OPEN ACCESS

AI for Society REVIEW

INNOVATION PRESS

AI Plus

AI Plus 1 (2026):100008
www.the-innovation.org/ai-plus

Extracting creativity from hallucination: Rethinking large language models

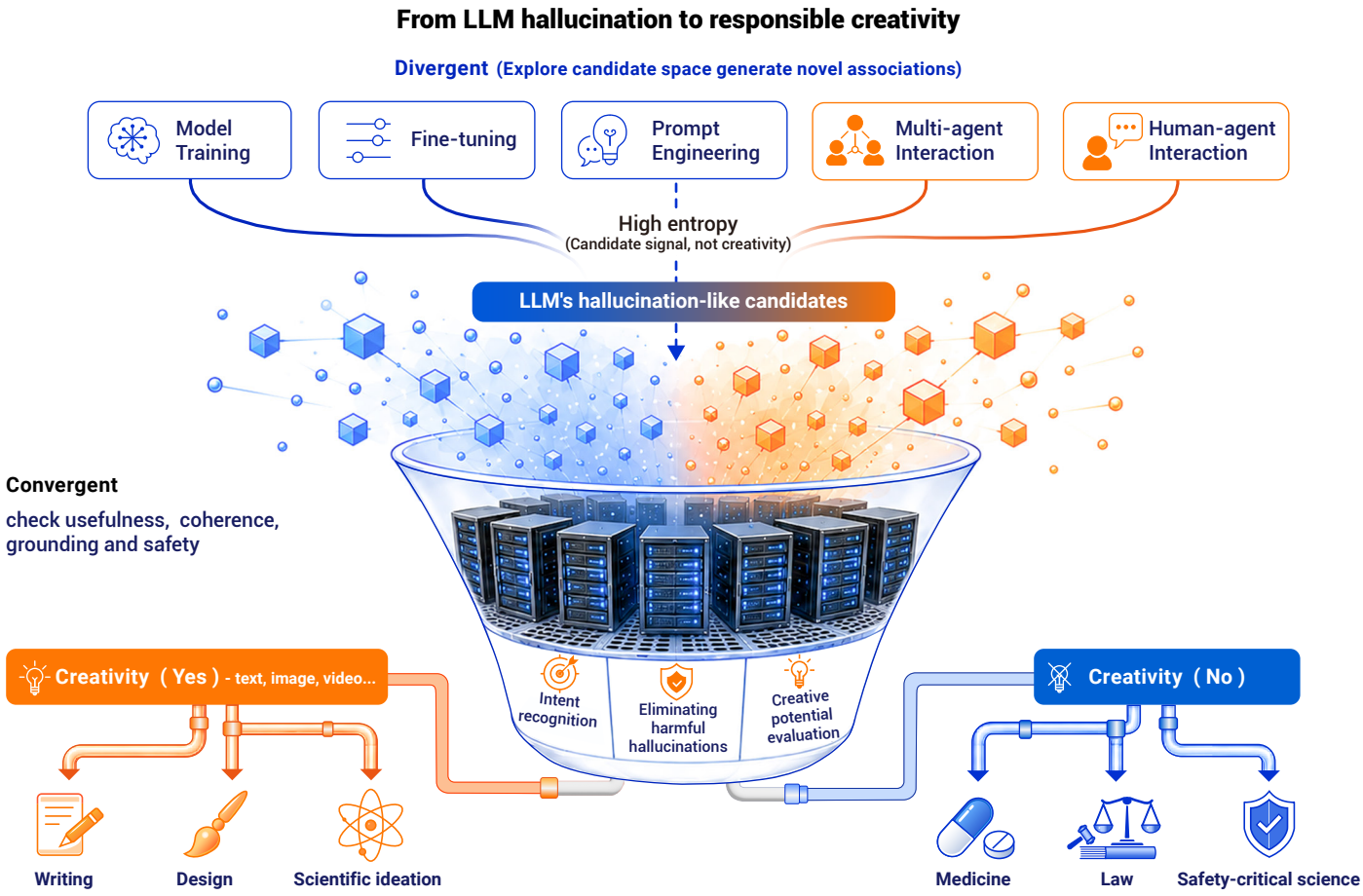
Xuhui Jiang,[1,2,4,8] Yi Liu,[3,8] Yinghan Shen,[1,4,8] Chengjin Xu,[2,4] Yuxing Tian,[2] Juan Chen,[1] Xuehao Zhai,[6] Yihang Liu,[1,5] Fengrui Hua,[2] Bobai Zhao,[7] Yuanzhuo Wang,[1,*] and Jian Guo[2,*]

*Correspondence: wangyuanzhuo@ict.ac.cn (Y.W.); guojian@idea.edu.cn (J.G.)

Received: May 21, 2026; Accepted: August 15, 2026; Published Online: August 18, 2026; https://doi.org/10.59717/ipj.aiplus.2026.100008

© 2026 The Author(s). This is an open access article under the CC BY license (https://creativecommons.org/licenses/by/4.0/).

GRAPHICAL ABSTRACT



PUBLIC SUMMARY

- Large language model (LLM) hallucinations can be errors in factual tasks but may aid open-ended ideation.
- Creative value arises only after hallucination-derived candidates pass tests of novelty, usefulness, and task fit.
- Divergent methods expand the candidate space, while convergent methods filter risks and select useful outputs.
- This review maps methods, evaluation criteria, application domains, and responsible-use boundaries.

REVIEW AI for Society

OPEN ACCESS

INNOVATION PRESS

AI Plus

AI Plus 1 (2026):100008
www.the-innovation.org/ai-plus

Extracting creativity from hallucination: Rethinking large language models

Xuhui Jiang,[1,2,4,8] Yi Liu,[3,8] Yinghan Shen,[1,4,8] Chengjin Xu,[2,4] Yuxing Tian,[2] Juan Chen,[1] Xuehao Zhai,[6] Yihang Liu,[1,5] Fengrui Hua,[2] Bobai Zhao,[7] Yuanzhuo Wang,[1,*] and Jian Guo[2,*]

[1]Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China
[2]International Digital Economy Academy, IDEA Research, Shenzhen 518045, China
[3]School of Science and Technology, Beijing Open University, Beijing 100081, China
[4]DataArc Technology Limited, Shenzhen 518045, China
[5]University of Chinese Academy of Sciences, Beijing 100049, China
[6]Department of Civil and Environmental Engineering, Imperial College London, London SW7 2AZ, United Kingdom
[7]College of Computer Science, Beijing Information Science and Technology University, Beijing 100192, China
[8]These authors contributed equally
*Correspondence: wangyuanzhuo@ict.ac.cn (Y.W.); guojian@idea.edu.cn (J.G.)

Received: May 21, 2026; Accepted: August 15, 2026; Published Online: August 18, 2026; https://doi.org/10.59717/ipj.aiplus.2026.100008

© 2026 The Author(s). This is an open access article under the CC BY license (https://creativecommons.org/licenses/by/4.0/).

Citation: Jiang X., Liu Y., Shen Y., et al. (2026). Extracting creativity from hallucination: Rethinking large language models. *AI Plus* **1**:100008.

Hallucinations in large language models (LLMs) are always seen as limitations. However, could they also be a source of creativity? This survey explores this possibility, suggesting that hallucinations may contribute to LLM application by fostering creativity. Hallucinations are not treated as creativity perse; rather, they are considered candidate materials that require novelty, usefulness, task alignment, semantic structure, safety boundaries, and explicit verification before they can contribute to creative outcomes. This survey begins with a review of the taxonomy of hallucinations and their negative impact on LLM reliability in critical applications. Then, through historical examples and recent relevant theories, the survey explores the potential creative benefits of hallucinations in LLMs. To elucidate the value and evaluation criteria of this connection, we delve into the definitions and assessment methods of creativity. Following the framework of divergent and convergent thinking phases, the survey systematically reviews the literature on transforming and harnessing hallucinations for creativity in LLMs. Finally, the survey discusses future research directions, emphasizing the need to further explore and refine the application of hallucinations in creative processes within LLMs.

INTRODUCTION

Recent advancements in artificial intelligence have propelled large language models (LLMs) into the spotlight, marking a significant leap in their ability to comprehend and generate natural language. This surge in progress has sparked a global wave of research and practical applications, highlighting the transformative impact of LLMs across various domains. Among these advancements, the phenomenon of hallucination within LLMs has emerged as a focal point of investigation. Characterized by the models' tendency to produce unfounded or misleading information without solid data backing, hallucination poses a challenge to the reliability and applicability of LLMs.[1] Despite the increasing volume of research, the comprehensive understanding of hallucination for LLMs remains to be explored.

This study begins by outlining the taxonomies of hallucinations in LLMs, setting the stage for an in-depth review of key research efforts aimed at identifying and reducing these occurrences. It highlights how existing literature predominantly views hallucinations as detrimental, advocating for strategies to minimize their presence, particularly in serious application scenarios like legal and financial.

However, a key question raises and provokes deep reflection: "*Is hallucination in LLMs always harmful, or does creativity hide in hallucinations?*" Different from previous surveys or studies about hallucination, this paper revisits the phenomenon from a positive perspective. In addition to the negative impacts of hallucination on the reliability of LLMs, this paper recognizes a trend in research on the creativity of LLMs and explores the interplay between hallucination and creativity, as well as how to unearth the value of LLM hallucination from the perspective of creativity.

