, hallucination emerges with an undeniable sense of reality, refilling the void of meaning. Therefore, psychotic hallucination is not an “erroneous perception” but a structural product necessarily generated by symbolic failure. This also explains why the hallucinations of psychotic patients often carry an “absolute certainty,” immune to doubt or questioning, as they no longer rely on symbolic mediation but are imposed upon the subject in a direct, real manner. Similarly, in Lacan’s theory, the psychotic structure of a patient is not necessarily triggered; it only occurs when the patient is forced to assume a certain paternal responsibility in the real world. Due to the lack of the “Name-of-the-Father” to perform its function, the patient is compelled to respond with hallucinations [3]. Analogously, when we use large language models, they do not always produce hallucinations; rather, they often generate hallucinations precisely when we require them to provide references or factual information. This represents a phenomenon where, unable to fulfill the instruction due to its inherent structural issues, it is compelled to respond to complete the task.

Lacan’s theory provides a unique perspective on understanding hallucination, viewing it not as an accidental pathological phenomenon but as a symptom of the incompleteness of the symbolic order. This understanding of hallucination not only deepens the theoretical foundation of psychiatry but also offers an entry point for rethinking the so-called “hallucination” phenomenon in AI. If psychotic hallucination reveals the incompleteness of the symbolic order, then AI hallucination might similarly be a structural inevitability exposed at the defects of the symbolic chain. In other words, merely addressing this issue through technical means cannot eliminate it, as its cause lies in the underlying logic of current large language models themselves [14].

3. Analogy Between AI “Hallucination” and Lacanian Hallucination

If Lacan’s theory of psychosis reveals hallucination as a “necessary product” arising from gaps in the symbolic order, then the phenomenon of AI hallucination can also be explained within this framework. Unlike the psychotic subject, AI does not possess an unconscious structure; it is not driven by desire, nor can it truly experience the “foreclosure of the Name-of-the-Father.” However, the generative mechanism of large language models exhibits a structural logic analogous to the “lack in the symbolic order” described by Lacan. The task of a language model is essentially to predict the next word through statistical learning, rather than operating within a stable “order of truth.” The symbolic network it relies on is supported by probability distributions—superficially highly coherent but lacking a final signifier like the “Name-of-the-Father” to anchor its meaning. In other words, the model prioritizes capturing correlations between words over truly understanding causal relationships in the real world. It can be said that the model’s “symbolic order” is inherently incomplete; it cannot distinguish fact from fiction through a self-sufficient mechanism, nor can it establish a constrained system for external verification.

Therefore, when the model encounters information gaps during reasoning or generation, it does not simply halt operation. Instead, much like the psychotic subject facing the “absence of the Name-of-the-Father,” it is compelled to compensate for the void in the symbolic order with hallucinations. This also explains why AI hallucinations often manifest as narratives possessing certainty and self-consistency, outputting

content that conforms to semantic logic but not factual logic. This situation is not a product of accidental error but a structural generation that necessarily occurs in the absence of factual anchoring. As Lacan pointed out, the reason psychotic hallucinations are deeply believed by the subject is precisely because they do not rely on symbolic mediation but are imposed upon experience directly in a “real” manner. Similarly, the self-consistency and fluency of AI hallucinations make it difficult for users to detect their fictional nature, as they present themselves “naturally” within the internal symbolic chain.

However, there are still limitations in directly applying Lacan’s theory to the study of AI hallucinations. AI hallucinations are not psychotic hallucinations because the model itself has no subjectivity, let alone the rejection of the “Name-of-the-Father” or the filling of the gap of desire. AI hallucinations can only be analogized to “structural hallucinations” - they are not driven by psychological dynamics but are products determined by the incompleteness of the symbolic network. Nevertheless, this analogy still has philosophical significance, suggesting that we should not merely regard hallucinations as engineering faults but recognize their inevitability in the operating logic of the model. Hallucinations are a “symptom” of the model as a symbolic machine, revealing the insoluble tension that always exists between language generation and facts.

