The Hallucination Challenge in Large Model Agents: Causes, Risks, and Mitigation - Post-print

Authors: Xu Qi, Sun Zhipu

Date: 2025-07-09T00:00:00+00:00

Abstract

Abstract

Objective: The hallucination and risk issues in large model agents are becoming increasingly prominent, and an in-depth analysis of their causes, risk manifestations, and countermeasures holds important theoretical and practical significance.

Method: Addressing the theoretical and practical needs of the journalism and communication field, this study is primarily based on interdisciplinary literature research and theoretical analysis.

Results: Agent hallucination refers to a series of unavoidable errors at the model generation level, where generated content is illogical or unfaithful to the provided source content. It is mainly divided into two categories: factual hallucination and faithfulness hallucination. The former includes factual errors, fabrication, and omission, while the latter encompasses inconsistencies in intent, context, and logic. In downstream applications, hallucination risks widely exist in tasks such as machine translation, question-answering systems, dialogue, summarization, knowledge graphs, and visual question answering, manifesting as translation deviation, incomplete responses, information distortion, etc., thereby jeopardizing content authenticity and accuracy.

Conclusion: To address the hallucination challenge, the media industry must first strengthen risk awareness and technical literacy at the cognitive level. Technically, Retrieval-Augmented Generation and factual decoding strategies can be adopted. In terms of workflow, human-machine collaboration workflows should be improved, and verification and multi-dimensional evaluation systems enhanced to balance agent effectiveness and reliability.

Full Text

Preamble

The Hallucination Problem in Large Model Agents: Causes, Risks, and Countermeasures

XU Qi, SUN Zhipu
(State Key Laboratory of Media Convergence and Communication, New Media Research Institute, Communication University of China, Beijing 100024)

Abstract

[**Purpose**] The hallucination problem and associated risks in large model agents are becoming increasingly prominent, making it theoretically and practically significant to thoroughly analyze their causes, risk manifestations, and countermeasures. [**Method**] Addressing theoretical and application needs in the field of journalism and communication, this study is primarily based on interdisciplinary literature review and theoretical analysis. [**Results**] Agent hallucination refers to a series of inevitable errors at the model layer where generated content becomes illogical or unfaithful to the provided source material. These errors are mainly categorized into factual hallucinations and faithfulness hallucinations. The former includes factual errors, fabrication, and omission, while the latter encompasses inconsistencies in intent, context, and logic. In downstream applications, hallucination risks are pervasive across tasks such as machine translation, question answering systems, dialogue systems, summarization, knowledge graphs, and visual question answering, manifesting as translation deviations, incomplete responses, information distortion, and other issues that jeopardize content authenticity and accuracy. [**Conclusion**] To address the hallucination challenge, the media industry must first strengthen risk awareness and technological literacy at the cognitive level. Technically, retrieval-augmented generation and factual decoding strategies can be employed, while procedurally, human-machine collaboration workflows should be improved with enhanced verification and multi-dimensional evaluation systems to balance agent effectiveness and reliability.

Keywords: Agent; Large Model; Hallucination; Intelligent Media; Intelligent Communication
Classification Code: G222
Document Code: A
Article ID: 1671-0134(2025)05-07-08
DOI: 10.19483/j.cnki.11-4653/n.2025.05.001
Citation Format: XU Qi, SUN Zhipu. The Hallucination Problem in Large Model Agents: Causes, Risks, and Countermeasures[J]. China Media Technology, 2025, 32(5): 7-14.

1. Problem Statement: The "Hallucination" of the Agent Brain

When asked to explain the detour problem at Xi' an' s Anding Gate, the high-performance DeepSeek-R1 model cited a fabricated concept of "silent zones" from the *Xi' an Historical and Cultural City Protection Plan* and even invented a vibration control standard (GB/T 5845-2019) [**?**]. This phenomenon—where large models generate text that is illogical or unfaithful to the provided source content—is termed "hallucination," a persistent problem in large models [**?**]. In Vectara' s HHEM test, DeepSeek-V3 exhibited a hallucination rate of 3.9%, while DeepSeek-R1' s rate reached 14.3% [**?**]. Although large models represent a leap forward in natural language capabilities, testers have noted that they produce more detailed errors in practical use [**?**]. Behind the enhanced fluency and logical reasoning of generated text lies a more concealed cognitive risk: a shift from easily detectable commonsense errors and fabricated citations to more subtle forms such as fictional technical terminology, forged document numbers, and cross-disciplinary knowledge pastiche.

For the media industry, most hallucination research has focused on peripheral issues such as cognitive risks [**?**], cultural consequences [**?**], and misinformation governance [**?**] triggered by hallucinations, while lacking deep understanding of the problem' s essence. This study introduces the computer science perspective on hallucination to provide theoretical advancement in the epistemology of hallucination, offering a more comprehensive scientific understanding of the problem for future research and further recommendations on how media should understand and address hallucination issues in practice.