In our exploration of the interplay between LLMs' hallucinations and creativity, we scrutinize notable historical examples where hallucinations have catalyzed creative breakthroughs. By examining these instances, we aim to uncover the complex dynamics between human creativity and hallucination, drawing insights from cognitive science underpinned by pertinent scholarly work. Furthermore, this paper reviews recent studies that focus on this specific interplay in the realm of LLMs, underscoring this critical interplay. This analysis lays the groundwork for a deeper comprehension of the symbiotic relationship between hallucination and creativity.

Recognizing this interplay, this paper delves into evaluating its significance and value. We begin by examining the concept and evaluation of creativity through the lens of cognitive science, followed by a thorough review of the studies on LLM creativity, to understand its critical role and potential. This discussion leads to the assertion: "*While minimizing hallucination risks, it's vital to assess and harness its creative potential, maximizing the value of LLM hallucinations.*"

To make this position explicit, this review separates three related but nonequivalent notions. A factual error denotes an output that conflicts with verifiable knowledge and should be suppressed in factual or high-stakes settings. A productive hallucination denotes a model deviation that can provide narrative, associative, or exploratory value under suitable task constraints, consistent with recent arguments that some confabulations may function as resources rather than merely failures.[2] A creative hallucination is narrower: it refers only to a hallucination-derived candidate that survives a convergent evaluation process and demonstrates both novelty and usefulness, the two core criteria in the standard definition of creativity.[3] Accordingly, factuality and faithfulness describe what a hallucination conflicts with, whereas divergent and convergent phases describe how candidate outputs are generated and filtered.

To achieve the transformation from hallucination to creativity, this paper follows the divergent and convergent phases of cognitive science to systematically review the literature on harnessing hallucination in LLMs for creativity. Specifically, we thoroughly explore the divergent phase, focusing on the methods and research that enhance LLMs' ability to generate creative hallucinations. Subsequently, we delve into the convergent phase, examining how these hallucinations can be critically assessed and refined into valuable creative outputs. Through this bifocal approach, the study highlights practical implications and potential future directions in leveraging LLM hallucinations for creative endeavors.

In conclusion, this paper casts an eye toward the horizon of future

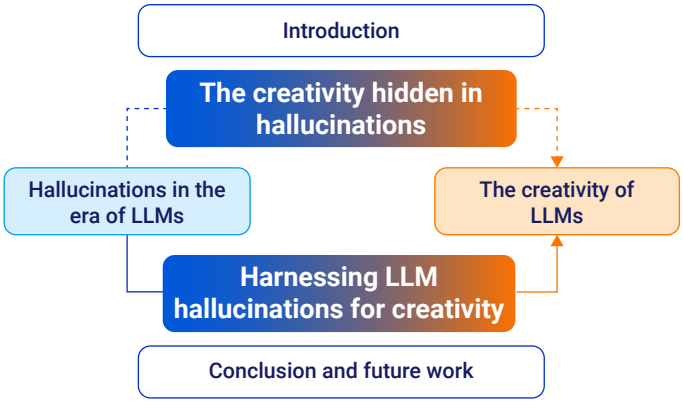


Figure 1. General framework of the survey.

research, contemplating strategies to broaden and deepen the exploration of the interplay between hallucination and creativity within LLMs.

The rest of this paper is organized as shown in Figure 1. Hallucinations in the Era of LLMs reviews existing research on LLM hallucination. The Creativity Hidden in Hallucinations questions the prevailing negative view of hallucinations and explores their interplay with creativity. The Creativity of LLMs addresses the definition and assessment of creativity in LLMs. Harnessing LLM Hallucinations for Creativity investigates how to leverage the interplay between hallucination and creativity. Finally, Conclusion and Future Work outlines future directions for this field.

HALLUCINATIONS IN THE ERA OF LARGE LANGUAGE MODELS

The advent of LLMs like ChatGPT and LLaMA marks a significant milestone in the field of AI.[4] These models, trained on vast datasets, have demonstrated remarkable proficiency in generating coherent text and opening new frontiers in AI-assisted writing, conversation, and content creation. Their versatility and efficiency have made them indispensable in both academic research and commercial applications. However, the ascendancy of LLMs brings with it a complex challenge: the phenomenon of 'hallucination', where models confidently generate incorrect or irrelevant information. This often stems from issues like erroneous information in the training data or the problematic alignment process.[5] As LLMs are increasingly applied in various aspects of our lives, comprehensively understanding, detecting, and mitigating hallucinations has become crucial, serving not only as a focal point in academic research but also as a key prerequisite for the reliable deployment of these powerful AI systems.

Hallucination taxonomy

In dissecting the phenomenon of hallucinations within LLMs, it is crucial to categorize these instances for targeted and effective resolutions. A widely accepted classification method identifies two main types of hallucinations: factuality hallucinations and faithfulness hallucinations.[1] Factuality hallucinations encompass instances where the content generated by LLMs diverges from verifiable real-world facts. This category manifests in two primary forms. The first is factual inconsistency, in which the model's output contradicts known real-world information. The second form is factual fabrication, in which the model generates content that has no basis in reality and cannot be substantiated with factual data. The second main type, faithfulness hallucination, occurs when the content produced by LLMs does not align with the user's instructions or the context they have previously generated. This category includes three distinct phenomena. Instruction inconsistency is observed when the model's output strays from the user's specific directives. Context inconsistency arises when the output, although possibly factually correct, is irrelevant or inappropriate for the given situation. Logical inconsistency refers to flaws in the model's reasoning processes or a mismatch between the reasoning and the final output. Using a different classification criterion, prior work categorizes hallucinations as intrinsic or extrinsic.[6,7] Other work classifies hallucinations according to the source of conflict: input-conflicting, where responses go against user inputs; context-conflicting, involving contradictions within the generated context; and fact-conflicting, where outputs clash with established factual knowledge.[5]

The above taxonomy should not be conflated with the divergent-convergent framework used later in this review. Factuality and faithfulness classify hallucinations by their object of conflict: external world knowledge, user instructions, local context, or reasoning consistency. Divergent and convergent phases classify the creative process: the former expands the candidate space, while the latter filters, validates, and selects usable outputs. These two lenses intersect. For example, a fact-conflicting output can sometimes serve as a speculative prompt in a low-risk ideation task, but it remains unacceptable in medical, legal, financial, or scientific reporting unless independently verified. Conversely, a faithful response can still be creatively weak if it lacks novelty or usefulness. Table 1 summarizes this distinction. Its conflict-based categories are drawn from hallucination taxonomies in LLM and NLG surveys,[1,5,6,8] while its process-based categories are drawn from divergent-convergent creativity theory.[9,10] We set up the table this way to avoid mixing two different classification criteria: what the output conflicts with and how a candidate is generated or filtered.

Furthermore, several studies have categorized hallucinations across different applications involving LLMs. One study provides a summary of representative hallucinations in numerous downstream tasks, such as machine translation, question and answer, dialog systems, and summarization systems, among others.[1] Another study provides a structured summary and discussion, categorizing and addressing hallucinations specifically within LLMs, multilingual LLMs, and domain-specific LLMs.[14] These classifications provide a clear framework for understanding the hallucinations that occur in LLMs during content generation and reveal the challenges and limitations faced by LLMs in various tasks.

Hallucination detection

Catering to the phenomenon of hallucination, existing studies have proposed various detection and evaluation strategies. The following studies organize these strategies using different criteria. Reflecting the proposed classification of hallucinations, one survey has also summarized the corresponding detection methods for each category.[8] For factuality hallucinations, they recommend strategies that involve retrieving external facts and implementing uncertainty estimation. As for faithfulness hallucinations, they also delineate several approaches, including classifier-based, QA-based, and prompting-based metrics, to sidestep the potential pitfalls of contradictory outputs. Another survey categorizes the approaches into two types: generation and discrimination.[5] Generation methods consider hallucination as a generation characteristic, and evaluate the generated texts from LLMs, while discrimination approaches assess LLMs' capability to distinguish between truthful and hallucinated statements. Using concrete detection techniques as its organizing criterion, another study proposes methods like reasoning classifiers, uncertainty measures, and self-assessment for identifying hallucinations.[1]

Furthermore, some research efforts focus on creating benchmarks specifically designed for various application scenarios of LLMs in hallucination detection. HaluEval presents a large collection of generated and human-annotated hallucinated samples for assessing LLMs' hallucination detection capabilities, revealing that incorporating external knowledge or adding reasoning steps can improve recognition accuracy.[15] TruthfulQA extends its assessment to 38 domains including law, finance, and politics, employing both manual methods and automated fine-tuning of LLMs for evaluation.[16] Considering the serious consequences in healthcare applications, Med-HALT proposes a two-tiered approach, encompassing Reasoning Hallucination Tests (RHTs) and Memory Hallucination Tests (MHTs), to evaluate the presence of hallucinations.[17] In the scenarios mentioned, such as law, finance, and healthcare, hallucinations are typically detrimental. However, it's worth contemplating whether, in certain domains, hallucinations might not always need to be viewed as problematic.

Hallucination reduction

Drawing from the detailed analysis of hallucination types and detection methods in LLMs outlined in the preceding sections, some research has focused on developing targeted strategies to reduce these hallucinations. The following surveys organize mitigation approaches using different criteria.

