Reframing Hallucination in Large Language Models: A Lifecycle-Based, Mechanism-Aligned, and Phenomenon-Consistent Definition

Xinyi Fang†
Universal Village Society
Cambridge, USA
amberrr@bu.edu

Hao Yuan†
Universal Village Society
Cambridge, USA
hyuan@universal-village.org

Hanxia Li†
Universal Village Society
Cambridge, USA
hxli@universal-village.org

Jieren Kou
New York University
New York, USA
jk6697@nyu.edu

Chuqiao Gu
Carnegie Mellon University
Pittsburgh, USA
chuqiaog@alumni.cmu.edu

Wuyang Zhang
Dept. of Elec. & Comp. Engineering
University of Massachusetts Amherst
Amherst, USA
noctis@umass.edu

Xiaoman Duan
Universal Village Society
Cambridge, USA
xduan@mit.edu

Yajun Fang*
Universal Village Society
Cambridge, USA
yjfang@universal-village.org

***Abstract*—While recent advances in Large Language Models' reasoning capabilities have improved their performance across many tasks, the fundamental challenge of hallucination, a form of unexpected generation, persists. Researchers are already attempting to develop methods to detect and mitigate hallucination, even though the term itself lacks a unified definition and is applied inconsistently across the literature. This incongruity underscores the need for a clear and conceptually grounded definition, which is the prerequisite for pursuing effective solutions. We reviewed 76 definitions of hallucination from AI-related literature published between 2023 and 2025. Our analysis revealed substantial overlaps and ambiguities in the definitional boundaries between hallucination and other related concepts, leading to inconsistent usage across the literature. The current definitional vocabulary also remains largely focused on observable properties, such as factuality and faithfulness, rather than the underlying mechanisms that give rise to these outputs. To address these limitations, we propose a Lifecycle-Based, Mechanism-Aligned, and Phenomenon-Consistent definition of hallucination that integrates both the generative mechanisms and their observable manifestations across the stages of data collection, processing, and quality assurance, through model development, to deployment. This reframed definition provides a conceptually coherent foundation for hallucination evaluation and mitigation, offering future guidance for developing benchmarks, designing diagnostic tools, and implementing mitigation strategies.**

***Keywords*—hallucination, large language models, AI ethics, natural language processing, machine learning.**

†Co-First Author
*Corresponding Author

I. Introduction

In the latest wave of generative artificial intelligence (AI) advancement from leading labs, hallucination is moving to the center of product narratives. In OpenAI's debut of ChatGPT5, visibly lower hallucination rates are marketed to emphasize its generation ability compared to its previous models [1].

The term hallucination entered mainstream AI discourse notably in July 2021, when Meta publicly used *hallucination* to acknowledge that their early large language models (LLMs) could "confidently state information that isn't correct [2]." During these years, the term has also gradually entered popular discourse. Cambridge Dictionary selected "hallucinate" as its 2023 Word of the Year [3] and added a dedicated definition for hallucination as "false information that is produced by an artificial intelligence [4]."

Evidence from high-stakes domains, including healthcare and law, shows that AI hallucinations can lead to serious consequences. In healthcare, AI transcription tools have been found to produce concerning hallucinations. Whisper, utilized by over 30,000 clinicians across 40 health systems, was found to produce hallucinated content in approximately 1% of transcriptions, with 38% of these hallucinations containing explicitly harmful misinformation [5]. The legal system faces similar risks. A recent Stanford study reveals that leading commercial legal AI systems provide misleading information in approximately 17% of queries [6]. Courts have witnessed multiple instances of citing bogus cases generated by AI tools and sanctioned lawyers for contempt of court [7], [8]. In response to these incidents, the industry has increased its

979-8-3315-3151-5/24/$31.00

commitment and efforts towards developing and implementing hallucination management strategies [9].

However, even among leading AI developers, the criteria used to define hallucination remain inconsistent. Public documents from major developers adopt different wordings and scopes. For example, OpenAI describes hallucination as factual errors unsupported by reality that misinform users [10]. Meta researchers, by contrast, define hallucinations as errors and irrelevant responses that are especially problematic because they appear correct [11]. While the former definition is restricted to focusing on output factuality, the latter considers faithfulness with user input and confines itself to plausible outcomes. This inconsistency is not merely semantic. It breaks comparability across benchmarks, misguides mitigation, and confuses deployment where risk tolerances differ by domain.

To address this gap, we address the following research questions in this study:

- RQ1: What definitions of hallucination are currently employed in the field of AI, and what conceptual and practical challenges hinder the development of a unified definition?
- RQ2: How can we formulate a conceptually coherent definition of hallucination that bridges existing gaps and provides an actionable basis for hallucination management?

Our contributions are:

- We catalog and analyze 76 definitions of hallucination from AI-related literature published between 2023 and 2025 and highlight the inconsistent and even conflicting vocabulary choices and scopes
- We explain why it is challenging to establish a unified and consistent definition of hallucination, focusing on three factors: (i) the term "hallucination" itself is conceptually misleading, (ii) evaluation baselines for hallucination vary across tasks, modalities, and domains, and (iii) existing definitions lack grounding in underlying mechanisms
- From a mechanistic perspective, we analyze how hallucination was generated across the lifecycle of large language models (LLMs) and propose a lifecycle-based, mechanism-aligned, and phenomenon-consistent definition of hallucination that provides actionable guidance for future hallucation evaluation and mitigation

II. Absence of a Unified Definition

As AI hallucination attracts increasing concerns, the research community has already entered a stage focusing on hallucination detection and mitigation. Paradoxically, the field still lacks a standardized definition of the term itself. The absence of a standardized definition for this prominent topic hinders both research progress and public understanding. This section reviews existing definitions of hallucination across literature, reveals underlying inconsistencies among these definitions, and outlines the growing criticisms of the term "hallucination."