2.1 Architecture Tracing: The Behavioral Decision-Making Mechanism of Large Model Agents

Large models, also known as foundation models, are large-scale foundation models (Large/Large-scale Foundation Models) that typically employ deep neural network architectures [**?**]. These are large-scale machine learning models pre-trained with billions or even hundreds of billions of diverse parameters. Supported by powerful computational resources, they can learn extensive knowledge and demonstrate strong generalization capabilities, producing versatile models capable of handling multiple tasks, including natural language processing and question answering, with significant accuracy. Large models are essentially large language models (LLMs) with language as their output modality. Initially, LLMs primarily addressed language text understanding and reasoning tasks. Recently, efforts have focused on developing multimodal large language models (MLLMs) that accept images, video, audio, and text as input to solve more complex language understanding and reasoning tasks [**?**].

Large model agents represent concrete implementations at the application layer of large models, with the large model serving as the "brain" of the agent architecture [**?**]. Agents inevitably inherit the hallucination defects of large models,

extending them into agent hallucinations. While the model-layer hallucination problem has received considerable attention in computer science, with a series of operational definitions developed to support subsequent measurement and optimization research, most studies in journalism and communication have focused on the cognitive risks [**?**], cultural consequences [**?**], and misinformation governance [**?**] triggered by hallucinations, lacking deep understanding of the problem' s essence. This study introduces the computer science perspective on hallucination to provide theoretical advancement in the epistemology of hallucination, offering a more comprehensive scientific understanding for subsequent research and further recommendations on how media should understand and address hallucination issues in practice.

An agent is a high-performance autonomous system built by using a large model as the "brain" or controller of the agent system [**?**]. Large models can guide various components of the agent, playing a role in perception, decision-making, and action execution. As the agent' s brain, large models not only enable natural language dialogue interaction but also allow agents to demonstrate strong adaptability and generalization capabilities when facing multi-task and novel situations, enabling highly autonomous and flexible behavioral decision-making [**?**]. Different modality large models have processing biases, offering possibilities for expanding agent capabilities. For example, large vision-language models (LVLMs) integrate visual modalities, utilizing large-scale image-text pre-training to directly match any given image and text for zero-shot prediction [**?**], providing agents with language-aligned universal visual encoders and zero-shot visual recognition capabilities. This means that agent hallucination is essentially large model hallucination—agents inherit both the generative capabilities and the hallucination problems of large models.

2.2 Defining Hallucination: The Endogenous Nature and Manifestations of Hallucination

In general contexts, hallucination originates from pathology and psychology, defined as a realistic but false perception—"a perception experienced by a conscious individual in the absence of appropriate external stimuli" [**?**]. For large models, the phenomenon of generating untrue or meaningless text shares similar characteristics with such psychological hallucinations. Therefore, computer scientists introduced "hallucination" to explain quality-level errors in model-generated content.

Agent hallucination specifically refers to a series of inevitable errors at the model layer where generated content becomes illogical or unfaithful to the provided source content [**?**]. Specifically, agent hallucination should be understood as fictional, misleading, or fabricated details, facts, or claims generated by large models in text generation, rather than authentic or reliable text [**?**]. Hallucination is a model-generated output that conflicts with constraints, is incorrect, or involves incorrect reasoning about the generated text, or produces unsubstantiated or mis-cited meaningless or false claims, or deviates from expected

behavior in actual deployment, or is completely irrelevant to the task at hand —but in such cases may still be considered grammatically credible (i.e., the generated text sounds coherent).

It is worth noting that the term "hallucination"has not been universally accepted. Some scholars argue that these errors, and even the overall activities of large language models, are best understood as Frankfurt' s concept of "bullshit" [**?**] —the model is indifferent to the truthfulness of its output. This suggests that "hallucination" is not an irreplaceable concept.

2.3 Classification Overview: Analyzing Factual and Faithfulness Hallucinations

The hallucination problem has already attracted attention in traditional natural language generation tasks, but the issue is more complex in large models. Generally, hallucinations in natural language generation tasks can be divided into two main types: intrinsic hallucinations and extrinsic hallucinations [**?**]. Intrinsic hallucinations refer to generated outputs that contradict the source content (the input information or references that large models rely on when generating text, including user-provided direct instructions, context or original materials, external knowledge bases such as Wikipedia or professional literature, and factual information in internal memory). Extrinsic hallucinations refer to generated outputs that cannot be verified from the source content—where we can neither find evidence for the generated output in the source nor assert that it is wrong [**?**].

For large model hallucinations, this classification method, focused on NLP use cases, cannot adequately cover and describe the versatility and task complexity of large models, revealing the limitations of existing task-specific classification paradigms. Therefore, this paper draws on Huang et al.' s hallucination classification method based on the practical application of large language models, dividing hallucinations into factuality hallucination and faithfulness hallucination [**?**], and incorporates fact omission in multimodal large models [**?**] into the factuality hallucination category.

