

# IS CREATIVITY HIDDEN IN HALLUCINATION?

Xuhui Jiang<sup>a,b,c,1</sup>, Chengjin Xu<sup>a,1</sup>, Fengrui Hua<sup>a</sup>, Yuxing Tian<sup>a</sup>, Yuanzhuo Wang<sup>b,\*</sup> and Jian Guo<sup>a,\*</sup>

<sup>a</sup>International Digital Economy Academy, IDEA Research, Shenzhen, 518045, China

<sup>b</sup>Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100072, China

<sup>c</sup>School of Computer Science and Technology, University of Chinese Academy of Science, Beijing, 100049, China

## ARTICLE INFO

### Keywords:

Hallucination

Creativity

Large Language Model

## ABSTRACT

Hallucinations are often perceived as a significant limitation in current large language models (LLMs). However, is this always the case? Could these apparent errors harbor seeds of creativity? This article proposes a novel viewpoint, suggesting that certain aspects of hallucinations could unexpectedly contribute to advancements in LLM research, particularly in fostering creativity. Through a detailed examination of specific instances, we illustrate how these perceived misfires might fuel creative thought. Additionally, we introduce a comprehensive framework designed to identify and capitalize on the creative opportunities that these hallucinations provide. Our goal is to reframe the conversation around LLM hallucinations, inspiring researchers to delve deeper into the potential of LLM creativity and recognize it as a valuable asset, rather than a mere flaw.

## 1 Recent Advances of Large Language Models

Recent advancements in artificial intelligence have spotlighted large language models (LLMs) like ChatGPT, known for their exceptional ability to understand and generate natural language. This breakthrough has spurred a global surge in research and applications in generative AI. With the increasing application of LLMs in a diverse range of fields, researchers have become intrigued by a specific phenomenon associated with these models: hallucinations. These are instances where the model's output veers away from factual accuracy or strays from the specific context of the input. Hallucinations often involve the creation of information or responses that are more aligned with the model's internal learned patterns, rather than being based on direct instructions or actual reality [1].

## 2 Hallucinations of Large Language Models

Traditionally, hallucinations are regarded as a harmful factor in LLM inference, and therefore current research [1, 2] primarily examines the limitations of hallucinations in LLMs, highlighting their impact on content accuracy, particularly in critical areas such as finance and law where precision is crucial. These studies focus on identifying and mitigating hallucinations. For instance, the research [1] categorizes hallucinations into two types: factuality and faithfulness. Factuality hallucinations contradict real-world facts, further divided into factual inconsistency and fabrication. Faithfulness hallucinations, on the other hand, deviate from user instructions or context, leading to inconsistencies in instructions, context, and logic. The research [2] extends this classification, noting hallucinations that oppose user input, previously generated context, or facts. This framework aids in understanding LLMs' hallucinations. Moreover, various

strategies for detecting and assessing hallucinations have been proposed [1, 2]. The research [1] discusses fact-based and behavior-based detection methods, including classifier metrics, question-answering metrics, and uncertainty estimation. The research [2] suggests generative and discriminative assessment methods to evaluate LLMs' propensity for hallucinations and their detection abilities. These methods also aim to mitigate hallucinations during pre-training and model inference.


## 3 The Value of Hallucinations

While existing studies have predominantly focused on identifying and mitigating hallucinations in LLMs, often perceiving these as drawbacks, a crucial question arises: **Is hallucination in LLMs always harmful, or is creativity hidden within this apparent error?** This paper explores this vital question, aiming to understand the hallucinations of LLMs and harness their potential creative value. We challenge the conventional view and propose a new perspective on LLMs' hallucinations.