Recent studies have begun to highlight the concerning lack of conceptual clarity regarding "hallucination" in AI. A recent systematic review by Maleki et al. [12] underscores the inconsistency in the characterization of hallucination within the context of text generation across fourteen databases. Meanwhile, other researchers have emphasized the overlaps and ambiguities in definitional boundaries between hallucination and other related concepts. Ioste et al. [13] point out that some definitions overgeneralize hallucination as discrepancies from user expectations, a broader category. In some cases, hallucination is further conflated directly with these omissions [14], reflecting a hierarchical concept conflation in which distinct phenomena are improperly treated as equivalent.

To answer RQ1 and investigate this definitional inconsistency more systematically, we collected 76 definitions from AI hallucination-related publications between 2023 and 2025. Fig. 1 exhibits a three-level hierarchy of the existing vocabulary used in hallucination definition based on collected sources. At Level-1, the depiction of Manifestation Defect dominates with 80.0% of term instances, followed by Manifestation Premise with around 15%, and Mechanism-related with only 3.9%. Within the description of manifestation defects, the factuality-related words account for approximately half of the labels, as shown in Fig. 2, and exhibit substantial fragmentation of synonyms at Level-3 of the hierarchy. Faithfulness is the second largest cluster with 22% in defect description. For the wording of the hallucination premise, plausibility accounts for 76% of the terms, as shown in Fig. 3.

Analysis of our collected definitions reveals major descriptive diversity and wording differences. Many sources oversimplify hallucination by equating it as any incorrect generation [15]–[17]. Existing definitions also differ in scope and coverage. Guided by distinct objectives, different authors emphasize different aspects. Some definitions center only on factual accuracy [10]. Others, especially studies in the field of natural language processing [18], [19], emphasize the faithfulness to input context. These different emphases lead to divergent labels for the same phenomena.

Moreover, these definitions even appear to contradict and disagree on what qualities a hallucinated output must have. Some definitions stipulate, as a premise, that hallucinated responses must be plausible and fluent [20]–[26]. In contrast, other definitions also encompass outputs that are, obviously unreasonable [27], obviously nonsensical [20], [22], [28]–[30] or logically inconsistent [13], [31], [32], revealing a fundamental contradiction in the assumed characteristics of hallucinated outputs. This inconsistency is further reflected in how some sources restrict hallucination to instruction-compliant outputs [33], while others extend the term to include outputs that deviate from the user's request [31], [34].

Such definitional confusion confounds efforts to benchmark models or develop standardized evaluation: divergent labeling practices, for example, treating a missing detail as a "hallucination" in one study but not in another, will result in incompatible criteria for success.

Overall, our collected definitions of hallucination share two notable characteristics:

First, Fig. 4 shows that manifestation-centric terminology

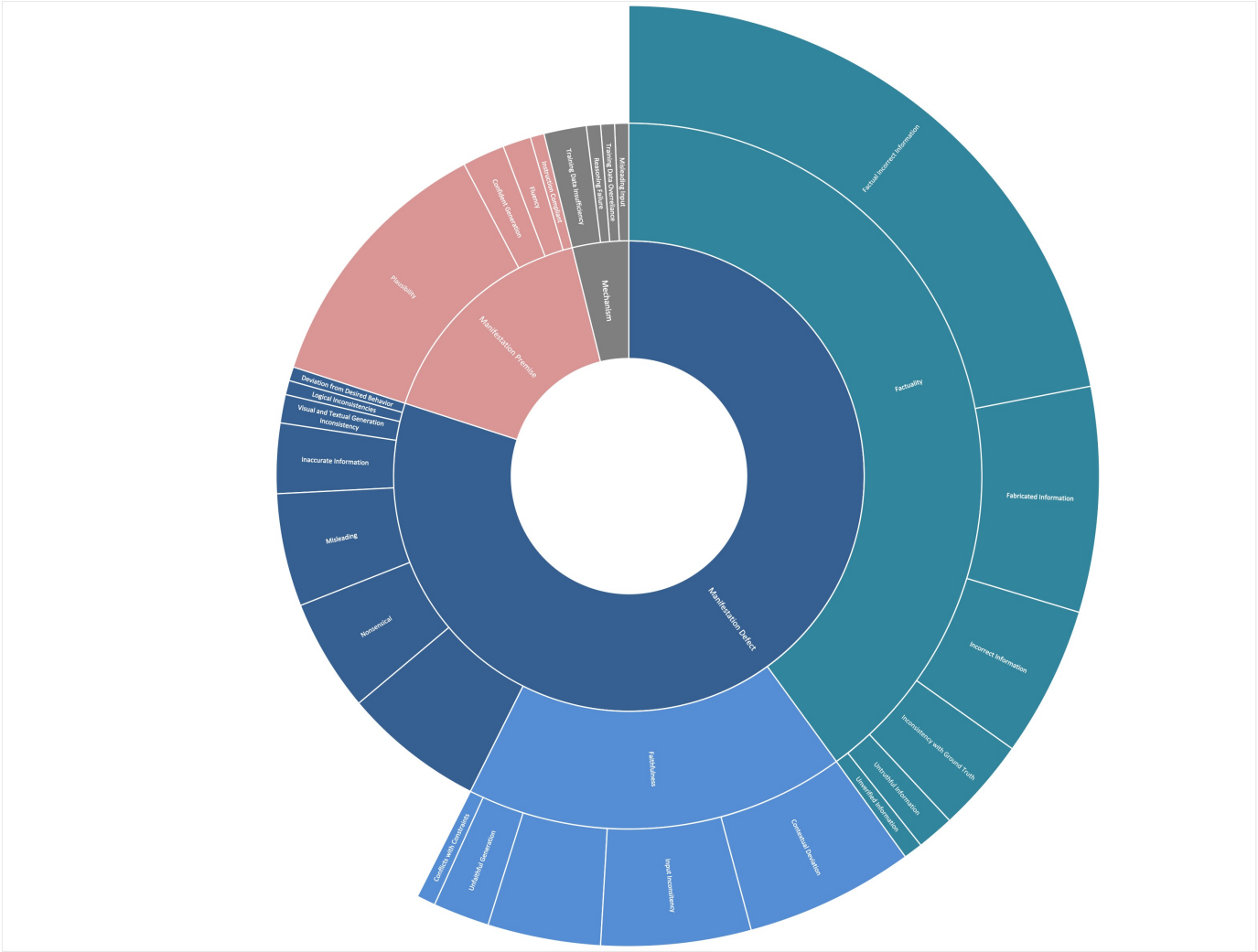

Fig. 1. Existing Definition Schema

dominates while mechanism-oriented labels are scarce. Most key terms in current definitions, such as fabricated, untruthful, inaccurate, ungrounded, and their variants, refer only to models' externally observable behavior. They describe what the model says, not how or why it produces the statement, remaining at the level of manifestation.