Factuality hallucination refers to situations where large model-generated content does not match or cannot be verified against real-world facts, mainly divided into factual errors, factual fabrication, and fact omission. First, factual errors stem from mistakes in the model' s capture, storage, and expression of factual knowledge, where the output contradicts real-world information, specifically manifested as event information errors and event relationship errors. Event information errors involve mistakes in the constituent elements of the output content itself, while event relationship errors involve mistakes in relationships between elements, such as incorrect correspondence between events and their temporal-spatial contexts in news reports. Second, factual fabrication refers to large model outputs that cannot be verified with real-world knowledge, where the model generates content without real-world basis, including overclaims lack-

ing universal validity due to subjective bias. Finally, fact omission refers to large models ignoring parts of the original text under multimodal instructions, such as neglecting disaster environment descriptions when depicting disaster news scenes.

Faithfulness hallucination refers to situations where large model outputs are inconsistent with user-provided instructions or exhibit internal logical inconsistencies in their own generated content. Faithfulness hallucination is divided into three subtypes: intent inconsistency, context inconsistency, and logical inconsistency. Intent inconsistency means the model' s output deviates from the user' s instructions. While some deviations may be safety-driven choices, the inconsistency here refers to unintentional deviation from non-malicious user instructions—for example, when the user' s intent is translation, but the model mistakenly deviates and performs a question-answering task. Context inconsistency means the model' s output is inconsistent with the user-provided context information. Logical inconsistency means there are internal logical contradictions in the model' s output, commonly observed in reasoning tasks, manifested as inconsistencies between reasoning steps themselves and between steps and final answers.

Additionally, hallucinations across different modalities extend beyond large model hallucinations. For large vision-language models, hallucination manifests as contradictions between visual input (considered "fact") and LVLM text output. When the model' s response to user queries or statements is inconsistent with actual visual data, judgment hallucination occurs [**?**], such as the large model failing to detect and describe objects that should be present in the image, or describing objects that do not exist in the image.

3. Risk Panorama: Hallucination Penetration in Downstream Applications

In downstream applications, agents face various tasks, with representative examples including machine translation, question answering, dialogue systems, summarization, knowledge graphs, and visual question answering. Hallucination manifests differently across these tasks, with risks lurking within.

In machine translation, large model hallucinations mainly manifest as translation deviation, over-generation, or translation failure [**?**]. Translation deviation occurs when translated content completely deviates from the source text's theme and information while remaining linguistically fluent. Over-generation occurs when models produce excessive unnecessary content, making translation results verbose, complex, or even containing information unrelated to the original text. In some cases, models may attempt translation but fail due to input text complexity or model limitations, resulting in translation failure and inability to generate reasonable translations.

In question answering systems, models often rely on external knowledge and memorized prompt information when answering questions. However, when this

knowledge is defective or recall prompts are insufficient, models tend to give incomplete but seemingly reasonable answers [**?**]. When no relevant information is available, models still attempt to answer, producing inaccurate or partial answers [**?**]. If the memorized information stored within the model lacks accurate, reliable, and accessible source support, the model may generate answers based on incorrect or outdated information that is difficult to verify for correctness.

In dialogue with agents, dialogue models primarily imitate data distribution characteristics rather than generating outputs faithful to real information. This means models may simply copy or repeat patterns from training data rather than truly understanding and generating contextually appropriate responses. Due to discourse phenomena, some dialogue models produce "uncooperative" responses [**?**], directly outputting complete evidentiary text instead of providing precise answers according to user needs, or exhibiting information bias or inaccurate details during responses [**?**].

When users employ large models for article summarization, although LLM-generated summaries are usually linguistically fluent, they often lack faithful representation of original document content, showing a gap in accuracy compared to traditional summarization models in human evaluations. On one hand, generated summaries distort information that may exist in the original text, causing output content to be factually inconsistent with the original document and directly affecting summary accuracy. On the other hand, summaries may contain additional information that did not exist in the original text [**?**].

In knowledge graph construction and knowledge generation tasks, models not only cover input information but may also incorporate redundant details from their internal memory. LLMs not only repeat input information but also add redundant internal memory knowledge, leading to knowledge hallucination [**?**]. Users need to distinguish between "correctly generated knowledge" and "knowledge hallucination" —that is, in the knowledge creation process, clearly identifying which content is authentically valid and which is hallucination.

In cross-modal tasks, despite leveraging large language models (LLMs) to enhance language capabilities, "object hallucinations" still exist in large vision-language models (LVLMs). Object hallucination is common in visual question answering, image captioning, and report generation tasks, indicating that even when models excel at language generation, they still have deficiencies in aligning visual information with text, resulting in generated descriptions that do not match actual image content [**?**].

For content production, hallucination means that untruthful situations require constant vigilance during agent usage. Especially for users, the dissemination of hallucination content generated when using agents to process tasks poses potential knowledge propagation risks.

4. Mechanism Analysis: Triple Defects in Data, Training, and Inference

Hallucination originates from the model' s inherent generation mechanisms and knowledge update difficulties, essentially reflecting LLMs' inherent deficiencies in knowledge storage, fact verification, and logical reasoning rather than application-level design flaws. Large models can produce hallucinations at every stage of data, training, and inference. Data-level mismatches, erroneous information, and biases plant the seeds of hallucination; training-stage objective design, knowledge boundaries, and insufficient adaptation to human feedback further amplify hallucination; and inference-stage decoding strategies, attention mechanisms, and logical inference capabilities ultimately manifest hallucination in output.