Historical examples, in which diverse forms of hallucinations have sparked revolutionary discoveries, offer insightful parallels and serve as a guide in understanding the potential creative value of hallucinations in LLMs. For instance, as shown in Figure 2(a), consider the notion of factuality hallucinations. Historically, the shift from a geocentric to a heliocentric model of the solar system was a monumental change in scientific thought. Initially, heliocentrism was dismissed as a factual error, much like how LLMs might generate seemingly erroneous information. However, just as Copernicus's heliocentric model eventually revolutionized astronomy, what LLMs produce as 'errors' can lead to novel ideas, challenging conventional wisdom. Similarly, faithfulness hallucinations in LLMs can be likened to the accidental discovery. Alexander Fleming's unintended experiment resulted in a groundbreaking medical breakthrough: penicillin. This mirrors how deviations from standard or expected outputs in LLMs, akin to an experiment gone awry, can yield

\*Corresponding author

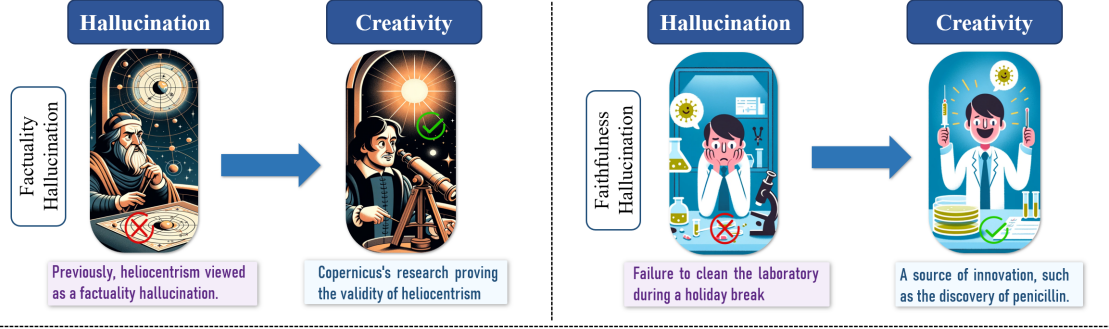
\*\*Principal corresponding author

 jiangxuhui19@ict.ac.cn (X. Jiang); xuchengjin@idea.edu.cn (C.

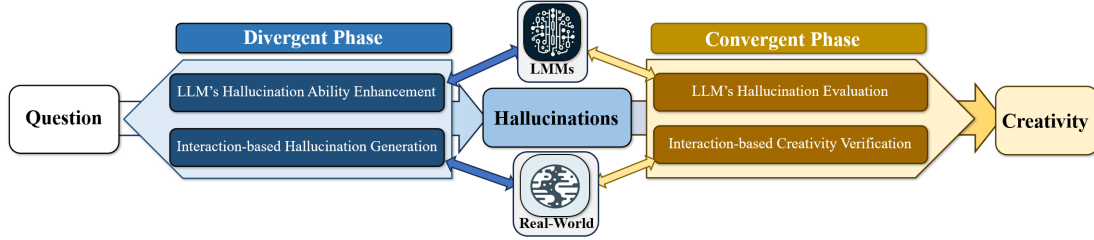
Xu); guojian@idea.edu.cn (J. Guo)

ORCID(s): 0000-0002-1741-0781 (X. Jiang)

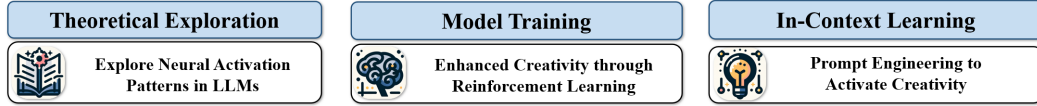
(a) Relation between hallucination and creativity



(b) Divergent-convergent framework for the hallucination-creativity transformation



(c) Potential research directions of leveraging hallucinations for creativity



**Figure 1:** (a) Case studies of relations between hallucination and creativity; (a) Framework for the hallucination to creativity transformation; (c) Future research directions of leveraging hallucinations for creativity.

unexpectedly beneficial results. These examples highlight how both factuality and faithfulness hallucinations in LLMs could be pivotal in driving innovation. By recognizing the potential value hidden in these 'errors,' we can better utilize LLMs for creative problem-solving in various fields.