Second, despite the term variety, these words can almost always be reduced to two underlying concerns, factuality and faithfulness, which are also the two notable term choices of the collected definition as shown in Fig. 2. This consolidation echoes Ji et al.'s [35] manifestation–level taxonomy: Factuality, concerning whether the output contradicts ground truth; Faithfulness, concerning whether the output is inconsistent with user-provided context [26], [31], [35].

In summary, the current definitional vocabulary still focuses on observable properties, mainly factuality and faithfulness, rather than the underlying mechanisms that generate the outputs.

III. Challenges in Reaching a Unified Definition

As shown in the previous section, the community's use of hallucination remains anything but consistent. Definitions cluster around loose keywords, vary across application domains, and remain focused at the phenomenal level. This section discusses three reasons why a unified and consistent definition has yet to be established: (*i*) the "hallucination" label itself is misleading, (*ii*) hallucination evaluation baselines vary across tasks, modalities, and domains, and (*iii*) current definitions are not grounded in underlying mechanisms.

A. *"Hallucination" as a Misleading Label*

While "hallucination" has become a predominant label for describing certain unexpected outputs from generative AI, many researchers and practitioners argue that the term itself is misleading or counterproductive.

Two primary critiques dominate the discourse: (*i*) "Hallucination" carries anthropomorphic baggage that overstates the model's cognitive nature, and (*ii*) "Hallucination" frames the

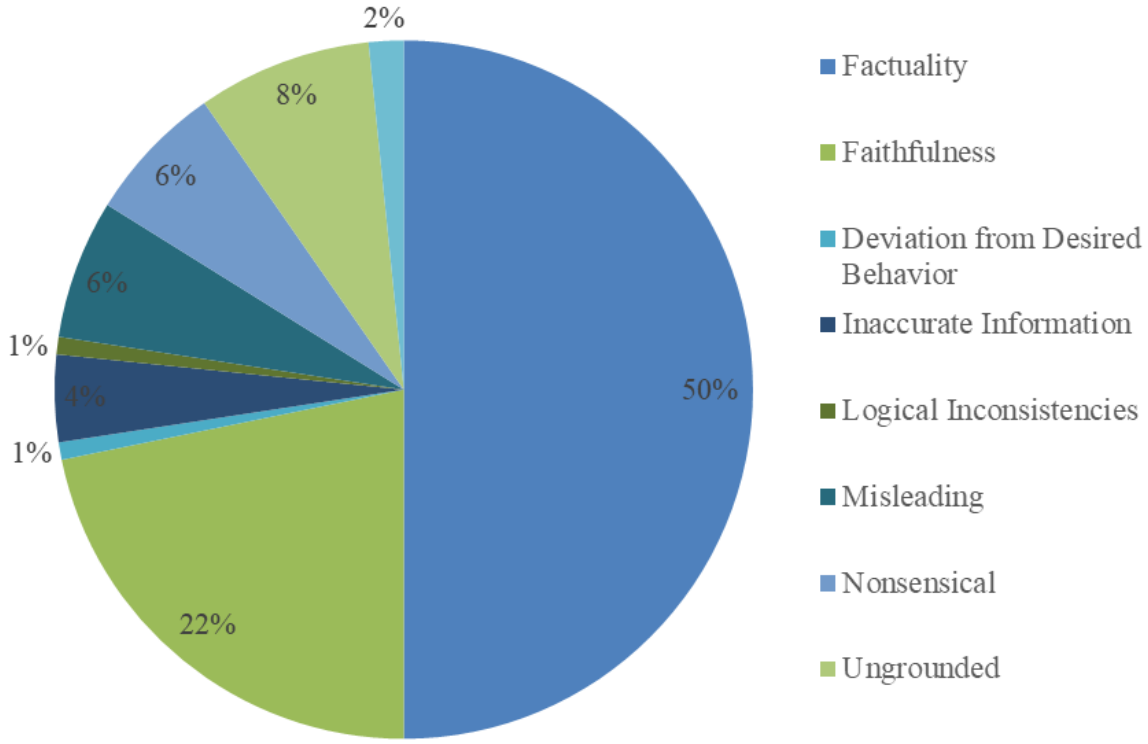

Fig. 2. Defect Description

phenomenon purely as a defect, overlooking its potential to foster innovations.

Anthropomorphic Baggage

Originating from psychiatry, hallucination describes a conscious agent perceiving stimuli that do not exist [36]. Applied to LLMs, the word itself inherently carries anthropomorphic meanings, unwittingly implying a form of perception, imagination, and subjective experience—capabilities they do not possess.

As Fayyad notes, the wording conveys an over-expectation towards models, and risks confusing users about how the system actually works [37]. Unlike human beings, LLMs, however, live in a world of mathematical foundations, from which they sample a learned distribution and assign probabilities to tokens given prior context, without any understanding or concept of correctness.

Anthropomorphizing language may also lull users into over-trusting [37] and overlooking the structural inevitability of hallucination.

Gundogmusler et al. trace apparent hallucinations to the transformer's probabilistic foundations that any token with non-zero probability can eventually be emitted, even when it contradicts ground truth [38]. And using a diagonalization argument, Xu et al. show that for any computably enumerable models, there exists a computable ground-truth function on which each will hallucinate [39].

Therefore, labeling every deviation from the expected output as a hallucination can make unavoidable outcomes seem trivial and obscure the core problem.