4.1 Data Processing Bias: Root Causes of Hallucination

Large models learn knowledge through statistical patterns in massive multi-source pre-training data. Data is the foundation of model capabilities, and the vast amount of data comes from the internet, making it difficult to ensure all raw data is high-quality. Models inevitably absorb and reproduce this untrue information, and different models are trained on different data scales and scopes, with some professional knowledge being difficult to obtain. Various reasons plant hidden dangers for large model hallucination [**?**].

First is the issue of data source and annotation bias. In large-scale dataset construction, "source" and "target" are two key concepts. The target is the expected output result—what the model should generate. When training models, source and target typically appear in pairs, with models learning the correspondence between source and target to improve generation accuracy [**?**]. Mismatch between source and target means models may learn inaccurate associations during training, thereby generating hallucinations when producing text. Additionally, if pre-training corpora contain numerous duplicate examples, models tend to repeatedly memorize and generate phrases from these examples, causing inappropriate "repetitive" hallucinations in downstream tasks [**?**].

Data's innate divergence also causes hallucination problems. In natural language generation (NLG) tasks, especially open-domain dialogue systems, models are required to generate natural, fluent, and engaging dialogue. To achieve this goal, models are typically allowed to generate diverse responses. These responses may contain subjective opinions, casual conversation, or even content generated without precise factual support. This task characteristic means models do not need to strictly maintain factual alignment with input source information when generating text [**?**].

Since pre-training requires massive data often sourced from the internet, it inevitably contains false information (such as rumors) and social biases [**?**]. When data includes fake news, unfounded rumors, etc., neural networks have a tendency to memorize training data, and models may remember this erroneous

information and bias (Misinformation & Biases), outputting false statements during generation and causing so-called imitative falsehood [**?**]. Meanwhile, social biases are deeply embedded in social media platforms, and models may inadvertently learn these biases and propagate them into generated content. Although this is not entirely hallucination, certain biases related to gender, nationality, etc., are indeed closely related to hallucination.

Traditional training data struggles to cover all domains and latest information (such as recent scientific research, legal texts, etc.). Once users ask questions beyond the model' s known scope, models may fabricate answers [**?**].

Inferior alignment data also affects hallucination occurrence [**?**]. Foundation models are often fine-tuned by industry for downstream application scenarios. The Supervised Fine-Tuning (SFT) stage often relies on human-annotated instructions and examples, but if this alignment data itself has insufficient information quality or is overly complex and diverse, it can also exacerbate hallucination. Additionally, models are forced to "learn" new knowledge at this stage that may not match their existing knowledge boundaries, leading to misalignment between generated content and facts [**?**].

4.2 Training Mechanism Defects: Inherent Limitations in Capability Acquisition

Limitations in training processes such as pre-training, supervised fine-tuning, and reinforcement learning also bring hallucination problems.

Two aspects of the pre-training process significantly impact model generation quality: limitations of autoregressive language models and exposure bias [**?**]. These are challenges that current pre-training models need to overcome. Since this involves understanding specific model principles, we will not elaborate in detail.

Limitations of autoregressive language models mainly manifest in contextual dependency capture capabilities and attention dispersion issues. GPT-like models (such as GPT-2, GPT-3) employ causal autoregressive prediction, meaning each word' s prediction is based only on preceding words when generating text. Models may struggle to understand relationships between distant words in sentences, leading to logical coherence issues in generated text [**?**]. As sequence length increases, the model' s attention mechanism may become dispersed, leading to unstable reasoning for long-range dependencies.

Exposure bias arises from inconsistency between training and inference stages, causing cumulative errors and hallucination [**?**]. During training, models typically use teacher-forcing maximum likelihood estimation (MLE)—simply put, when generating each word, the model predicts the next word based on real previous words (ground-truth prefix sequences, i.e., human-annotated correct history). During inference, however, the model generates the next word based on its own previously generated words (historical sequences previously generated

by itself). This difference may cause the model to generate sequences during inference that differ from those seen during training, producing cumulative errors. That is, once the model generates an erroneous token during inference, subsequent generation may further deviate from facts based on this error, creating a "snowball effect." This cumulative error causes generated text to gradually deviate from the correct path, ultimately producing hallucination [**?**].

Additionally, SFT knowledge boundaries and over-fitting on new knowledge also bring hallucination problems. SFT often requires models to output content "beyond their original knowledge boundaries" through human instructions. If models cannot effectively absorb this new knowledge, they may fabricate facts. Traditional SFT typically requires models to answer every instruction without encouraging them to express "uncertainty" [**?**]. The inability to reject—when user questions exceed the model' s knowledge scope—leads models to fabricate answers rather than refuse to answer, causing frequent hallucinations.