Expanding upon these historical insights, our focus shifts to recent studies that delve into the relationship between hallucinations in LLMs and creativity. In the realm of cognitive science, the hallucinations experienced by humans are seen as crucial for simulating creative thought processes. For example, research [3] on human creativity indicates that creative thinking involves both the activation of the left prefrontal cortex, known for its role in imaginative thinking and the engagement of the right hippocampus, essential for memory processing. Such insights underscore the idea that creativity is not merely about retrieving information but recombining and expanding upon existing knowledge, which is similar to hallucination. Further exploration [4] reveals parallels between the brain's activation during the creation of novel musical sequences and the improvisation of new melodies, suggesting a link between creativity and random processes. This insight is significant, hinting at ways to enhance creative thought. Drawing parallels to human cognition, LLMs also blend vast training data in novel ways, producing original content for complex tasks. Although their outputs may sometimes stray from factual accuracy, these deviations can reflect a creative leap akin to aspects of human creativity. For instance, research [5] shows that LLMs,

when responding to ambiguous or open-ended questions, often generate novel ideas. While these responses might be factually inconsistent, they can ignite new discussions and innovative problem-solving methods. Similar to the human brain, this comparison suggests that LLMs' creative output can stem from a blend of existing knowledge and innovative data processing, even when manifested as hallucinations. Thus, these phenomena in LLMs may be more than mere errors; they could be gateways to unexplored realms of creativity and innovative thought.

#### 4 Hallucination And Creativity Evaluation

Given the recognition of hallucinations as potential catalysts for creativity in LLMs, a pivotal question arises: **How can we effectively harness hallucinations for creative thought while also mitigating their risks?** We aim to deepen our understanding and use of these phenomena to enhance the creativity and efficiency of LLMs in various applications without compromising accuracy. This requires methods that not only induce hallucinations but also critically evaluate them, transforming them into innovative sources. This endeavor involves a dual approach: first, to stimulate, and second, to critically assess hallucinations, aiming to transform them into inventive outcomes. In terms of framework development, recent research [6] defines creativity as a blend of generative and evaluative processes, encompassing both divergent and convergent thinking. Inspired by this, we propose a two-stage generative framework,

as depicted in Figure 1 (b). This framework consists of two distinct phases: a divergent phase and a convergent phase. During the divergent phase, the emphasis is on generating topic-relevant hallucinations by linking loosely associated concepts to foster novel idea creation. Current LLMs are still evolving in this realm [7], highlighting the necessity for further research in this direction, whether through specialized prompts, LLM training enhancements, or interactions with agents and real-world contexts. The convergent phase, in contrast, focuses on refining these hallucinations into concrete creative contributions. This involves precise identification and categorization of hallucinations, selecting those that are constructive and align with user requirements, through simulations or interactions with real-world scenarios. Drawing on established frameworks for assessing human creativity, which consider factors such as fluency, flexibility, originality, and elaboration [8], this methodology can be adapted to effectively evaluate the creative output of LLMs.

## 5 Future Directions

Summarizing the above, we propose future research directions, illustrated in Figure 1 (c), for leveraging the phenomenon of hallucinations to enhance the creative capabilities of LLMs. These directions aim to transform hallucinations from potential obstacles into creative catalysts, starting from the perspectives of theoretical exploration, training, and in-context learning.

In the realm of theoretical explorations, drawing from studies on the neural mechanisms of human brain creativity [4], future studies can explore similar neuronal activation patterns in LLMs. Particularly, considering the neural activation patterns associated with creative thinking and hallucinations in the human brain, we can analogize these findings to explore the activation patterns of attention mechanisms in LLMs to understand and enhance their creativity and hallucination phenomena.