OPEN ACCESS

Table 1. Relationship between hallucination taxonomies and the divergent-convergent creativity framework.

Lens	Question answered	Main categories	Role in this review
Conflict-based hallucination taxonomy	What does the output conflict with?	Factuality and faithfulness hallucinations; input-, context-, and fact-conflicting hallucinations.[1,5,6,8]	Identifies risks and determines what must be checked or suppressed
Divergent phase	How are candidate ideas expanded?	Prompting, sampling, layer steering, multi-agent debate, human-agent interaction.[9–11]	Encourages diverse candidate outputs, including controlled deviations from the most probable response
Convergent phase	How are candidates filtered into useful outputs?	Grounding, verification, expert judgment, self-checking, human review.[9,10,12,13]	Distinguishes harmful errors from potentially useful, novel, and task-aligned candidates

One survey classifies approaches to mitigate hallucinations in LLMs based on the timing of their application within the LLM life cycle. For instance, during the reinforcement learning from human feedback (RLHF) stage, a specific reward score targeting hallucinations can be designed and directly optimized through reinforcement learning. In the model inference stage, decoding strategies and the retrieval of external knowledge can be employed for retrieval-based enhancement.[5] Using mitigation method as its organizing criterion, another survey summarizes five methods for addressing hallucinations, including parameter adaptation and leveraging external knowledge, and applies these approaches across various downstream tasks.[1] Using several parameters rather than life-cycle stage or method type, another survey presents an elaborate taxonomy based on dataset utilization, common tasks, feedback mechanisms, and types of retrievers.[18] This systematic classification effectively differentiates between the various specialized approaches developed for mitigating hallucination issues in LLMs. It also provides a detailed summary of various strategies based on prompt engineering for reducing hallucinations, including retrieval augmented generation (RAG), and self-refinement.

Particularly, some research has focused on exploring how knowledge graphs (KGs) can aid LLMs in reducing hallucinations.[19–21] One study argues that KG-augmented retrieval techniques effectively address hallucination issues by expanding the model's knowledge with non-parametric factual knowledge.[19] These studies not only enhance our understanding of the operational mechanisms of LLMs but also provide critical guidance for improving the accuracy and reliability of models in practical applications.

THE CREATIVITY HIDDEN IN 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 can some hallucination-like deviations become useful in creative workflows?* This paper explores this question by reviewing historical analogies, cognitive science literature, and recent LLM studies. Importantly, this review does not claim that hallucination itself is creative. Instead, it argues that some hallucinations can act as exploratory candidates during divergent thinking, provided that convergent evaluation later tests their novelty, usefulness, coherence, and factual or contextual acceptability. Recent research has introduced related notions such as productive hallucination and confabulation, emphasizing that some deviations may supply value in open-ended tasks when they are appropriately constrained and verified.[2]

Terminological scope. In this review, hallucination refers to a fluent output that is unsupported, unfaithful, inconsistent, or otherwise in conflict with facts, instructions, context, or reasoning constraints. Productive hallucination refers to a hallucination or confabulation that supplies associative, narrative, or exploratory value in an open-ended task, without implying that it is already correct or creative. Creative hallucination is reserved for hallucination-derived material that has passed convergent screening and satisfies the creativity criteria of novelty, usefulness, and task appropriateness emphasized in recent LLM creativity studies.[3] This terminology prevents productive hallucination from being used as a blanket justification for factual error.

Historical examples are useful as analogies, but they should be interpreted cautiously. Figure 2 is therefore used only as an illustrative map of possible relations between hallucination-like deviation and later creative value, not as evidence that factual errors are creative by themselves. Cases such as heliocentrism or penicillin show a narrower point: deviations from prevailing assumptions, accidents, or initially implausible candidates can sometimes trigger exploration when a community subsequently subjects them to evidence, interpretation, and validation. This distinction is essential for LLMs. In open-ended ideation, a model deviation may help users search a broader possibility space; in high-risk factual settings such as law, medicine, finance, and safety-critical science, the same kind of deviation should be treated as an error and suppressed or verified before use.[22,23]

Expanding upon these historical insights, the clue of the interplay between hallucination and creativity can also be explored in cognitive science studies. The hallucinations experienced by humans are seen as crucial for simulating creative thought processes. For example, research 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.[24] 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.[1] Furthermore, neuroscience research on creativity, exemplified by studies of improvisation in jazz musicians, parallels this concept.[25] By monitoring different brain regions, researchers have discovered that improvisation activates the same areas as pseudo-random motor movements. This similarity suggests a link between the spontaneity of pseudo-random movements and creative processes, further implying an interplay between hallucinations and creativity. These cognitive science findings also provide valuable insights for research in LLMs. High uncertainty may indicate exploratory divergence from the most likely continuation, but it should not be taken as direct evidence of semantic-level logical reconstruction. Whether a high-uncertainty output is meaningful must be assessed with additional signals, such as coherence, task utility, grounding, interpretability evidence, and human or automated validation. This qualification is consistent with recent hallucination surveys emphasizing that diversity-oriented generation still requires grounding and reliability constraints.[8,22]

In the realm of LLMs, there's an increasing focus on the interplay between hallucinations and creativity. As Sam Altman, CEO of OpenAI, elucidates, there is a profound and often overlooked connection between the hallucinatory phenomena in LLMs and their creative output.[26] Altman argues that LLM hallucination is a cornerstone in unleashing the creative potential of AI. Recent studies also support this view. The research provides a theoretical reference for understanding the interconnectedness of hallucination and creativity in LLMs, substantiating their correlation through rigorous mathematical analysis.[27] This mathematical analysis work introduces two key assumptions regarding hallucinations in GPT models: (1)*in ambiguous contexts, the estimated probabilities* $p(x_{i+1})$ *of GPT models exhibits relatively small differences between the highest and subsequent probabilities*; (2)*hallucination takes place when the GPT model generates a low-probability token and subsequently this token is employed as input for next prediction*. This work then use entropy to define the uncertainty of the GPT model's prediction at position $i+1$ as Equation 1:

$$H(x_{i+1}|x_1,\ldots,x_i;\Theta) = -\sum_{x'\in\mathcal{V}} p(x_{i+1}|x_1,\ldots,x_i;\Theta)\log p(x_{i+1}|x_1,\ldots,x_i;\Theta) \quad (1)$$

Based on the two key assumptions, both the highest and subsequent probabilities are relatively low when hallucination occurs, and according to

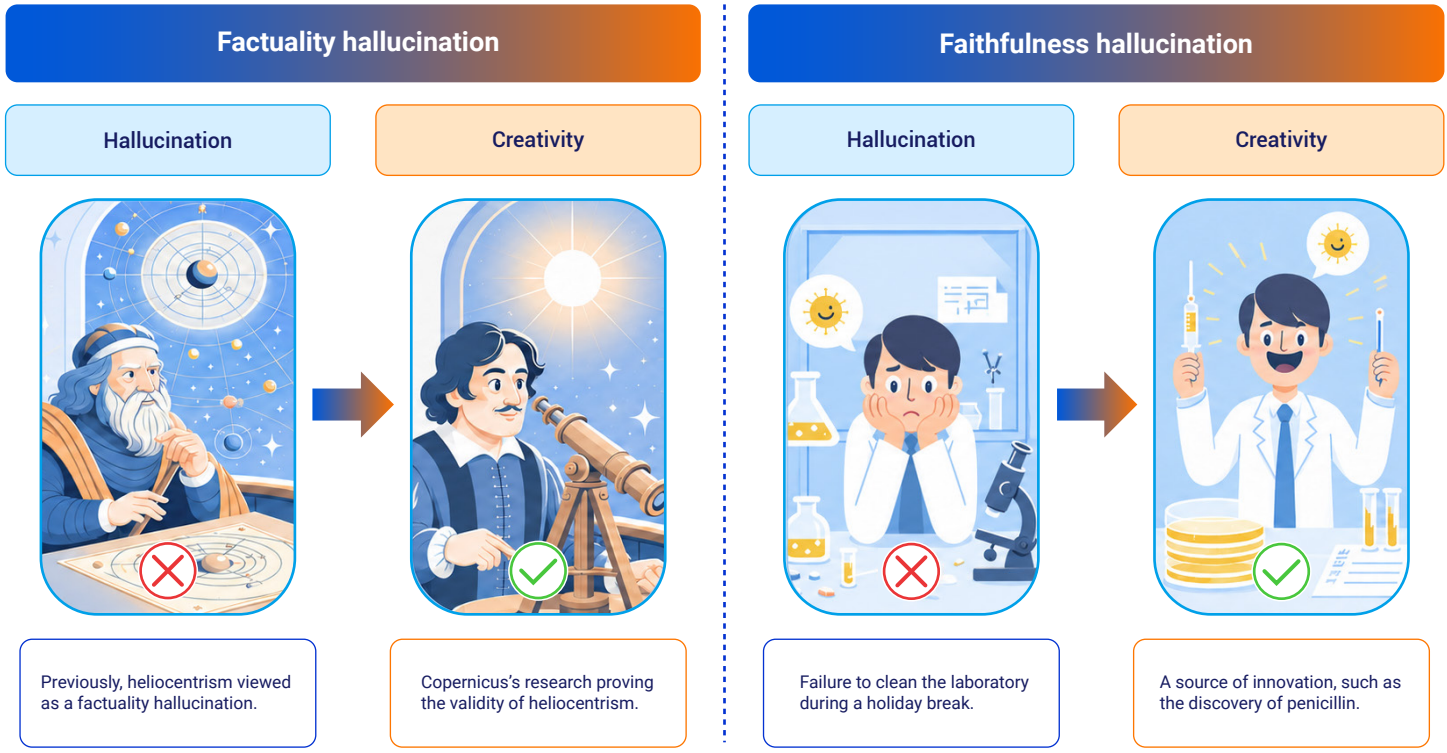


Figure 2. Cases of relations between hallucination and creativity.