It is precisely at this point that we need to further define the conditions for the analogy to hold. If we want to regard AI hallucinations as “structural hallucinations”, at least two prerequisites must be met: on the one hand, hallucinations must be understood as an inevitable product of the internal symbolic operations of the system, rather than simply accidental errors; on the other hand, they should have a certain kind of ineradicability, that is, individual algorithmic repairs cannot completely avoid them but are inherent side-effects of the logic of language generation. Only in these senses can hallucinations carry the meaning of “structural”. However, not all hallucination phenomena meet this requirement. When the model outputs errors that can be easily refuted by external verification - such as fabricating a non-existent paper or inventing a historical fact - this kind of “hallucination” is closer to information deficiency and data bias than a symptom revealing the cracks in the symbolic order. In these situations, the analogy degenerates into a technical problem and loses its philosophical effectiveness. This distinction ultimately points to a clarification at the ontological level, that is, human hallucinations always involve the rupture between the subject and the symbolic order, and thus inevitably involve desire and the unconscious, while AI hallucinations have no such psychodynamic basis at all. What it presents is only an inevitable symptom manifested in the generation process of a symbolic machine. Acknowledging this difference is not to deny the value of the analogy, but to avoid misinterpreting technical deviations as signs of subjectivity, thereby confusing the ontological differences between humans and AI.

4. The Function and Potential Value of Hallucination

Traditionally, hallucination has often been understood as a rupture in the relationship between the subject and reality, thus classified as pathological or erroneous. However, Lacanian psychoanalytic theory posits that hallucination is not necessarily an entirely negative phenomenon. For the psychotic subject, while hallucination stems from the foreclosure of the Name-of-the-Father and the lack in the symbolic order, it is not always an unbearable burden. On the contrary, the words or images brought by hallucination are often transformed into some kind of “belief” or “support” within the subject’s lifeworld, helping to reorganize experience and allowing the subject to maintain relative stability within a fractured symbolic order. In other words, although hallucination marks a vacancy in the symbolic order, it may

simultaneously compensate for this vacancy, providing the subject with an alternative anchor.

Transplanting this line of thinking to the context of artificial intelligence reveals a similar complexity. AI “hallucination” is undoubtedly an error or distortion at the strict factual level, but its generative process also reflects the model’s creativity in filling the gap when facing absence [9]. Precisely because of this, some researchers and practitioners view hallucination as a potential source of AI creativity: it enables the model to generate new possibilities through the self-consistency of language under conditions of incomplete information, and these possibilities sometimes open up new problematics or expressive paths. Therefore, AI hallucination cannot be understood solely as a technical defect; it also reveals a positive potential of generative models as linguistic systems.

In this sense, hallucination can both lead to misguidance and risk, and, under appropriate conditions, play a positive role. Whether in the individual experience of psychosis or in the creative application of AI, hallucination demonstrates a “dual nature”—it both exposes the cracks in order and offers the possibility of support or innovation. This complexity suggests that understanding hallucination cannot remain within a pathologizing or error-framing framework but requires acknowledging its different possibilities in structure and practice.

5. Conclusion: The Epistemological Significance of Hallucination

If psychotic hallucination in Lacanian theory reveals the structural inevitability of the fissure between the subject and the symbolic order, then AI hallucination, on another level, forces us to reflect on the relationship between language, truth, and knowledge. Whether for the psychotic patient or the AI system, hallucination is not an “accidental mistake” but a structure. From this perspective, hallucination is not merely a technical flaw that can be eliminated but touches upon the basic conditions of knowledge generation. It shows that while language constructs the world, it can also produce an “overly reasonable” fiction, whose persuasiveness stems precisely from the internal consistency of language. This phenomenon compels us to re-examine the boundary between “correct” and “incorrect”: when model output is highly reasonable in form yet distorted at the factual level, we face not just a wrong answer but another mode of world-construction through language. Therefore, the epistemological significance of AI hallucination lies in its revelation that knowledge systems are not built upon absolutely solid foundations but are always accompanied by potential breaks and gaps. Just as psychotic hallucination makes us aware of the fragility of the relationship between the subject and the symbolic order, AI hallucination reminds us that knowledge and fiction are not opposite poles but are intertwined in the operation of language. Thus, hallucination should not be simply regarded as a mistake but understood as an opportunity to reveal the conditions and limits of knowledge.