In the realm of model training, Pressing's theory of improvisational creativity [9] suggests that improvisation is an acquired skill, requiring extensive training to reach a professional level. Based on this theory, we can design a reinforcement learning framework for LLMs aimed at enhancing their creative abilities. This framework simulates the neural mechanisms of the human brain in improvisation, particularly the activation of the prefrontal cortex and motor cortex, emphasizing divergent thinking and cognitive flexibility. Through deliberate practice and closed-loop feedback mechanisms, this framework will enable LLMs to accumulate expertise in specific creative tasks and continually self-improve through generation-evaluation cycles. The model will be trained to flexibly utilize its extensive knowledge base and integrate user feedback to optimize creative outputs. This approach not only enhances the model's innovative capabilities but also advances our understanding and simulation of human creative thinking.

In the realm of in-context learning, by setting specific prompts and contexts, LLMs will continuously learn and adapt in actual contexts to enhance their understanding and

response capabilities for complex creative tasks. For example, giving a prompt like: "*prompt: to come up with something clever, humorous, original, compelling, or interesting*" has been validated as effective in research [10]. Moreover, compared to other methods, this approach is more readily implementable when users have creative needs, enabling LLMs to explore and respond to users' creative requirements more deeply, thereby exhibiting a higher level of creativity in practical applications.

In summary, this paper reconsiders the phenomenon of hallucinations in LLMs from the perspective of creativity. Specifically, the paper reviews relevant cognitive science and artificial intelligence research and finds the connection between hallucinations and creativity. This connection points out that future research on LLMs should not only aim at the challenges caused by hallucinations but also strive to unearth their opportunities for creativity. Finally, the paper proposes a general framework to realize hallucination-creativity transformation, and discusses valuable future research directions from theoretical, model training, and in-context learning, providing a comprehensive view on harnessing the creative potentials of hallucinations in LLMs.

## 6 Declaration of Interests

The authors declare no competing interests.

## References

- [1] Hongbin Ye, Tong Liu, Aijia Zhang, Wei Hua, and Weiqiang Jia. Cognitive mirage: A review of hallucinations in large language models. *arXiv preprint arXiv:2309.06794*, 2023.
- [2] Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. Siren's song in the ai ocean: A survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*, 2023.
- [3] Mathias Benedek, Emanuel Jauk, Andreas Fink, Karl Koschutnig, Gernot Reishofer, Franz Ebner, and Aljoscha C Neubauer. To create or to recall? neural mechanisms underlying the generation of creative new ideas. *NeuroImage*, 88:125–133, 2014.
- [4] Roger E. Beaty. The neuroscience of musical improvisation. *Neuroscience & Biobehavioral Reviews*, 51:108–117, 2015.
- [5] Steven R Rick, Gianni Giacomelli, Haoran Wen, Robert J Laubacher, Nancy Taubenslag, Jennifer L Heyman, Max Sina Knicker, Younes Jeddi, Hendrik Maier, Stephen Dwyer, et al. Supermind ideator: Exploring generative ai to support creative problem-solving. *arXiv preprint arXiv:2311.01937*, 2023.
- [6] Jeff Pressing. Psychological constraints on improvisational expertise and communication. In *the course of performance: Studies in the world of musical improvisation*, pages 47–67, 1998.
- [7] Mika Koivisto and Simone Grassini. Best humans still outperform artificial intelligence in a creative divergent thinking task. *Scientific reports*, 13(1):13601, 2023.
- [8] Paul J Silvia, Beate P Winterstein, John T Willse, Christopher M Barona, Joshua T Cram, Karl I Hess, Jenna L Martinez, and Crystal A Richard. Assessing creativity with divergent thinking tasks: exploring the reliability and validity of new subjective scoring methods. *Psychology of Aesthetics, Creativity, and the Arts*, 2(2):68, 2008.
- [9] Jeff Pressing. Improvisation: Methods and models. In *Physical Theatres: A Critical Reader*, pages 66–78. Routledge, 2007.
- [10] Roger Beaty, Bridget Smeekens, Paul Silvia, Donald Hodges, and Michael Kane. A first look at the role of domain-general cognitive and creative abilities in jazz improvisation. *Psychomusicology: Music, Mind, & Brain*, 23, 11 2013.