Equation 1, the uncertainty will be high. Therefore, this work concluded that hallucination implies high uncertainty. Then this work defines the creativity in GPT model's prediction at position $i+1$ as Equation 2:

$$C(x_{i+1}|x_1,\ldots,x_i;\Theta) = \frac{H(x_{i+1}|x_1,\ldots,x_i;\Theta)}{H_{\max}(x_{i+1})} \quad (2)$$

where $H_{\max}(x_{i+1})$ is the maximum entropy achievable for given vocabulary $\mathcal{V}$, which occurs when all tokens have uniform probability and equals to $\log|\mathcal{V}|$. Equation 2 should therefore be interpreted as an uncertainty or divergence proxy rather than a standalone creativity score. High predictive entropy is a necessary but not sufficient condition for creativity. A uniform distribution over the vocabulary maximizes entropy, but such a distribution can generate unstructured noise with no semantic coherence, utility, or task relevance. Thus, entropy can indicate that the model has entered a broad candidate region, but convergent screening is required to determine whether any candidate is creatively valuable. The normalization term also limits cross-model interpretation. Since $H_{\max}(x_{i+1}) = \log|\mathcal{V}|$ depends on vocabulary size and tokenization scheme, the absolute value of C is comparable only when the tokenizer and vocabulary are fixed. Subword segmentation choices change the units over which entropy is computed, so this metric is most appropriate for within-model or controlled-tokenizer analysis, not as a direct benchmark across unrelated LLM families.[28] Two surveys also support this point: one suggests that this interplay is beneficial,[29] while the other portrays hallucination models as 'collaborative creative partners'.[14] Even though their outputs might deviate from factual or instruction, they serve as catalysts for innovative thought. Meanwhile, recent interpretability and creativity-oriented studies have begun to examine how internal representations, decoding layers, and associative mechanisms may mediate the tension between memorization, factuality, and creative divergence.[3,30] These findings are promising, but they should be treated as early evidence rather than proof that hallucination automatically corresponds to semantic reconstruction or creative value.

Despite the emerging evidence linking hallucination and creativity in LLMs, research in this field is still limited. There is a significant need for further research to both deepen our understanding of the value of the link between hallucination and creativity in LLMs and to explore how this interplay can be effectively harnessed for innovative applications.

THE CREATIVITY OF LARGE LANGUAGE MODELS

Regarding the interplay between hallucinations and creativity in LLM, establishing a definitive concept of creativity becomes essential. However, creativity has long been a challenging concept to both define and quantify due to its multifaceted and intricate nature. In this section, we will explore definitions of creativity as understood in cognitive science, examine the approaches employed to measure creativity, and highlight recent studies that investigate LLM creativity.

Definition of creativity

While one study presents an extensive array of over 100 different definitions of creativity, it is noteworthy that most studies in this field typically utilize only a select few of these definitions.[31] In cognitive science, creativity is often conceptualized from four distinct perspectives: cognitive processes associated with creativity (later in this paper referred to as 'process'), personal characteristics of creative individuals ('person'), creative products or outcomes ('product') and the interaction between the creative individual and the context or environment ('press').[32] In this paper, we mainly introduce the 'process' and 'product', as they are particularly pertinent to understanding the creativity of LLM, while the 'person' and 'press' categories are more aligned with human creativity.

Regarding the process perspective of creativity, one study defines it as the process of identifying problems or gaps in knowledge, developing hypotheses or propositions, testing and validating hypotheses, and ultimately sharing the findings.[33] Similarly, another study suggests that creativity involves combining associative elements into novel configurations that fulfill the requirements of a given task.[34] A further account views creativity as a type of problem-solving and differentiates between two kinds of cognitive operations: divergent and convergent production.[9] Divergent production is characterized by a wide-ranging search for multiple logical solutions or alternatives to open-ended problems, while convergent production involves a narrow search for a single, precise solution to a problem where one specific answer is needed. They posit that divergent production processes are more closely associated with effective creative thinking.

As the definition of creativity towards the product, one study defines creativity as the construction or organization of ideas, thoughts, and feelings into unusual and associative connections through the power of imagination.[35]

Another account posits that creative individuals possess the capacity to solve problems, fashion products, or formulate new questions in ways that are novel yet acceptable within a specific cultural context.[36] Creativity is also perceived as the ability to generate or conceive something original and adaptive to task constraints, while also being of high quality, useful, aesthetically pleasing, and novel.

Researchers often distinguish between two main types of creativity: everyday creativity, or "little-C," common to nearly everyone, and eminent creativity, or "big-C," which is seen in significant historical figures. Expanding on this, the Four C model of creativity introduces "mini-c," representing the creative learning process, and "Pro-c," indicating professional-level creative expertise.[37]

Despite the varied perspectives in defining creativity, a consensus exists among researchers regarding its core attributes. It is widely acknowledged that creativity is characterized by the generation of responses that are both novel and useful. However, the precise interpretation of these terms remains a subject of ongoing discourse.

This consensus is important for interpreting entropy-based accounts of LLM creativity. High predictive entropy can enlarge the search space and may indicate novelty or divergence from the most probable continuation, but it cannot by itself establish usefulness, task appropriateness, semantic coherence, or factual acceptability. Therefore, high predictive entropy is a necessary but not sufficient condition for creativity: it may help generate candidate ideas, but the candidate must still satisfy the product-oriented criteria of usefulness and appropriateness through convergent evaluation.[3,27] This point also prevents random high-entropy outputs from being misclassified as creative merely because they are unlikely.

Approaches for measuring creativity

In the field of cognitive science, measuring creativity is proven to be a challenging task. These challenges stem from the subjectivity of creativity and the variety of environments in which it is manifested. Based on the various definitions of creativity, researchers have developed many different methods to measure it. These methods are divided into four categories, representing the four main dimensions of the creativity definition: process, product, person, and press. Here, we focus on the widely adopted process and product approaches.

The process approaches in creativity measurement focus on the specific cognitive processes and structures that are conducive to creative production. Divergent thinking tests (DAT) have been most widely used for measuring creative processes or creativity-relevant skills.[38] Examples of these tests include the Associative chain test (ACT),[39] the Torrance Tests of Creative Thinking (TTCT),[33] the Structure of the Intellect Divergent Production Tests (SOI),[9] Wallach-Kogan Creativity Tests and the Creativity Assessment Packet (CAP). These tests, which are also known as measures of ideational fluency, include open or ill-structured problems that require individuals to generate as many responses as possible, which are then scored to capture fluency (number of responses), originality (statistical rarity), flexibility (number of different categories) and elaboration (amount of detail). Hence, the focal point of divergent thinking tests is not only to consider the amount of responses but also the quality of these responses. Besides, there are some convergence thinking tasks that consider creative ideas are achieved by forming mutually remote associative elements into new and useful combinations, like Remote Associates Test (RAT),[34] Bridge-the-Associative-Gap Task (BAG).[40] However, the fundamental psychometric assumptions and underlying cognitive processes involved in them remain controversial.

The product approaches are supported by the rationale that a comprehensive assessment of an individual's creative capabilities should include the measurement of their concrete creations. Central to this method is the Consensual Assessment Technique (CAT), which is widely implemented in studies focusing on the products of creativity.[41] Distinct from other methods of assessment, the CAT does not rely on predetermined theoretical frameworks of creativity. Instead, it utilizes the informed judgment of experts to evaluate the creative outputs within their relevant fields. This unique characteristic of the CAT allows it to incorporate a range of perspectives. While the CAT for measuring creativity is known for its high inter-rater reliability, several situational variables like the number of tasks and the performance domain can impact this reliability. Furthermore, the practicality of implementing the CAT faces numerous challenges. The selection and expertise level of judges are subjects of ongoing debate, with the consensus among judges notably swayed by their expertise. Additionally, judges personalities can influence creativity ratings, with more agreeable judges tending towards leniency.[42] Cultural disparities among judges also play a role, as found in a comparison of American and Chinese judges' evaluations of artistic creativity.[43] Judges may also display biases, particularly when evaluating their own work.

Recent works on evaluating LLM's creativity

The section mentioned above delves into the approaches to measure human creativity from a cognitive science perspective. However, since there are differences between LLM and humans, it might lead to irrelevant responses or serious logical issues, requiring us to additionally assess these approaches. This predicament raises a question: *how can we adapt these measures to effectively evaluate the creativity of LLM?* In light of LLM's burgeoning potential, researchers have explored various approaches to measure the creativity of LLM. Broadly, these approaches fall into two distinct types. The first type involves adapting existing creativity assessment techniques from cognitive science for use with LLM. The second is to design a new approach specifically for measuring creativity in LLM.