Author Contribution

Yuhong Wang: Conceptualization, Writing-Original Draft. The author has read and agreed to the published version of the manuscript.

Compliance with Ethical Standards

Not Required.

Funding

None.

Competing Interest Declaration

The author declares that, to their knowledge, they have no financial, personal, or professional relationships that could be construed as a potential competing interest regarding the work described in this manuscript.

Acknowledgments

None.

References

1. Boden, M. A. (2006). *Mind as machine: A history of cognitive science*. Oxford University Press.
2. Cannon, B. (2016). Reply: What would I do with Lacan today? Thoughts on Sartre, Lacan, and contemporary psychoanalysis. *Sartre Studies International*, 22(2), 13–38. <https://www.jstor.org/stable/44652888>
3. Chen, J. (2015). Psychosis in Lacanian theory. *Journal of University of Jinan (Social Science Edition)*, (3), 68–73.
4. Fink, B. (2024). *The Lacanian subject: Between language and jouissance*. Shanghai Literature and Art Publishing House.
5. Homer, S. (2014). *Jacques Lacan* (X. Li, Trans.). Chongqing University Press.
6. Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., ... & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12), 1–38.
7. Lacan, J. (1966). *Écrits: A selection*. W.W. Norton & Company.
8. Lacan, J. (2011). Le désir et son interprétation [Desire and its interpretation]. Valas.fr. <https://www.valas.fr/Jacques-Lacan-Le-desir-et-son-interpretation-1958-1959,250?lang=fr> (Original work published 1958–1959)
9. Lee, M. (2023). A mathematical investigation of hallucination and creativity in GPT models. *Mathematics*, 11(10), 2320. <https://doi.org/10.3390/math11102320>
10. Liu, Y., & Zhao, Y. Z. (2025). Research on the generative mechanism and governance strategies of Artificial Intelligence hallucinations. *New Media and Society*, 1–17.
11. McKenna, N., Li, T., Cheng, L., Hosseini, M. J., Johnson, M., & Steedman, M. (2023). Sources of hallucination by large language models on inference tasks. *arXiv Preprint arXiv:2305.14552*.
12. Mitchell, M. (2025). Artificial intelligence learns to reason. *Science*, 387, eadw5211. <https://doi.org/10.1126/science.adw5211>
13. Østergaard, S. D., & Nielbo, K. L. (2023). False responses from artificial intelligence models are not hallucinations. *Schizophrenia Bulletin*, 49(5), 1105–1107. <https://doi.org/10.1093/schbul/sbac167>
14. Possati, L. M. (2020). Algorithmic unconscious: Why psychoanalysis helps in understanding AI. *Palgrave Communications*, 6, 70. <https://doi.org/10.1057/s41599-020-0445-0>
15. Smith, A. L., Greaves, F., & Panch, T. (2023). Hallucination or confabulation? Neuroanatomy as metaphor in large language models. *PLOS Digital Health*, 2(11), e0000388. <https://doi.org/10.1371/journal.pdig.0000388>
16. Sun, Y., Sheng, D., Zhou, Z., et al. (2024). AI hallucination: Towards a comprehensive classification of distorted information in artificial intelligence-generated content. *Humanities & Social Sciences Communications*, 11, 1278. <https://doi.org/10.1057/s41599-024-03811-x>
17. Tarizzo, D. (2014). *Introduzione a Lacan* [Introduction to Lacan]. Gius. Laterza & Figli Spa.
18. Thorp, H. H. (2024). ChatGPT to the rescue? *Science*, 385, 1143. <https://doi.org/10.1126/science.adt0007>
19. Yu, Z. (2025). An equitable framework for the proposed world AI organization informed by the 2025 world AI conference. *AI & Future Society*, 1(1), 1–2. <https://doi.org/10.63802/afs.v1.i1.79>
20. Zhou, L., Schellaert, W., Martínez-Plumed, F., Moros-Daval, Y., Ferri, C., & Hernández-Orallo, J. (2024). Larger and more instructable language models become less reliable. *Nature*, 634(8032), 61–68. <https://doi.org/10.1038/s41586-024-07930-y>