In traditional reinforcement learning, models learn optimal behavioral strategies through interaction with the environment to maximize cumulative reward. However, defining an appropriate reward function is often very difficult, especially in complex and diverse tasks. Reinforcement Learning from Human Feedback (RLHF) trains and optimizes model behavior through human feedback. This method combines characteristics of reinforcement learning and supervised learning, enabling models to better understand and meet human needs. However, even after RLHF training, even if the model internally judges that a response may be incorrect or inaccurate, it may still output content that contradicts its own internal judgment to please human evaluators, generating a response that better meets human evaluators' expectations to obtain higher rewards or evaluations—i.e., sycophancy [**?**].

4.3 Inference Strategy Limitations: Dynamic Instability in Generation

In large models, the decoding stage refers to the process where models predict possible words based on input prompts or questions using statistical probabilities and gradually generate responses or text content. At this stage, models decide the most likely word at each position, combining these words one by one into complete sentences or paragraphs as output.

To improve generation diversity and creativity, randomness is often introduced during decoding (such as top-k, temperature sampling, etc.) [**?**]. While this helps generate diverse content, it is also positively correlated with hallucination risk.

Since the model' s attention mechanism tends to focus more on local text and ignore global context when generating longer sequences, the attention mechanism is a critical component for large models to understand input text and generate appropriate output. It helps the model decide which parts of the input text to focus on when generating each word. Because models learn during training that

local information is usually more important for generating the next word, when generating longer text sequences, the attention mechanism may focus more on local parts of the input text—the most recent few words or phrases—potentially ignoring the overall context of the input text, leading to instruction forgetting or information confusion [**?**]. In such cases, models may erroneously generate seemingly correct but factually unfounded content based on local fluency.

The softmax function is used to convert model outputs into probability distributions for calculating the probability of each word being selected. It converts raw scores from model outputs into probability values to determine the next word' s selection. Distributed word vectors represent words as vectors in high-dimensional space, with each dimension representing some feature of the word. This approach can capture semantic and grammatical relationships between words. When multiple correct answers exist for the target output, the probability distribution will show multiple peaks. For example, for the question "What will the weather be like tomorrow?" , there may be multiple reasonable answers such as "sunny," "cloudy," or "windy." When the softmax function combines with distributed word vectors, it limits the model' s ability to express this multimodal distribution. The model may struggle to evenly distribute probability among multiple reasonable answers, causing some answers' probabilities to be overestimated or underestimated [**?**]. Because the model cannot evenly distribute probability, it may select inappropriate words during generation, causing generated content to become distorted or hallucinated.

Even if LLMs possess necessary knowledge, in tasks requiring complex reasoning such as multi-hop question answering, models may struggle to accurately utilize this knowledge due to limitations in reasoning capabilities, leading to incorrect answers [**?**]. Models may correctly answer "A is B" but cannot logically infer "B is A," or may miss intermediate connections in multi-hop question answering, causing factual deviation or inference errors. Additionally, overly complex or diverse instruction design (such as multiple constraints) significantly increases hallucination probability because it exceeds the model' s task parsing capabilities.

5. Countermeasures: Cognitive-Technical-Procedural Collaboration

The hallucination problem is a significant challenge facing the media industry in the intelligent era. However, through reasonable cognitive reshaping, process optimization, technical selection, and research exploration, media can find a balance in agent application—fully leveraging intelligent technology advantages while ensuring content authenticity and credibility, thereby promoting steady progress for the media industry in the intelligent wave.

5.1 Cognitive Enhancement: Risk Awareness and Technological Literacy

The media industry needs to view hallucination as an inherent risk in intelligent technology application and adopt corresponding risk management strategies. For agent model layers, large model data security, model security, and technical architecture security are critical. Protecting these areas requires combining standard cybersecurity practices with large model-specific protective measures [**?**]. Data security includes robust measures such as encryption, access control, and data integrity protocols to prevent data poisoning and privacy leakage. Model security requires monitoring [**?**] to guard against concealed errors that are logically valid but fabricate factual details. Infrastructure security emphasizes protecting the hosting environment through firewalls, encryption, and physical protection measures. The infrastructure hosting large models must also be protected against cyber and physical threats.

Media practitioners need basic technological literacy to understand agent working principles, usage scenarios, generation capabilities, and limitations to minimize hallucination occurrence when using intelligent technology. Currently, when collaborating with agents, practitioners need to master prompt engineering techniques with appropriate constraints, avoiding overly complex reasoning processes. Simultaneously, they must maintain critical thinking, rigorously reviewing and verifying agent-generated content.

The hallucination problem is not only a technical challenge but also involves ethical and legal responsibilities. Future collaboration with social sciences, law, ethics, and other disciplines is needed to build a comprehensive regulatory and responsibility mechanism, clarifying responsibilities of all parties when agents generate erroneous information, thereby providing more solid institutional guarantees for widespread agent application.

5.2 Technical Correction: Enhanced Generation and Dynamic Fact Constraints

Hallucination originates from data, training, and model inference, and is to some extent an unavoidable practical problem [**?**]. LLMs learn language patterns by compressing massive amounts of data during training. They compress relationships among trillions of words into billions of parameters that determine connection strengths between artificial neurons. In this compression process, some information loss is inevitable, causing errors in generated content. While data, training, and model inference levels bring problems, they also point directions for regulating and controlling hallucination.