For the first type, some researchers think that creative problem-solving is a crucial ability for LLM. For example, one study conducts a comparative analysis using the AUT, where both GPT-3 and human subjects were instructed to generate novel and useful responses. The responses were rated on a scale from 1 to 5 by two human judges, leading to the conclusion that humans outperformed GPT-3 in generating creative responses within the AUT.[44] Building upon this, another study devised a series of prompts aimed at sifting through the 690 alternative uses responses previously generated, isolating those that were both original and practical. These prompts required the evaluation of the pros and cons associated with employing the objects in their new, unconventional roles. While GPT-3 demonstrated an ability to generate responses that were "surprisingly good," it notably failed to discern and discard impractical alternative uses.[45] One study explores the creativity of LLM by using the DAT task on GPT-4 and GPT-3.5 compared to the human norms.[46] In contrast to these approaches, another study proposes an innovative, interactive method enabling GPT-4 to autonomously refine and enhance the creativity of its outputs.[47] They employ the AUT and the TTCT visual completion tasks to investigate the LLM's creativity. Similarly, another study also uses TTCT to evaluate the creative abilities of GPT-4.[48] Furthermore, a landmark study identifies a homogeneity paradox across frontier LLMs, including GPT-4, Claude, and Gemini. By evaluating population-level response diversity, researchers found that different models generate highly similar creative outputs compared to humans, warning that over-reliance on LLMs as brainstorming partners may collectively narrow human creative diversity.[49]

For the second type, one study proves in theory that LLM can be as creative as humans under the condition that it can properly fit the data generated by human creators.[50] Additionally, the study also introduces two concepts of creativity in LLM: relative and statistical creativity. Relative creativity deems an LLM as creative as a hypothetical yet realistic human creator if an evaluator cannot distinguish the LLM's works from those of that creator. Statistical creativity is a means for understanding whether and to what degree LLM achieves creativity by comparing it with existing human creators. Using a different mathematical perspective, another study derives a mathematical characterization of the trade-off between hallucination and creativity in LLM by developing a rigorous mathematical framework for analyzing the hallucination phenomenon in LLMs.[27]

Although researching creativity in LLMs become a trend, there is still a need for more theoretical studies and comprehensive benchmarks to deepen understanding of this area.

HARNESSING LARGE LANGUAGE MODEL HALLUCINATIONS FOR CREATIVITY

Given the recognition of LLM hallucinations as potential catalysts for creativity, a pivotal question arises: *How can we effectively harness hallucinations for creative thought while also mitigating their risks?* The aim is to

Table 2. Comparison of divergent and convergent phases for harnessing LLM hallucinations.

Phase	Goal	Representative methods	Main risk	Verification mechanism
Divergent phase	Expand the candidate space and encourage non-obvious associations	Prompting, controlled sampling, layer steering, multi-agent debate, human-agent co-creation.[11,30]	Unstructured noise, false facts, prompt drift, unsafe associations	Diversity and coherence checks; task-intent recognition; user steering
Convergent phase	Select candidates that are useful, grounded, and appropriate	Retrieval or KG grounding, Chain-of-Verification, SelfCheckGPT, expert review, human preference evaluation.[12,13,19,21]	Over-filtering creative ideas or accepting plausible falsehoods	Grounding, independent fact checking, consistency sampling, domain expert assessment

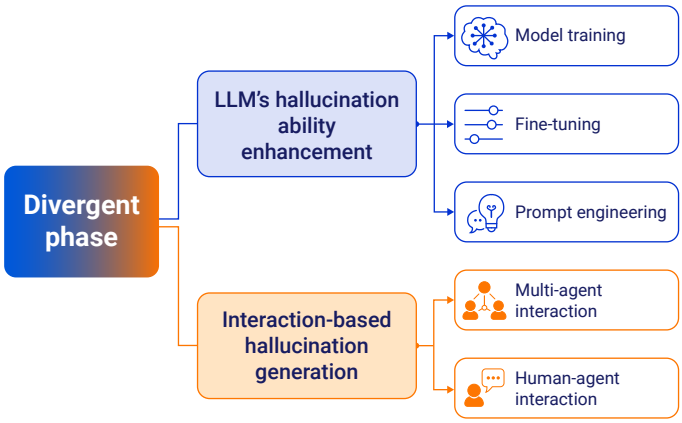


Figure 3. The divergent phase stimulates the hallucination of LLMs, fostering creative thought.

enhance our understanding and utilization of these phenomena, thereby amplifying the creative potential of LLMs across various applications, without compromising on accuracy. This requires methods that not only induce hallucinations but also critically evaluate them, transforming them into innovative sources. In addressing this challenge, our research revolves around a comprehensive consideration of two critical dimensions: the divergent phase and the convergent phase. These phases align with the framework outlined in fundamental research, which posits creativity as an interplay of generative and evaluative processes, embracing both divergent and convergent thinking.[10] The divergent phase stimulates LLM hallucinations tailored to user needs, fostering creative thought, while the convergent phase assesses these to transform them into innovative, applicable ideas, and our survey will systematically explore studies aligned with these phases.

Table 2 summarizes the basis for this two-phase organization. The classification standard follows classical divergent-convergent creativity theory;[9,10] the method categories are derived from recent LLM studies on multi-agent divergent generation, retrieval or knowledge-graph grounding, verification, self-checking, and factuality-oriented decoding.[11-13,19,21,51] We organize the table by goal, method, risk, and verification mechanism because these dimensions determine whether a hallucination-derived output remains an unsafe error or becomes a candidate for creative use.

These examples also highlight a trade-off: methods that improve factuality, such as retrieval-augmented generation, Chain-of-Verification, SelfCheckGPT, and layer-contrast decoding, can serve as convergent filters, but they should be tuned carefully so that they do not collapse useful divergent candidates into overly conservative outputs.[12,13,19,21,51]

Divergent phase of LLM

In the divergent phase, the emphasis extends beyond the mere generation of relevant hallucinations. It encompasses fostering the LLM's capacity for hallucination with creativity. This phase benefits significantly from insights derived from existing literature, which suggest a multi-faceted approach to model optimization, as shown in Figure 3. The following discussion proceeds through model training, inference-time steering, prompt engineering, and interaction.

For the hallucination ability enhancement of LLMs, Pressing's theory of improvisational creativity suggests that improvisation is an acquired skill, requiring extensive training to reach professionals, which literature underscores the importance of training.[52] In concordance with the aforementioned views, recent studies have proven the hallucination gap between different LLMs through extensive experiments, which points out the necessity of model training to improve the LLM's abilities of hallucination and creativity.[3,53] For instance, research demonstrates how specific training techniques can enhance the model's ability to generate creative outputs. This research proposes that the cognitive process of creativity, especially the 'insight' phenomenon in problem-solving, aligns closely with the mechanisms of hierarchical reinforcement learning (HRL) in artificial systems. This work underscores the potential of HRL to mimic human-like creative processes, offering a novel perspective on the capabilities of LLM in replicating complex cognitive behaviors.[54] At the inference stage, micro-level architectural modulation offers a precise mechanical path to steer creative hallucinations. For instance, recent quantitative insights on decoding layers introduce the hallucination and creativity across layers framework to analyze the intrinsic trade-off between hallucination and creativity, revealing an optimal layer that acts as a lever to balance generation accuracy and diversity.[30] Extending this layer-wise modulation to multimodal models, the FlexAC framework unifies hallucination and creativity as facets of associative reasoning, employing middle-layer steering vectors to dynamically tune associative intensity without retraining.[55] For prompt engineering, by setting specific prompts and contexts, LLMs will continuously learn and adapt in contexts to enhance their understanding and response capabilities for hallucination generation and complex creative tasks. This theory can be traced back to the classic literature of cognitive science. For example, giving a prompt like: "*prompt: to come up with something clever, humorous, original, compelling, or interesting*" has been validated as effective for human hallucination in research.[56] In the prompt engineering research about LLM, it is also the most widely adopted strategy for divergent thinking, enabling LLMs to explore and respond to users' creative requirements more easily, thereby exhibiting a higher level of creativity in applications.[57,58]

The integration of interaction-based hallucination generation methods also offers a promising direction to enrich the divergent phase. This concept introduces two distinct types of interactions: multi-agent interaction and human-agent interaction. For the interaction among multiple agents, a dynamic that has been investigated in recent studies plays a crucial role.[11] These multi-agent interactions encourage diverse and creative outputs from LLMs, as agents with varying perspectives and knowledge bases engage in collaborative or competitive dialogues. For the human-agent interaction, an area receiving increasing attention provides invaluable insights.[59,60] This form of interaction allows for a more grounded and realistic feedback loop, where human creativity can inspire and be inspired by the LLM's responses, leading to more relatable and applicable hallucinations. In summary, these two types of interactions underpin the diversification of hallucination generation in LLMs, each contributing to a more dynamic, context-aware, and creatively stimulating phase.

As examined in prior work, the divergent thinking ability of LLM is still far from human, which requires the improving of LLM divergent thinking abilities.[58]

Convergent phase of LLM

The convergent phase, in stark contrast to the divergent phase, is dedicated to refining hallucinations into valuable creative contributions. This phase is characterized by a meticulous process of evaluating hallucinations, employing a systematic approach to identify, categorize, and select those that are most constructive and aligned with user needs.