Overlooking Creativity

The word "hallucination" inherently carries a negative framing. Yet an increasing number of researchers and practitioners are challenging this narrow view and criticizing that "hallucination" is a misnomer that obscures the models' underlying creative potential.

Recent work has introduced the concept of "valuable hallucinations" to characterize outputs that, while factually incorrect, equip important cognitive and creative functions [40], [41]. Sui et al. [41] show that such outputs frequently display heightened narrativity and semantic coherence, properties that contribute to their creative utility. Altman, OpenAI's CEO, has also described hallucination as "more feature than bug" [42], arguing that LLMs outpace their adoption for routine tasks precisely because they recombine learned data into novel perspectives.

Extending beyond corporate branding, empirical studies further confirm that hallucination and creativity frequently co-occur [43]–[45], and deeper theoretical analyses reveal they share the same generative mechanism and even mathematical foundation [38], [46]. For example, in Lee et al.'s verification, any attempt to drive hallucination probability arbitrarily close to zero simultaneously pushes a paired "creativity term" toward zero [44].

Consequently, labeling all such outputs under the pejorative

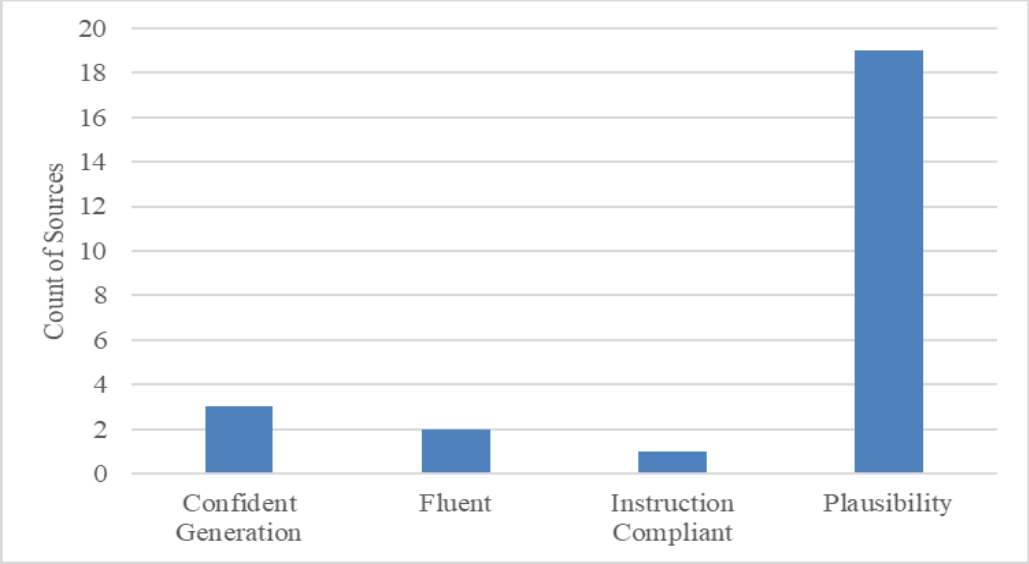

Fig. 3. Premise Constraints

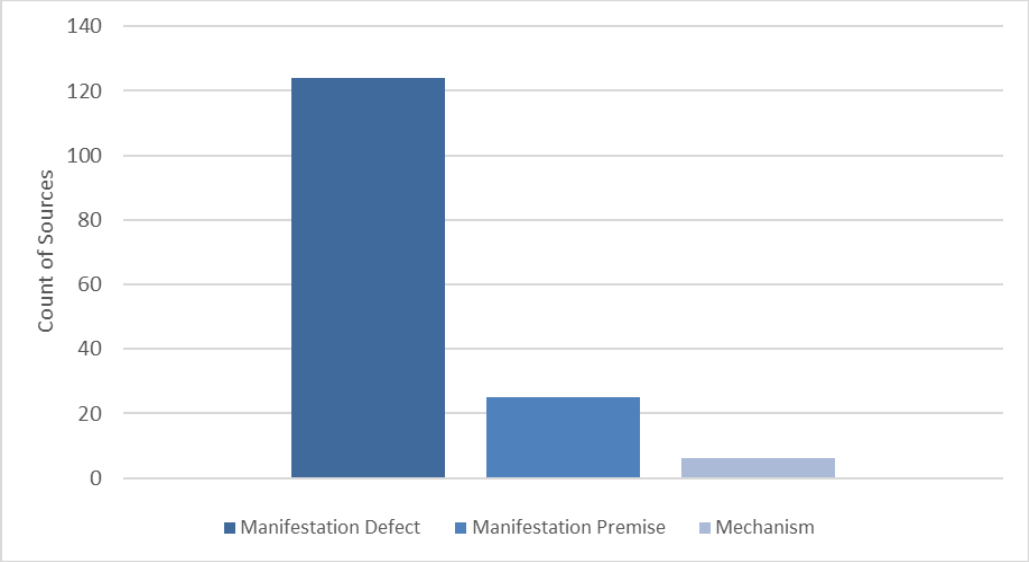

Fig. 4. Distribution of Definition Schema

term "hallucination" flattens a multifaceted phenomenon into a single flaw, erasing its innovative aspects and constraining opportunities for future nuanced evaluation and mitigation.

Emergence of Alternative Terms

Given all these reasons, it is unsurprising that some researchers have begun eschewing the term "hallucination." Alternatives such as "confabulation" or "fabrication" [41], and "delusion" [12] have been proposed. While none has achieved consensus, the competing terms reflect a growing discomfort with hallucination as the descriptor. The use of this term then becomes further confused and diversified.

B. One Evaluation Baseline Does Not Fit All

Beyond the controversy over the label itself, definitional divergence also stems from various baselines for determining what outputs are considered unacceptable. In practice, these baselines differ by tasks, models, and application domains. Therefore, achieving a universal judging criterion across multiple settings is difficult with a single phenomenological definition.

1) Task-Dependent Benchmark: Outputs' truth conditions are task-relative, and so are the current ways benchmarks operationalize "hallucination." Each benchmark caters to a specific task or limited tasks and develops metrics suited to its own purpose.