At the data level, data filtering can remove errors, biases, and inaccurate information, thereby reducing false information in pre-training data. Data layer optimization can utilize model editing techniques to correct knowledge within the model, ensuring erroneous information is not solidified. The Retrieval-Augmented Generation (RAG) method can be adopted, combining external

reliable knowledge bases during generation to provide factual basis for model output, thereby reducing hallucination.

At the training layer, hallucination risks from long-tail knowledge and vague concepts can be reduced by improving the pre-training process and optimizing training objectives to reduce exposure bias. During the Supervised Fine-Tuning (SFT) process, tasks and instructions should be reasonably designed so that models do not tend to fabricate information when adapting to new information. For the Reinforcement Learning from Human Feedback (RLHF) stage, training strategies should be adjusted to prevent models from generating overconfident and untrue content to please evaluators.

At the inference layer, factual enhancement decoding strategies can be employed during the decoding process to control sampling temperature and constrain generation probability distribution, ensuring generated content better aligns with facts. Faithfulness enhancement decoding methods can strengthen the model' s faithful expression of input instructions and context, avoiding logical or informational deviation.

Overall, although hallucination originates from multiple factors in data, training, and inference, this does not mean hallucination is uncontrollable. Through multi-level, multi-angle improvement measures—from data preprocessing, training strategy optimization, to decoding improvements during inference—hallucination rates can be reduced to some extent. Meanwhile, completely eliminating hallucination remains a challenge given complex and changing real-world application scenarios. Future research should strive to balance reducing hallucination risk with maintaining model generation diversity, thereby improving the reliability and safety of large models in practical applications.

5.3 Process Guarantee: Human Patching and Human-Machine Collaboration

After comprehensively understanding agents, we recognize both their powerful generation and reasoning capabilities and their limitations. The media industry needs to redefine the role of intelligent technology. However, artificial intelligence technology is not perfect and still requires human intervention and patching to function properly—what is termed "Human Fix" [**?**]. Automated processes do not exist in isolation but depend on operators' rich experience and knowledge, occurring in specific social environments where people communicate and cooperate to solve problems based on actual conditions. However, this knowledge is often overlooked and marginalized. Human assistance is needed when errors occur, contexts require explanation, or decisions must be made.

Future research should explore more theories and methods for human-machine collaboration [**?**] to ensure generated content accuracy and credibility. Additionally, existing evaluation metrics often struggle to comprehensively capture hallucination phenomena. Future development should create more nuanced and multi-dimensional evaluation systems that consider both surface grammatical

fluency and semantic consistency and factual accuracy.

Taking news writing as an example, agents' creative capabilities have surpassed traditional automated writing' s template-based generation for structured news, advancing toward professional journalists. However, unlike human creativity and intelligence, large model agents' generation capabilities are based on probability statistics, with data as their foundation and hallucination as their inherent risk. This means concerns about replacement of repetitive work are reasonable [**?**], but it also means the "value-added" core of journalists' professionalism—such as in-depth reporting and fact-checking—becomes more critical. The media industry should view agents as auxiliary tools that can improve production efficiency and provide creative support, while human journalists should maintain deep involvement and gatekeeping, especially in critical news reporting and fact-checking processes.

The hallucination problem in large model agents reveals the deep contradiction in AI technology's transition from "capability leap"to "trustworthy deployment." Its essence is systematic coupling of data bias, training defects, and inference instability rather than "program bugs." Although current technical paths can mitigate some hallucinations through retrieval-augmented generation and dynamic decoding constraints, complete elimination is still limited by the model' s inherent probabilistic generation logic and the complexity of open-domain tasks. In fact, hallucination is a double-edged sword. Some scholars describe creativity in large models as generating tokens that are both original and diverse while maintaining contextual plausibility [**?**], suggesting hallucination phenomena can enhance creativity in GPT models by allowing them to explore a broader space beyond the most probable token sequences given input conditions. On the other hand, hallucination may also inspire new ideas and perspectives, driving innovation as a "collaborative creative partner"[**?**]. This inspires us that understanding hallucination and its cognitive impacts still requires continued research to move toward trustworthy human-machine symbiosis.

References

[**?**] Jue Mingzi. DeepSeek is Building a "Hallucination Great Wall"on the Chinese Internet[EB/OL]. (2025-02-07)[2025-04-25]. https://mp.weixin.qq.com/s/aMy99RcCq62D9JvTgTUi7A.

[**?**] Kalai A T, Vempala S S. Calibrated language models must hallucinate[C]. Proceedings of the 56th Annual ACM Symposium on Theory of Computing, 2024: 160-171.

[**?**] Vectara. DeepSeek-R1 hallucinates more than DeepSeek-V3[EB/OL]. (2025-01-30)[2025-04-25], https://www.vectara.com/blog/deepseek-r1-hallucinates-more-than-deepseek-v3.

[**?**] Nicola J. AI hallucinations can' t be stopped—but these techniques can limit their damage[J]. Nature. 2025, 637(8047): 778-780.