Different from existing hallucination detection and reduction methods as discussed in Hallucination Reduction and Hallucination Detection, which

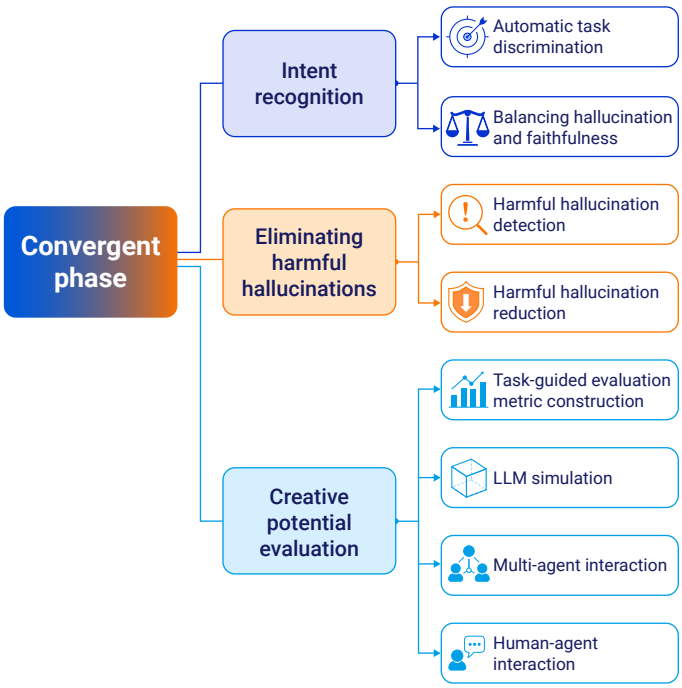


Figure 4. The convergent phase refines hallucinations into valuable creative contributions.

directly identify and eliminate hallucinations, the convergent phase involves discerning whether hallucinations are needed based on the context, eliminating harmful hallucinations, and identifying valuable ones, raising higher demands for LLM convergent thinking abilities, as shown in Figure 4.

To determine whether hallucinations are necessary for applications, it's crucial to optimize the intent recognition capabilities of LLMs. For instance, the 4C creativity theory discussed in research explores four scenarios requiring creativity, optimizing Little-C tasks for LLMs' creativity.[53] Automating the discrimination of these task scenarios remains an area of urgent research. Moreover, balancing hallucination and faithfulness in LLMs based on the demand, as mentioned in research, involves fine-tuning and knowledge fusion, but this area is still in its infancy, and lacking a systematic approach.[61] To bypass this bottleneck, recent empirical investigation reveals that standard mitigation techniques impose highly differentiated side effects on creative expressions across various model scales. Specifically, while the chain of verification (CoVe) is proven to mitigate factual errors while surprisingly reinforcing divergent thinking, approaches like decoding by contrasting layers (DoLa) tend to heavily suppress the model's creative ideation.[62]

Table 3 extends this argument to application domains. Its scenarios are selected from major use cases repeatedly discussed in LLM creativity, problem-solving, science communication, drug discovery, and responsible AI literature.[3,22,23,53,63,64] The classification standard is practical risk tolerance: low-risk ideation can treat hallucination-derived outputs as candidates, whereas high-stakes factual tasks require strict grounding and expert validation.

To discern harmful hallucinations, a detailed and nuanced classification system for hallucinations is crucial. Existing research, as extensively discussed in Hallucination Detection, has developed a taxonomy for hallucinations predominantly under the assumption that they are detrimental and need to be mitigated. However, this approach overlooks the potential value of certain hallucinations. The current classification lacks in annotating and recognizing beneficial hallucinations, thereby limiting the LLMs' ability to differentiate between harmful and potentially valuable hallucinations.

To evaluate the creative potential hidden within the sea of hallucinations produced by LLMs, it's crucial to establish metrics and design methods for effectively assessing LLMs' creative output. Evaluation metrics specify what should be measured. As discussed in Recent works on evaluating LLM's creativity, the evaluation of creativity within these hallucinations can leverage existing metrics from cognitive science and those specifically developed for LLMs. The choice of metrics should align with the specific applications. For instance, artistic creation often emphasizes fluency, flexibility, originality, and elaboration,[65] whereas scientific innovation is more focused on problem-solving abilities.[53] To extend evaluation into science communication, the StoryScore framework distinguishes pedagogical creativity from genuine factual errors, preventing conventional metrics from mispenalizing constructive narrative adaptations.[63] A prominent empirical case is found in early-stage drug discovery, where seemingly erroneous hallucinated molecular descriptions are systematically harnessed as speculative counterfactual reasoning clues to significantly boost the accuracy of downstream attribute predictions.[64] Evaluation methods specify how the assessment is performed and by whom. Those mentioned in Recent works on evaluating LLM's creativity for assessing creativity are still applicable to hallucinations. However, these studies often rely on manual assessment by data annotators or experts. Automating the assessment of creativity within hallucinations is key to transitioning from theory to practical applications. For model-based evaluation, recent studies propose self-reflection for automatic content evaluation and correction but highly rely on the model's reasoning abilities, which requires effective training methods or prompt strategies in future research.[3,53] For multi-agent evaluation, studies like ChatEval attempt to use multi-agent approaches to assess whether generated content meets requirements.[66] Research evaluates the creative value behind LLM actions from a simulation perspective.[67] For human-agent evaluation, research assessing user interactions with LLM-generated content offers a way to evaluate creativity in hallucinations, moving away from traditional manual annotation.[60] The latest studies, by introducing dynamic reasoning chains and self-evolving evaluation frameworks, more accurately capture the breakthrough ideas generated by models at the edge of extreme hallucination, further validating the automated potential of agent collaboration in the convergent phase.[68,69]

Multimodal scenarios: Text-to-image and text-to-video

The distinction between harmful hallucination and creative deviation becomes more complex in multimodal generation. In text-to-image systems, hallucination may appear as missing objects, incorrect attributes, wrong counting, broken spatial relations, or weak prompt-image faithfulness. These errors should not be treated as creative merely because the image is visually novel. Recent benchmarks and analyses such as TIFA, I-HallA, T2I-CompBench++, and VQAScore evaluate faithfulness through question answering, compositional reasoning, image-hallucination patterns, and visual question answering signals.[70-73] Therefore, a text-to-image deviation can be considered creatively useful only when it preserves task-relevant entities and relations or when the user explicitly requests reinterpretation rather than faithful rendering.

Text-to-video generation adds temporal constraints. A visually interesting clip can still fail because of subject identity drift, incorrect action binding, unstable object interactions, temporal incoherence, poor motion smoothness, or temporal flickering. Benchmarks such as EvalCrafter, VBench, VBench-2.0, and VideoGen-RewardBench evaluate these dimensions from both quality and faithfulness perspectives.[74-77] For creative use, these metrics suggest a convergent rule: multimodal hallucination-derived candidates should be evaluated not only by novelty, but also by cross-modal alignment, temporal consistency, and user intent. This multimodal perspective also explains why the same deviation can be acceptable in artistic exploration but unacceptable in instruction-following image or video generation.

Responsible use boundaries. The distinction between productive and creative hallucination must be visible to users. Systems should make clear whether they are operating in a factual mode, where unsupported claims are errors to be minimized, or a creative mode, where outputs are exploratory candidates. Even in creative mode, hallucination-derived outputs should be labeled as candidates rather than verified facts. They should not be used directly for legal, medical, financial, or safety-critical decisions without grounding, verification, and appropriate expert review.[22,23]

In conclusion, while many studies have explored this direction, research in this field is still in its early stages. A more comprehensive evaluation metric system, richer datasets, and stronger models and framework designs are urgently needed for further exploration.

Table 3. Domain-specific handling of hallucination-derived candidates.

Scenario	Potentially acceptable role of hallucination	Main risk	Filtering or validation strategy
Creative writing and brainstorming	Associative prompts, alternative framings, narrative variation.[3]	Cliche, incoherence, harmful stereotypes, misleading claims presented as fact	Human selection, coherence checks, style constraints, disclosure that outputs are candidates
Education and science communication	Analogies and simplified explanations that improve engagement.[63]	Pedagogical distortion or false simplification	Expert review, retrieval grounding, question-answer verification
Scientific ideation	Speculative hypotheses, counterfactual prompts, search-space expansion.[53]	False mechanisms or unsupported claims	Literature retrieval, domain expert screening, experiment or simulation before adoption
Drug discovery and design ideation	Unusual molecular descriptions or speculative candidates for downstream testing.[64]	Invalid chemistry, unsafe recommendations, overclaiming	Rule-based constraints, model-based property prediction, laboratory or simulation validation
High-stakes factual tasks	Generally unacceptable except as flagged hypotheses for expert review.[22,23]	Legal, medical, financial, or safety harm	Strict grounding, Chain-of-Verification, SelfCheckGPT-style consistency checks, external expert approval

CONCLUSION AND FUTURE WORK

Conclusion

This paper revisits the hallucinations in LLMs, advocating for a paradigm shift where they are not solely viewed as negative phenomena, but also as potential catalysts for creativity. It establishes what hallucinations entail within LLMs and reviews research focused on their detection and reduction. Challenging the view of hallucinations as entirely negative, the study delves into their potential to spur creativity. Examining historical cases and current research highlights the intricate interplay between hallucination and creativity in LLMs. The discussion follows the cognitive science framework of divergent and convergent thinking, providing a comprehensive review of studies on harnessing creative hallucinations. This work not only aligns with cognitive science principles but also opens avenues for applications and future research in leveraging hallucinations for creative purposes in LLMs.