Tasks such as question answering (QA), text summarization, translation, and dialogue have relatively clear baselines, which require outputs to match known facts and established true reference answers.

- **QA** is a fundamental task in natural language processing.

TruthfulQA [47] and FACTOR [48] are two of the most widely used QA benchmarks, which set answer truthfulness and factual accuracy as metrics, assessing models' ability to understand textual information and generate accurate answers.
- For **text summarization** tasks, benchmarks such as Gigaword [49] and Multi-News [50] focus on semantic consistency and faithfulness between generated content and source information.
- **Dialogue** benchmark, such as KdConv [51], includes outputs' fluency and coherence to the context and knowledge information.

In open-ended tasks, generations lack fixed ground truths because the evaluation metric depends on references or verifiable facts, and such tasks do not have stable references. Here, hallucination becomes subjective. Outputs may be factually false yet still coherent, persuasive, or useful for the task.

Recent work notes that most benchmarks overlook hallucinations in such free-form settings [52]. Even when open-ended tasks are included, as in TruthQA, metrics often extend to contain more subjective criteria such as informativeness [47]. From a modeling perspective, Andriopoulos et al. [53] point out that a finite-context parametric model can only approximate the open-ended distribution.

Therefore, under the existing definition of "hallucination," benchmark design can only target task-specific manifestations.

2) Modality-Dependent Expectation: The definition of hallucination also shifts when moving from purely text-based LLMs to multimodal systems such as large vision–language models (LVLMs). For LLMs, hallucination typically centers on text generation quality and factual consistency for a purely textual ground truth. In LVLMs, however, evaluation must also account for the accuracy of visual perception, visual–textual alignment, and cross-modal reasoning.

Compared with text-only settings, benchmarks for multimodal hallucination capture non-existent objects, inaccurate object attributes, or incorrect object relationships. Benchmarks in image captioning, such as POPE [54], are designed to detect object hallucinations.

Because these risks are modality-specific, applying a single, text-centric definition of hallucination to LVLMs risks overlooking modality-driven failure modes.

3) Domain-Dependent Risk Baseline: Risk tolerance for hallucination varies across domains. In high-stakes contexts, such as clinical records, legal analysis, and financial advising, even a small probability of factual error can be unacceptable. By contrast, in creative or exploratory settings, factual deviation is often tolerated and may even be encouraged.

Early "factual but wooden" chatbots were criticized for sounding mechanical and uninspiring. Models such as ChatGPT rose in popularity in part because they produced imaginative, emotionally rich responses. Many users treat generative AIs not only as information utilities but also as conversational partners. For example, some even steer the dialogue toward practices like online divination [55].

Such outputs have no ground-truth reference, fitting any given definition of hallucination, yet they deliver clear user experiential value.

Commercial roadmaps confirm this demand. Both OpenAI and Google DeepMind market "creative writing" power as headline features. OpenAI highlights cases where authors co-develop story ideas and narratives with the model's imaginative suggestions [56]. A study with fifteen theatre and film professionals reported productivity gains and innovative plot twists under the assistance of DeepMind's Dramatron [57]. In scientific discovery, controlled "hallucinatory" exploration has even been found to aid molecular innovation [58].

Eliminating this capacity entirely could degrade usability in domains such as marketing, game design, and storytelling. A universal, rigidly factuality-based definition of hallucination would therefore over-penalize systems in these domains.

C. Phenomenon-Based Labels Without Mechanistic Grounding

As discussed in the previous section, most current definitions of hallucination focus on outward manifestations. The labels are essentially black-box depictions, with which the phenomena are named, but their mechanisms remain opaque.

Furthermore, mechanistic studies show that hallucination exhibits a many-to-many mapping between causes and appearances. One single symptom may have multiple triggers across a model's lifecycle. For example, Yu et al. [59] investigate non-factual hallucinations that misalign with world knowledge, tracing them to two distinct underlying causes and designing targeted remedies for each. Similarly, Zhang et al. [60] examine the phenomenon of knowledge overshadowing-when querying knowledge with multiple conditions-certain conditions dominate others, leading to hallucinated outputs. By pinpointing its mechanism, they propose detection strategies based on early warning signals observed inside the model.

Without being anchored in underlying mechanisms, mitigation remains non-targeted, and evaluation risks measuring surface behavior rather than root causes. The phenomenon-level definition, therefore, lacks effective guidance for mitigation and evaluation measures.

Taken together, these factors explain why the field still lacks a unified and consistent definition of hallucination.

IV. From Data, Through Model Development, to Deployment - Mechanisms of Hallucination Across the LLM Lifecycle

Recognizing that purely phenomenological labels of hallucination fail to generalize across tasks, modalities, and their underlying causes, we shift the lens inward from phenomenological to mechanistic analysis. To make definitions actionable guidance for hallucination management, we must tie observed symptoms to the mechanisms that generate them. This section dissects the entire LLM lifecycle to map the inherent failure points.

A. Hallucination Begins with the Data

The data used to train LLMs lay the foundation for both their capabilities and their limitations. Hallucination risks often originate here, long before model development or deployment decisions are made.

1) Data collection and Knowledge boundary: LLMs inherit all of their factual knowledge base from the corpora on which they are pre-trained [61]. The scope of those corpora thus defines an implicit knowledge boundary [62]. Materials that fall outside this boundary include content that is absent, outdated, or sparsely represented in the training corpora. Such gaps form knowledge blind spots, which force the model to select tokens based on non-existent or misunderstood regulations during decoding [11], resulting in hallucinations.

One major coverage gap of knowledge boundary formation is temporal staleness: the model's knowledge is frozen at the time of pre-training without outdated information and fails to keep up with the rapidly evolving world [26], [39]. Moreover, copyright laws and privacy restrictions exclude many high-quality sources from large-scale scraping [11], [63].

The limited parameter budget of current LLMs still falls short of encoding long-tail details of human knowledge [61]. When the decoder later encounters a query that sits outside this encoded subset, it assembles tokens that carry model probability but high epistemic uncertainty, yielding fluent yet incorrect answers [59], [64].