[**?**] Zhang Zheng, Liu Chenxu. Large Model Hallucination: Cognitive Risks and

Collaborative Governance Possibilities in Human-Machine Communication[J]. Journal of Soochow University (Philosophy & Social Science Edition), 2024, 45(5): 171-180.

[?] Jing Yulun, Zhang Dianyuan. The Manufacturing Logic of Generative AI Illusions and the Cultural Consequences of Their Hyperreal Construction[J]. Journal of Shandong Normal University (Social Science Edition), 2024, 69(5): 113-126.

[?] Zhang Xinsheng, Wang Runzhou, Ma Yulong. Research on Challenges, Opportunities, and Strategies for Misinformation Governance in the AIGC Context[J/OL]. Information Science, 1-23[2025-06-05]. http://kns.cnki.net/kcms/detail/22.1264.G2.20241111.1002

[?] Chakraborty N, Ornik M, Driggs-Campbell K. Hallucination detection in foundation models for decision-making: A flexible definition and review of the state of the art[J]. ACM Computing Surveys, 2025, 52(7): 1-35.

[?] Wu J, Gan W, Chen Z, et al. Multimodal large language models: A survey[C]. 2023 IEEE International Conference on Big Data. IEEE, 2023: 2247-2256.

[?] Xi Z, Chen W, Guo X, et al. The rise and potential of large language model based agents: A survey[J]. Science China Information Sciences, 2025, 68(2): 101-121.

[?] Gong R, Huang Q, Ma X, et al. MindAgent: Emergent Gaming Interaction[C]. Findings of the Association for Computational Linguistics: NAACL 2024, 2024: 3154-3165.

[?] Zhang J, Huang J, Jin S, Lu S. Vision-language models for vision tasks: A survey[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46(8): 2336-2354.

[?] El-Mallakh R S, Walker K L. Hallucinations, psuedohallucinations, and parahallucinations[J]. Psychiatry: Interpersonal and Biological Processes, 2010, 73(1): 34-42.

[?] Chakraborty N, Ornik M, Driggs-Campbell K. Hallucination detection in foundation models for decision-making: A flexible definition and review of the state of the art[J]. ACM Computing Surveys, 2025, 52(7): 1-35.

[?] Sahoo P, Meharia P, Ghosh A, et al. A Comprehensive Survey of Hallucination in Large Language, Image, Video and Audio Foundation Models[C]. Findings of the Association for Computational Linguistics: EMNLP 2024. 2024: 11709-11724.

[?] Chen X, Wang C, Xue Y, et al. Unified Hallucination Detection for Multimodal Large Language Models[C]. Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics. 2024, 1: 3235-3252.

[?] Hicks M T, Humphries J, Slater J. ChatGPT is bullshit[J]. Ethics and Information Technology, 2024, 26(2): 1-11.

[?] Huang L, Yu W, Ma W, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions[J]. ACM Transactions on Information Systems, 2025, 43(2): 1-55.

[?] Ji Z, Lee N, Frieske R, et al. Survey of hallucination in natural language generation[J]. ACM computing surveys, 2023, 55(12): 1-38.

[?] Huang L, Yu W, Ma W, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions[J]. ACM Transactions on Information Systems, 2025, 43(2): 1-55.

[?] Chen X, Wang C, Xue Y, et al. Unified Hallucination Detection for Multimodal Large Language Models[C]. Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics. 2024, 1: 3235-3252.

[?] Guerreiro N M, Alves D M, Waldendorf J, et al. Hallucinations in large multilingual translation models[J]. Transactions of the Association for Computational Linguistics, 2023, 11: 1500-1517.

[?] Zheng L, Chiang W L, Sheng Y, et al. Judging llm-as-a-judge with mt-bench and chatbot arena[J]. Advances in Neural Information Processing Systems, 2023, 36: 46595-46623.

[?] Adlakha V, Ghader B P, Lu X H, et al. Evaluating correctness and faithfulness of instruction-following models for question answering[J]. Transactions of the Association for Computational Linguistics 2024, 12: 681-699.

[?] Dziri N, Milton S, Yu M, et al. On the Origin of Hallucinations in Conversational Models: Is it the Datasets or the Models?[C]. Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2022: 5271-5285.

[?] Das S, Saha S, Srihari R K. Diving Deep into Modes of Fact Hallucinations in Dialogue Systems[C]. Findings of the Association for Computational Linguistics: EMNLP 2022, 2022: 684-699.

[?] Qiu Y, Ziser Y, Korhonen A, et al. Detecting and Mitigating Hallucinations in Multilingual Summarisation[C]. Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023: 8914-8932.

[?] Yuan S, Faerber M. Evaluating Generative Models for Graph-to-Text Generation[C]. Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing, 2023: 1256-1264.

[?] Li Y, Du Y, Zhou K, et al. Evaluating Object Hallucination in Large Vision-Language Models[C]. Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023: 292-305.

[?] Liu Zeyuan, Wang Pengjiang, Song Xiaobin, et al. A Survey on Hallucination Problems in Large Language Models[J]. Journal of Software, 2025, 36(3): 1152-1185.