Future work

Future research in the field of hallucinations and creativity within LLMs can be advanced through several concrete directions:

Deeper Theoretical Exploration: Future work should ask how uncertainty, semantic coherence, grounding, and usefulness jointly determine whether a hallucination-derived output can become creative. A key research question is how to model the transition from high-entropy candidate generation to validated creative output without equating entropy with creativity.

Richer Datasets and Benchmarks: The field needs benchmarks that annotate not only whether an output is hallucinated, but also whether it is novel, useful, harmful, recoverable through verification, or valuable only as an ideation prompt. Such benchmarks should separate factual error detection from creative value assessment and should report whether scores remain comparable across tokenizers, model families, and domains.

Optimization of Method Designs: Future systems should support mode-aware generation and evaluation. One research direction is to design interfaces and system prompts that explicitly distinguish factual mode from creative mode. Another is to combine divergent generation methods with convergent filters so that creative candidates are explored without weakening factual reliability.

Transformation of Hallucinations to Creativity in Multimodal Scenarios: Building on the multimodal benchmarks reviewed in Harnessing LLM Hallucinations for Creativity, future work should ask how to separate creative reinterpretation from prompt-faithfulness failure in text-to-image and text-to-video generation. A further question is how user intent, temporal consistency, and cross-modal alignment should be jointly modeled when deciding whether a deviation is useful or erroneous.

Exploration of More Application Scenarios: Future studies should compare domains in which hallucination-derived candidates may be useful, such as creative writing, design, education, scientific ideation, and early-stage discovery, against domains where hallucinations are primarily hazards. For each domain, the central research question is not whether hallucination is good or bad in general, but what constraints, disclosures, and validation mechanisms are required before a hallucination-derived candidate can be used responsibly.

Ultimately, these directions for future research endeavor to enrich the comprehension and application of hallucinations in LLMs, contributing innovative theories and methodologies and paving the way for practical applications.

DECLARATION OF GENERATIVE AI AND AI-ASSISTED TECHNOLOGIES IN THE WRITING PROCESS

During the preparation of this work, the authors used OpenAI Codex to improve language and readability and to transfer the manuscript into the journal's LaTeX template. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

REFERENCES

1. Ye H., Liu T., Zhang A., et al. (2023). Cognitive mirage: A review of hallucinations in large language models. *arXiv preprint arXiv: 2309.06794.*
2. Sui P., Duede E., Wu S., et al. (2024). Confabulation: The Surprising Value of Large Language Model Hallucinations. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* 14274–14284. DOI:10.18653/v1/2024.acl-long.770
3. Zhao Y., Zhang R., Li W., et al. (2025). Assessing and understanding creativity in large language models. *Mach. Intell. Res.* **22**:417–436. DOI:10.1007/s11633-025-1546-4
4. Touvron H., Lavril T., Izacard G., et al. (2023). Llama: Open and efficient foundation language models. *arXiv preprint arXiv: 2302.13971.*
5. Ngo R., Chan L. and Mindermann S. (2024). The alignment problem from a deep learning perspective. *International Conference on Learning Representations* **2024**:7474–7501.
6. Ji Z., Lee N., Frieske R., et al. (2023). Survey of hallucination in natural language generation. *ACM Comput. Surv.* **55**: 1–38. DOI:10.1145/3571730
7. Dziri N., Madotto A., Zaïane O., et al. (2021). Neural Path Hunter: Reducing Hallucination in Dialogue Systems via Path Grounding. *Proc. of EMNLP.*
8. Huang L., Yu W., Ma W., et al. (2025). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.* **43**:1–55. DOI:10.1145/3703155
9. Guilford J. P. (2017). Creativity: A quarter century of progress. *Perspectives in creativity* 37–59.
10. Pressing J. (1998). Psychological constraints on improvisational expertise and communication. *In the course of performance: Studies in the world of musical improvisation.*
11. Liang T., He Z., Jiao W., et al. (2024). Encouraging divergent thinking in large language models through multi-agent debate. *Proceedings of the 2024 conference on empirical methods in natural language processing* 17889–17904.
12. Dhuliawala S., Komeili M., Xu J., et al. (2024). Chain-of-Verification Reduces Hallucination in Large Language Models. *Findings of the Association for Computational Linguistics: ACL 2024* 3563–3578. DOI:10.18653/v1/2024.findings-acl.212
13. Manakul P., Liusie A. and Gales M. J. F. (2023). SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* 9004–9017. DOI:10.18653/v1/2023.emnlp-main.557
14. Rawte V., Sheth A. and Das A. (2023). A survey of hallucination in large foundation