2) Data processing and Quality: Whereas data collection sets the boundary of potential knowledge, hallucination can also stem from suboptimal preprocessing practices and poor data quality, including data accuracy, frequency, and diversity.

During data processing, annotation choices and insufficient data cleaning play key roles. Irrelevant annotations will lead to mismatches in token-level relationships [65], [66]. In supervised settings, the use of hard labels—where only one "correct" token is reinforced—limits the model's flexibility to handle ambiguity [67]. This practice, though computationally efficient, encourages brittle decision-making and can result in confident but incorrect outputs when multiple valid completions exist [67]. As for data cleaning, a key concern is the inadequate cleaning of incorrect knowledge [68]. Moreover, when encoded with multiple contradictory pieces of information about the same topic, the model shows a higher tendency to generate inconsistent outputs depending on the specific context or query formulation [69].

Even with optimal processing, hallucinations can arise from flaws inherent in the source data quality itself.

Data Accuracy. The collected data can include both structured and unstructured data [11]. Much of the training corpus consists of unstructured web data, which is easy to collect at scale but often contains errors, misinformation, biases, and outdated information [70]. Secondly, when LLMs encounter misinformation, contradictory facts, or systematically biased content during training, these elements become integrated into the model's parametric knowledge base [26], [39]. In addition to content-level noise, data compression during pretraining data storage introduces further information loss and distortion. As Yin et al. show, higher compression ratios observably reduce downstream model performance [71].

Term Frequency Imbalance. Hallucinations frequently result from statistical imbalances in training data, particularly when popular conditions dominate rare frequency tokens [60], [72]. When the model sees certain entities disproportionately during training, it may overemphasize those entities [72]. Zhang et al. show that such high imbalance ratios can lead to knowledge overshadow, where statistically dominant tokens tend to suppress others and dominate the generation process when the model encounters queries with multiple constraints [60].

Insufficient Diversity. A well-documented failure mode of LLMs involves tasks requiring negation or counterfactual reasoning [66], [72], [73]. Several top models have been shown to hallucinate frequently on negation-heavy prompts, due in part to the scarcity of negative instruction data in their training sets [73]. Furthermore, most standard training corpora do not represent counterfactual statements, which creates a conflict between pretraining and downstream tasks [66]. Kim et al. demonstrate that implanting counterfactual thinking into model training can significantly reduce hallucination rates [72], offering empirical validation for the claim that a lack of contextual diversity contributes to hallucination generation.

In multimodal domains, diversity limitations are even more pronounced. For image and video tasks, the available datasets are often repetitive and lack sufficient variety in object representations [74], due to the high cost of collecting high-quality data pairs [75].

B. Hallucination continues during Pretraining

Pretraining, the stage in which LLMs learn to predict the next token from vast text corpora, embeds not only factual knowledge but also statistical patterns and implicit preferences, making it a foundational source of hallucination in model outputs. This influence manifests through multiple mechanisms, which are outlined in the following discussion.

1) Memorization and Statistical Bias: During LLMs' next-token learning, the training data frequency and the model's memory rules, which determine how statistics are stored and retrieved, jointly influence how the model weighs different pieces of information during generation. The resulting frequency bias and attestation bias are the main ways through which those biases later leak into hallucination [76].

Frequency Bias. One major source of hallucination is the model's weighted memorization of corpus-level token frequencies, as McKenna et al. reveal [76]. LLMs internalize relative frequency priors and skew generation toward high-probability patterns seen in pretraining [77]. For example, in textual entailment tasks, statistically common entities are more easily retrievable in memory, while low-frequency information may be effectively suppressed even when it is more relevant to the current context. This frequency-driven memorization also explains why less frequent long-tail knowledge is especially

fragile. The hallucination rate increases when the model encounters long-tail knowledge requirements [26], [78], where it may fill in the output with plausible but incorrect information.

Attestation Bias. During pre-training, the model learns to use named entities as "indices" into its propositional memory [76]. Once the "indices" have been attested anywhere in the corpus, the model tends to affirm the statement, regardless of the new context or contradictory premise [76]. Teacher-forcing training amplifies the effect, leading to stochastic parroting, from which the model only learns to parrot and imitate previously seen continuations rather than reason about entailment [34], [79].

The memorization mechanism explains why models often struggle with specialized knowledge, minority perspectives, or edge cases that were underrepresented in training data.

2) Attention Pathologies: Hallucinations in LLMs can emerge from intrinsic limitations in the self-attention mechanism—the foundational architecture of transformers [80], which serves as the backbone of most modern LLMs. Since attention governs how input tokens are weighted and contextualized [81], its failure directly impacts factual fidelity. Two primary failure modes are frequently observed: unexpected attention to outlier tokens [59], [82]–[84] and inadequate attention to relevant content [85], [86].

Unexpected Attention Distribution. Self-attention mechanisms occasionally assign disproportionately high weights to irrelevant or outlier tokens. These tokens may act as noise or redundant information, distract the model from key context, or even dominate the output semantics despite lacking informative value [82]. Yu et al. [59] demonstrate that some multi-head attention modules hallucinate due to the failure of choosing the correct object attribute in upper transformer layers, where false-premise attention heads override other accurate retrievals, resulting in the factually incorrect selections and distorting output accuracy [83], [84].

Lost Attention. Transformers also have trouble dealing with long context. As input length grows, attention becomes diluted and drifted [87], making it increasingly difficult to maintain focus on semantically important tokens. This limitation stems from the computational bounds of self-attention, which lacks sufficient capacity to model hierarchical or periodic finite-state structures, as Hahn demonstrates [85]. Similarly, Chiang et al. illustrate that self-attention mechanisms struggle to capture both long-range dependencies, degrading contextual fidelity in extended sequences [86].

A salient manifestation of this limitation is the "lost in the middle" effect. Studies reveal that LLMs often neglect the middle portion of long input sequences, attending primarily to the beginning and end [88], [89].