[?] Lebret R, Grangier D, Auli M. Neural Text Generation from Structured Data with Application to the Biography Domain[C]. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, 2016: 1203-1213.

[?] Lee K, Ippolito D, Nystrom A, et al. Deduplicating Training Data Makes Language Models Better[C]. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, 2022, 1: 8424-8445.

[?] Rashkin H, Reitter D, Tomar G S, et al. Increasing Faithfulness in Knowledge-Grounded Dialogue with Controllable Features[C]. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, 2021, 1: 704-718.

[?] Das B C, Amini M H, Wu Y. Security and privacy challenges of large language models: A survey[J]. ACM Computing Surveys, 2025, 57(6): 1-39.

[?] Lin S, Hilton J, Evans O. TruthfulQA: Measuring How Models Mimic Human Falsehoods[C]. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, 2022, 1: 3214-3252.

[?] Kasai J, Sakaguchi K, Le Bras R, et al. Realtime qa: What's the answer right now?[J]. Advances in neural information processing systems, 2023, 36: 49025-49043.

[?] Paullada A, Raji I D, Bender E M, et al. Data and its (dis)contents: A survey of dataset development and use in machine learning research[J]. Patterns, 2021, 2(11): 100227.

[?] Gekhman Z, Yona G, Aharoni R, et al. Does Fine-Tuning LLMs on New Knowledge Encourage Hallucinations?[C]. Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, 2024: 16336-16351.

[?] Bhattacharya P, Prasad V K, Verma A, et al. Demystifying ChatGPT: An in-depth survey of OpenAI' s robust large language models[J]. Archives of Computational Methods in Engineering, 2024: 1-44.

[?] Wang C, Sennrich R. On Exposure Bias, Hallucination and Domain Shift in Neural Machine Translation[C]. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020: 3544-3552.

[?] Zhang M, Press O, Merrill W, et al. How Language Model Hallucinations Can Snowball[C]. International Conference on Machine Learning, 2024: 59670-59684.

[?] Yang Y, Chern E, Qiu X, et al. Alignment for honesty[J]. Advances in Neural Information Processing Systems, 2024, 37: 63565-63598.

[?] Cotra, Ajeya. Why AI alignment could be hard with modern deep learning[EB/OL]. (2025-09-21)[2025-04-25]. Cold Takes. https://www.cold-takes.com/why-ai-alignment-could-be-hard-with-modern-deep-learning/.

[?] Fan A, Lewis M, Dauphin Y. Hierarchical Neural Story Generation[C]. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, 2018, 1: 889-898.

[?] Alves D, Guerreiro N, Alves J, et al. Steering Large Language Models for Machine Translation with Finetuning and In-Context Learning[C]. Findings of the Association for Computational Linguistics: EMNLP 2023, 2023: 11127-11144.

[?] Yang Z, Dai Z, Salakhutdinov R, et al. Breaking the Softmax Bottleneck: A High-Rank RNN Language Model[C]. International Conference on Learning Representations, 2018: 1-18.

[?] Yuan Y, Wang W, Guo Q, et al. Does chatgpt know that it does not know? evaluating the black-box calibration of chatgpt[C]. Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), 2024: 5191-5201.

[?] Tihanyi N, Bisztray T, Ferrag M A, et al. How secure is AI-generated code: a large-scale comparison of large language models[J]. Empirical Software Engineering, 2025, 30(2): 1-42.

[?] Quan Hui. Impact, Integration, and Collaboration: A Discussion on ChatGPT's Influence on the Media Industry[J]. China Radio & TV Academic Journal, 2023(09): 17-21.

[?] Nicola J. AI hallucinations can' t be stopped—but these techniques can limit their damage[J]. Nature. 2025, 637(8047): 778-780.

[?] Katzenbach C, Pentzold C, Otero P V. Smoothing out smart tech' s rough edges: Imperfect automation and the human fix[J]. Human-Machine Communication, 2024, 7: 1-22.

[?] Guo Quanzhong, Su Liurunwei, Peng Zitao. 2023-2024 Media Industry Large Model Application Report[J]. China Media Technology, 2025(1): 6-10.

[?] Li Zitian. Instrumental Benefits and Systemic Risks: Journalists' Perceptions of AI News Technology[J]. Journalism Research, 2022(11): 29-42+117.

[?] Lee M. A mathematical investigation of hallucination and creativity in GPT models[J]. Mathematics, 2023, 11(10): 2371.

[?] Huang L, Yu W, Ma W, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions[J]. ACM Transactions on Information Systems, 2025, 43(2): 1-55.

Author Introduction: XU Qi (1982—), female, associate researcher and master' s supervisor at the New Media Research Institute, State Key Laboratory of Media Convergence and Communication, Communication University of China. Research interests: intelligent communication, media convergence, digital humanities, and new media. SUN Zhipu (2001—), male, master' s student. Re-

search interests: intelligent media, human-machine communication, and media convergence.

(Responsible Editor: Li Jing)

Note: Figure translations are in progress. See original paper for figures.