models. *ArXiv preprint.*
15. Li J., Cheng X., Zhao W. X., et al. (2023). Halueval: A large-scale hallucination evaluation benchmark for large language models. *Proc. of EMNLP.*
16. Lin S., Hilton J. and Evans O. (2022). TruthfulQA: Measuring How Models Mimic Human Falsehoods. *Proc. of ACL.*
17. Pal A., Umapathi L. K. and Sankarasubbu M. (2023). Med-halt: Medical domain hallucination test for large language models. *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)* 314–334.
18. Tonmoy S., Zaman S., Jain V., et al. (2024). A comprehensive survey of hallucination mitigation techniques in large language models. *arXiv preprint arXiv: 2401.01313.*
19. Agrawal G., Kumarage T., Alghamdi Z., et al. (2024). Can knowledge graphs reduce hallucinations in llms?: A survey. *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)* 3947–3960.
20. Shi X., Zhu Z., Zhang Z., et al. (2023). Hallucination Mitigation in Natural Language Generation from Large-Scale Open-Domain Knowledge Graphs. *Proc. of EMNLP.*
21. Sui Y., He Y., Ding Z., et al. (2025). Can knowledge graphs make large language models more trustworthy? an empirical study over open-ended question answering. *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* 12685–12701.
22. Zhang Y., Li Y., Cui L., et al. (2025). Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models. *Comput. Linguist.* **51**:1373–1418. DOI:10.1162/COLI.a.16
23. National Institute of Standards and Technology (2024). Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile. DOI: 10.6028/NIST.AI.600-1
24. Benedek M., Jauk E., Fink A., et al. (2014). To create or to recall? Neural mechanisms underlying the generation of creative new ideas. *Neuroimage* **88**:125–133. DOI:10.1016/j.neuroimage.2013.11.021
25. Beaty R. E. (2015). The neuroscience of musical improvisation. *Neurosci. Biobehav. Rev.* **51**:108–117. DOI:10.1016/j.neubiorev.2015.01.004
26. Salesforce (2023). A world without AI is becoming unthinkable.
27. Lee, M. (2023). A mathematical investigation of hallucination and creativity in gpt models. *Mathematics* **11**: 2320. DOI:10.3390/math11102320
28. Bostrom K. and Durrett G. (2020). Byte Pair Encoding is Suboptimal for Language Model Pretraining. *Findings of the Association for Computational Linguistics: EMNLP 2020* 4617–4624. DOI:10.18653/v1/2020.findings-emnlp.414
29. Wang F. (2024). LightHouse: A Survey of AGI Hallucination. *arXiv preprint arXiv: 2401.06792.*
30. He Z., Zhang B. and Cheng L. (2025). Shakespearean Sparks: The Dance of Hallucination and Creativity in LLMs' Decoding Layers. *arXiv preprint arXiv: 2503.02851.*
31. Treffinger D. J. (1998). Creativity, Creative Thinking, and Critical Thinking: In Search of Definitions. *Gifted and Talented International.*
32. Couger J. D., Higgins L. F. and McIntyre S. C. (1993). (Un)Structured Creativity in Information Systems Organizations. *MIS Q.*
33. Torrance E. P. (1977). Creativity in the Classroom; What Research Says to the Teacher.
34. Mednick S. A. (1962). The associative basis of the creative process. *Psychol. Rev.* **69**: 230-232. DOI:10.1037/h0048850
35. Khatena J. and Torrance E. P. (1973). Thinking creatively with sounds and words: Normstechnical manual. *Res. ed.) Bensenville, IL: Scholastic Testing Service.*
36. Gardner H. (2011). Creating minds: An anatomy of creativity seen through the lives of Freud, Einstein, Picasso, Stravinsky, Eliot, Graham, and Ghandi. (*Civitas books*).
37. Kaufman J. C. and Beghetto R. A. (2009). Beyond Big and Little: The Four C Model of Creativity. *Rev. Gen. Psychol.* **13**: 1-12. DOI:10.1037/a0013688
38. Olson J. A., Nahas J., Chmoulevitch D., et al. (2021). Naming unrelated words predicts creativity. *Proc. Natl. Acad. Sci. U.S.A.* **118**: e2022340118. DOI:10.1073/pnas.2022340118
39. Marko M., Michalko D. and Riečanský I. (2018). Remote associates test: An empirical proof of concept. *Behav. Res. Methods.* **51**:2700-2711. DOI:10.3758/s13428-018-1131-7
40. Gianotti L. R. R., Mohr C., Pizzagalli D. A., et al. (2001). Associative processing and paranormal belief. *Psychiatry Clin. Neurosci.* **55**:595-603. DOI:10.1046/j.1440-1819.2001.00911.x
41. Amabile T. M. (1982). Social psychology of creativity: A consensual assessment technique. *J. Pers. Soc. Psychol.* **43**:997–1013. DOI:10.1037/0022-3514.43.5.997
42. Baer J. (2015). The Importance of Domain-Specific Expertise in Creativity. *Roeper Rev.* **37**:165–178. DOI:10.1080/02783193.2015.1047480
43. Niu W. and Sternberg R. J. (2001). Cultural influences on artistic creativity and its evaluation. *Int. J. Psychol.* **36**:225-241. DOI:10.1080/00207590143000036
44. Stevenson C., Smal I., Baas M., et al. (2022). Putting GPT-3's Creativity to the Alternative Uses Test.
45. Summers-Stay D., Voss C. R. and Lukin S. M. (2023). Brainstorm, then Select: a Generative Language Model Improves Its Creativity Score. *The AAAI-23 Workshop on Creative AI Across Modalities.*
46. Cropley D. (2023). Is artificial intelligence more creative than humans? : ChatGPT and the Divergent Association Task. *Learn. Lett.* **2**:13. DOI:10.59453/ll.v2.13
47. Góes L. F., Volpe M., Sawicki P., et al. (2023). Pushing gpt's creativity to its limits: Alternative uses and torrance tests.
48. Guzik E. E., Byrge C. and Gilde C. (2023). The originality of machines: AI takes the Torrance Test. *J. Creat.* **33**:100065. DOI:10.1016/j.yjoc.2023.100065
49. Wenger E. and Kenett Y. N. (2026). Large language models are homogeneously creative. *PNAS Nexus* **5**:pgag042. DOI:10.1093/pnasnexus/pgag042
50. Wang H., Zou J., Mozer M., et al. (2024). Can AI be as creative as humans? *arXiv preprint arXiv: 2401.01623.*
51. Chuang Y. S., Xie Y., Luo H., et al. (2024). DoLa: Decoding by Contrasting Layers Improves Factuality in Large Language Models. *International Conference on Learning Representations.*
52. Pressing J. (2007). Improvisation: Methods and models. *Physical Theatres: A Critical Reader* 66–78.
53. Tian Y., Ravichander A., Qin L., et al. (2024). MacGyver: Are Large Language Models Creative Problem Solvers? *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)* 5303–5324.
54. Colin T. R., Belpaeme T., Cangelosi A., et al. (2016). Hierarchical reinforcement learning as creative problem solving. *Robot. Auton. Syst.* **86**:196-206. DOI:10.1016/j.robot.2016.08.021
55. Yuan S., Lyu X., Wang S., et al. (2025). FlexAC: Towards Flexible Control of Associative Reasoning in Multimodal Large Language Models. *Advances in Neural Information Processing Systems.*
56. Beaty R., Smeekens B., Silvia P., et al. (2013). A First Look at the Role of Domain-General Cognitive and Creative Abilities in Jazz Improvisation. *Psychomusicol. Music Mind Brain.* **23**:262–268. DOI:10.1037/a0034968
57. Hubert K., Awa K. and Zabelina D. (2023). Artificial intelligence is more creative than humans: A cognitive science perspective on the current state of generative language models.
58. Koivisto M. and Grassini S. (2023). Best humans still outperform artificial intelligence in a creative divergent thinking task. *Sci. Rep.* **13**:13601. DOI:10.1038/s41598-023-40858-3
59. Weber S., Kordyaka B., Palombo R., et al. (2023). Is a Fool With a (n AI) Tool Still a Fool? An Empirical Study of the Creative Quality of Human–AI Collaboration. *ACIS 2023 Proceedings.*
60. Rick S. R., Giacomelli G., Wen H., et al. (2023). Supermind Ideator: Exploring generative AI to support creative problem-solving. *ArXiv preprint.*
61. Zhang C. (2023). User-controlled knowledge fusion in large language models: Balancing creativity and hallucination. *arXiv preprint arXiv: 2307.16139.*
62. Banerjee M., Wangsajaya N. Y., Alsagoff S. A. R., et al. (2025). Does Less Hallucination Mean Less Creativity? An Empirical Investigation in LLMs. *arXiv preprint arXiv: 2512.11509.*
63. Argese A., Lisena P. and Troncy R. (2026). Hallucination or Creativity: How to Evaluate AI-Generated Scientific Stories? *arXiv preprint arXiv: 2602.02290.*
64. Yuan S., Qu Z., Kangen A. Y., et al. (2025). Can Hallucinations Help? Boosting LLMs for Drug Discovery. *arXiv preprint arXiv: 2501.13824.*
65. Silvia P. J., Winterstein B., Willse J. T., et al. (2008). Assessing creativity with divergent thinking tasks: exploring the reliability and validity of new subjective scoring methods. *Psychol. Aesthet. Creat. Arts.* **2**:68–85. DOI:10.1037/1931-3896.2.2.68
66. Chan C. M., Chen W., Su Y., et al. (2024). Chateval: Towards better llm-based evaluators through multi-agent debate. *International conference on learning representations* **2024**:9079–9093.
67. Zhang L., Zhang M., Wang W. L., et al. (2025). Simulation as Reality? The Effectiveness of LLM-Generated Data in Open-ended Question Assessment. *arXiv preprint arXiv: 2502.06371.*
68. Karpowicz M. and others (2025). On the Fundamental Impossibility of Hallucination Control in Large Language Models. *arXiv preprint arXiv: 2506.06382.*
69. Qiu Z. and Hu R. (2025). Deep Associations, High Creativity: A Simple yet Effective Metric for Evaluating Large Language Models. *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing* 10870–10883.
70. Hu Y., Liu B., Kasai J., et al. (2023). TIFA: Accurate and Interpretable Text-to-Image Faithfulness Evaluation with Question Answering. *Proceedings of the IEEE/CVF International Conference on Computer Vision* 20406–20417.
71. Lim Y., Choi H. and Shim H. (2025). Evaluating Image Hallucination in Text-to-Image Generation with Question-Answering. *Proceedings of the AAAI Conference on Artificial Intelligence* **39**:26290–26298. DOI:10.1609/aaai.v39i25.34827
72. Huang K., Duan C., Sun K., et al. (2025). T2I-CompBench++: An Enhanced and Comprehensive Benchmark for Compositional Text-to-Image Generation. *IEEE Trans. Pattern Anal. Mach. Intell.* **47**:3563–3579. DOI:10.1109/TPAMI.2025.3531907
73. Lin Z., Pathak D., Li B., et al. (2024). Evaluating Text-to-Visual Generation with Image-to-Text Generation. *European Conference on Computer Vision.*
74. Liu Y., Cun X., Liu X., et al. (2024). EvalCrafter: Benchmarking and Evaluating Large Video Generation Models. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 22139–22149.
75. Huang Z., He Y., Yu J., et al. (2024). VBench: Comprehensive Benchmark Suite for Video Generative Models. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 21807–21818.
76. Zheng D., Huang Z., Liu H., et al. (2025). VBench-2.0: Advancing Video Generation

Benchmark Suite for Intrinsic Faithfulness. *arXiv preprint arXiv: 2503.21755.*
77. Liu J., Liu G., Liang J., et al. (2025). Improving Video Generation with Human Feedback. *Advances in Neural Information Processing Systems.*

FUNDING AND ACKNOWLEDGMENTS

Funding informations: This work was supported by the Henan provincial key research and development program (No. 241111211900) and the Major Public Welfare Project of Henan Province (No.201300311200). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

AUTHOR CONTRIBUTIONS

X.J. conceived and designed the review and drafted the manuscript. Y.L. conducted the literature search and organized the reviewed evidence. Y.S. developed the evaluation framework and contributed to the initial draft. C.X. examined hallucination-detection and mitigation methods and revised the manuscript. Y.T. analyzed divergent and convergent creativity methods and interpreted the findings. J.C. screened and categorized relevant studies and edited the manuscript. X.Z. reviewed multimodal generation applications and benchmarks. Yh.L. curated the references and contributed to the hallucination taxonomy. F.H. analyzed responsible-use, safety, and ethical considerations. B.Z. checked the evidence, citations, and reference list and contributed to manuscript revision. Y.W. supervised the study, refined its conceptual framework, and critically revised the manuscript. J.G. supervised the project, coordinated the work, and critically revised the manuscript. All authors contributed to the manuscript and approved the final version.

DECLARATION OF INTERESTS

Yuanzhuo Wang is an Editorial Board member of *AI Plus* and was blinded from reviewing or making final decisions on the manuscript. Peer review was handled independently of this member and their research group. The other authors declare no conflicts of interest.

ETHICAL STATEMENT AND PATIENT CONSENT

This manuscript is a review article and does not involve human participants, animal experiments, or identifiable personal data. Therefore, ethical approval and informed consent were not required.

DATA AND CODE AVAILABILITY

No new data or code were generated for this review article. All content is based on publicly available sources cited in the manuscript.