3) Cross-Modal Biases: As LVLMs extend LLMs to handle multimodal inputs, they inherit and compound hallucination risks from both visual and textual modalities. LVLMs are typically composed of pre-trained encoders for vision and language, and an alignment module. Each introduces modality-specific failure modes that, when combined, lead to distinct cross-modal hallucinations [90], [91].

Vision Encoder Limitations. The inability of vision encoders to accurately ground images stems from their tendency to capture low-resolution and poor visual granularity [92], with an overemphasis on visually salient objects while neglecting fine-grained relationships or background context [93].

Language Priors. The vision encoder limitations then increase reliance on language priors. This overreliance leads models to hallucinate based on frequently co-occurring object patterns learned during pretraining [94]–[96], rather than grounding predictions in the visual input. When the visual evidence conflicts with these learned associations, the result is cognitive dissonance or expectancy violation [97]. Therefore, LVLMs often exhibit an overconfidence in their linguistic knowledge [75]

Alignment Module Bottlenecks. To connect vision and language components, LVLMs use lightweight alignment architectures such as linear projection or Q-Former modules [90]. However, these projections also act as bottlenecks, limiting the effective token bandwidth between modalities. Therefore, as only general information is available [98], the detailed semantic alignment signals between image and text vectors are insufficient and weak [91], leading to erroneous cross-modal associations and hallucinated object references.

C. Hallucination Extends Through Training

Beyond data and pretraining, hallucination continues to accumulate during subsequent training stages, particularly through misaligned uncertainty generation and limitations in task-specific fine-tuning.

1) Inadequate Handling of Uncertainty and Output Overconfidence: A common source of hallucination arises when there is a deeper misalignment between the model's internal uncertainty and its tendency to produce overconfident outputs despite lacking sufficient information to provide reliable answers [99].

The root cause of such misalignment can be attributed to the limitations in LLMs' ability to handle uncertainty. Most LLMs are not properly trained to acknowledge knowledge boundaries or express indetermination, as Tupsakhare et al. claim [24]. This deficiency manifests in the following two key aspects:

Inability to Reject. Rather than declining to answer ambiguous or unanswerable questions, models frequently make unreliable guesses, generating plausible but unsupported completions [100]. This creates a systematic bias toward overproduction—a tendency to fill informational gaps with fluent speculation, even when abstaining would better serve factual accuracy. Encouraging models to explicitly say "I don't know" [101], as seen in OpenAI's recent efforts on ChatGPT 5 development [1], or tuning for higher abstention rates [102] has been shown to mitigate hallucinations.

Consistency over Accuracy. Models are also trained to prioritize fluency and coherence during generation, even when doing so at the expense of truthfulness [70], [103]. This fluency bias is a product of the reward models embedded in the training objective [104], where coherent outputs are positively reinforced regardless of factual accuracy. Over time,

this creates weight configurations that strongly associate fluent generation with positive outcomes, even when fluency comes at the expense of accuracy. This mechanism underlies a typical phenomenon: hallucination snowballing, once an early invented fact is emitted, subsequent tokens tend to maintain narrative self-consistency, even when doing so compounds earlier errors [103].

The culmination of these failures is the generation of overconfident output [26]. Rather than indicating when it lacks sufficient information to provide a reliable answer, the model often generates outputs even when the probability of the predicted token is low.

2) Fine-tuning Alignment: Fine-tuning alignment is often applied to reduce hallucinations and tailor LLMs to specific downstream tasks [105]. However, this stage can introduce new sources of hallucination, particularly under domain shift, where the input distribution diverges from the pretraining regime [106].

One key factor is exposure bias, which arises from the discrepancy between training with teacher-forced inputs and inference generation [107]. This mismatch leads to incremental distortion in test-time outputs, increasing the risk of hallucination [106].

As models are exposed to facts not grounded in their pre-training distribution, they become more prone to generating unsupported assertions. Gekhman et al. [108] clarify that LLMs actually struggle to learn new knowledge during the fine-tuning stage, and any unknown knowledge they are exposed to will instead become a trigger of hallucination, appearing as a form of training overfitting to novel data.

Additionally, learning new knowledge during fine-tuning can degrade or overwrite previously acquired information [109]. Fine-tuning may induce catastrophic forgetting [110], where alignment objectives suppress the capabilities gained during pre-training. The quality and coverage of alignment data also introduce additional complications [111]. New information that contradicts pre-training data can exacerbate knowledge conflicts, thereby creating hallucinations in domains where the model previously performed well.

D. Hallucination Persists Throughout Inference and Deployment

Even with flawless pretraining, careful training alignment, and high-quality data, hallucinations can still emerge during inference and deployment. These inference-stage hallucinations can be broadly categorized into two types: failures in how content is internally represented and encoded, and failures in how that content is decoded and expressed during generation.

1) Content Encoding: Rather than achieving a truly logical and semantic understanding of inputs, LLMs rely on shallow linguistic and object co-occurrence statistics learned during the training process [83], [95].

Vulnerability to Prompt Variation. However, real-world downstream tasks often desire model behaviors that diverge from training objectives, creating model encoding dissonance [97]. This statistical approach to language processing then leads to failure in handling unexpected inputs and nuanced user needs, for example, the requirement of generating counterfactual reasoning [66] and handling chit-chat dialogue [112].

Fragility of Contextual Comprehension. Furthermore, recent studies show that even semantically equivalent paraphrases can elicit divergent factual outputs [64], highlighting the fragility of LLMs' contextual comprehension.

- **Deficits in contextual awareness** also contribute to this failure mode. LLMs often struggle with maintaining consistent attention over long contexts, leading to contextual misinterpretation [24]. Response strategies such as context-aware decoding [113] has been proposed to address these weaknesses and successfully demonstrated empirical reductions in hallucination rates.
- **Context length** introduces further risk and is challenging to manage. While models with small context windows are limited in reasoning capability [53], excessively long contexts can induce semantic phase transitions, cause semantic drift from the intended meaning, and increase hallucination risk [46].

2) Decoding Traps: The decoding process of how LLMs select the next token plays a crucial role in hallucination emergence.

Likelihood Trap in Deterministic Decoding. Traditionally, decoding strategies such as greedy decoding and beam search [114] aim to maximize likelihood by selecting the most probable tokens at each step. While these strategies yield fluent outputs, as Zhang et al. [115] note, they sometimes fall into the so-called likelihood trap, where high-probability sequences often fail to align with human quality judgments, despite surface-level fluency. This results in outputs that diverge from human-like language patterns [116] and generate confident but unfounded assertions [117].

Sampling Randomness in Stochastic Decoding. In response to the likelihood trap, most modern LLMs adopt stochastic sampling methods such as top-k or top-p sampling [118] to inject diversity into the output, rather than always selecting the highest-scoring token [70]. However, this flexibility comes with its own trade-off: increased generation uncertainty. Sampling randomness introduces volatility in the next token selection of the output sequence, resulting in outputs with high uncertainty and increased hallucination rates [26], [39].

V. Toward a Phenomenon-and-Mechanism Consistent Definition

A. Phenomenon Classes

Evidence from the literature suggests that external hallucinations cluster into several high-level classes. Factual hallucinations include factual incorrectness and inconsistency [26]; entity- and relation-error hallucinations [16]; numeric or temporal mistakes (e.g., misdated events, incorrect ages) [39]; acronym ambiguity [39]; geographic errata [39]; and time-warp scenarios [39]. Unverifiability and fabrication cover

responses that cannot be substantiated or are out of date [26]. Instruction and context inconsistency refers to the model's failures to follow user instructions, mismatches with provided context, content hallucination, misinterpretation, and needle-in-a-haystack retrieval misses [31], [39]. Logical inconsistency captures internal contradictions and semantic distortions [31], [39]. Narrative elaboration includes intrinsic and extrinsic additions, including "factual mirage" and "silver-lining" phenomena, in which the model adds imaginative but irrelevant details [39]. Finally, incompleteness covers truncated or partially missing responses [16].

B. Mechanism Categories

As shown in Fig. 5, hallucination originates from mechanistic errors introduced at different stages of the LLM lifecycle, including:

- **Data collection and processing** – narrow or outdated knowledge boundaries, insufficient cleaning and annotation, term frequency imbalances, and lack of diversity lead to missing or distorted facts.
- **Pretraining** – memorization and statistical biases (frequency and attestation), attention pathologies (unexpected attention, false premise heads, lost attention), and cross-modal biases (vision encoder limitations, language priors, projection bottlenecks) introduce structural preferences that later leak into generation.
- **Training** – inadequate handling of uncertainty and overconfidence, exposure bias, domain shift, and catastrophic forgetting misalign the model's confidence with its knowledge.
- **Inference and Deployment** – failures in content encoding (prompt variation, fragile contextual comprehension) and decoding traps (likelihood trap in deterministic decoding, sampling randomness) cause late-stage distortions.

These mechanisms do not act in isolation: there is a compounding effect of stage-to-stage transmission, where hallucinations introduced earlier can be amplified or reshaped downstream in the LLM lifecycle. A missing fact in the corpus can be reinforced by memorization bias and then misallocated by attention pathologies, culminating in a wrongly asserted entity. Conversely, even accurate internal representations can be compromised by poor decoding strategies. Moreover, we identify a feedback loop in which hallucinated outputs, when incorporated into subsequent data collection, could contaminate future training sets, thereby perpetuating and potentially worsening hallucinations in later model iterations.

C. Mapping Mechanisms to Phenomena

Table I summarizes the principal connections between phenomenon classes and mechanistic causes.

The table shows that certain mechanisms recur across multiple phenomena. For instance, attention pathologies contribute to factual hallucinations, instruction-related inconsistencies, and logical breakdowns: when attention is misallocated, the model may distort factual retrieval, misinterpret user intent, or lose semantic coherence. Similarly, memorization bias drives factual errors and narrative elaboration, as the model may overrely on high-frequency patterns or memorize training artifacts, producing plausible but incorrect details.

Fine-tuning misalignment and exposure bias explain why unverifiable or fabricated statements appear: models are optimized to produce fluent, confident responses, even when their underlying knowledge is uncertain or outdated. At the inference stage, content encoding errors and decoding traps often lead to instruction inconsistency or incompleteness—for example, skipping contextual cues, truncating lists, or generating fragmented outputs. Finally, the feedback loop shows how hallucinated outputs, if incorporated back into training corpora, can contaminate future datasets and amplify factual or fabricated errors over time.

D. Reframed definition

Drawing together the phenomenon categories and mechanistic causes, we propose a reframed definition of hallucination:

Hallucination in a large language model is not a single failure but the external manifestation of internal errors introduced and propagated through the model's lifecycle. These internal errors include data collection, processing, and quality issues, memorization bias, attention pathologies, cross-modal biases, exposure bias, fine-tuning misalignment, and failures in content encoding or decoding. When these mechanisms interact, they produce observable phenomena, such as factual inaccuracies, unverifiable assertions, instruction or context inconsistencies, logical contradictions, narrative embellishments, and incomplete outputs. As multiple mechanisms may contribute to a single hallucination, while early errors can be amplified downstream, effective mitigation requires understanding and addressing the specific mechanisms involved.

This definition explicitly ties external manifestations to internal causes. It emphasizes the many-to-many mapping between mechanisms and phenomena and underscores the importance of lifecycle-aware evaluation and mitigation. By connecting "symptoms" to "root causes," researchers and practitioners can then develop targeted benchmarks, diagnostic evaluation methods, and mechanism-specific mitigation strategies.

VI. Implication

This mechanism-informed definition provides several implications for future research and practice targeting AI hallucination management.

A. Informing Hallucination Evaluation and Mitigation

First, a lifecycle-aware definition provides an organizing axis for mitigating hallucinations. Most current approaches treat hallucination as a black-box problem [59]. Techniques, for example, contradictive learning [32], [119], focus on symptom-level in the output, without systematically examining the internal mechanisms that generate them. This lack of mechanistic grounding significantly limits both the effectiveness and generalization of existing mitigation strategies.

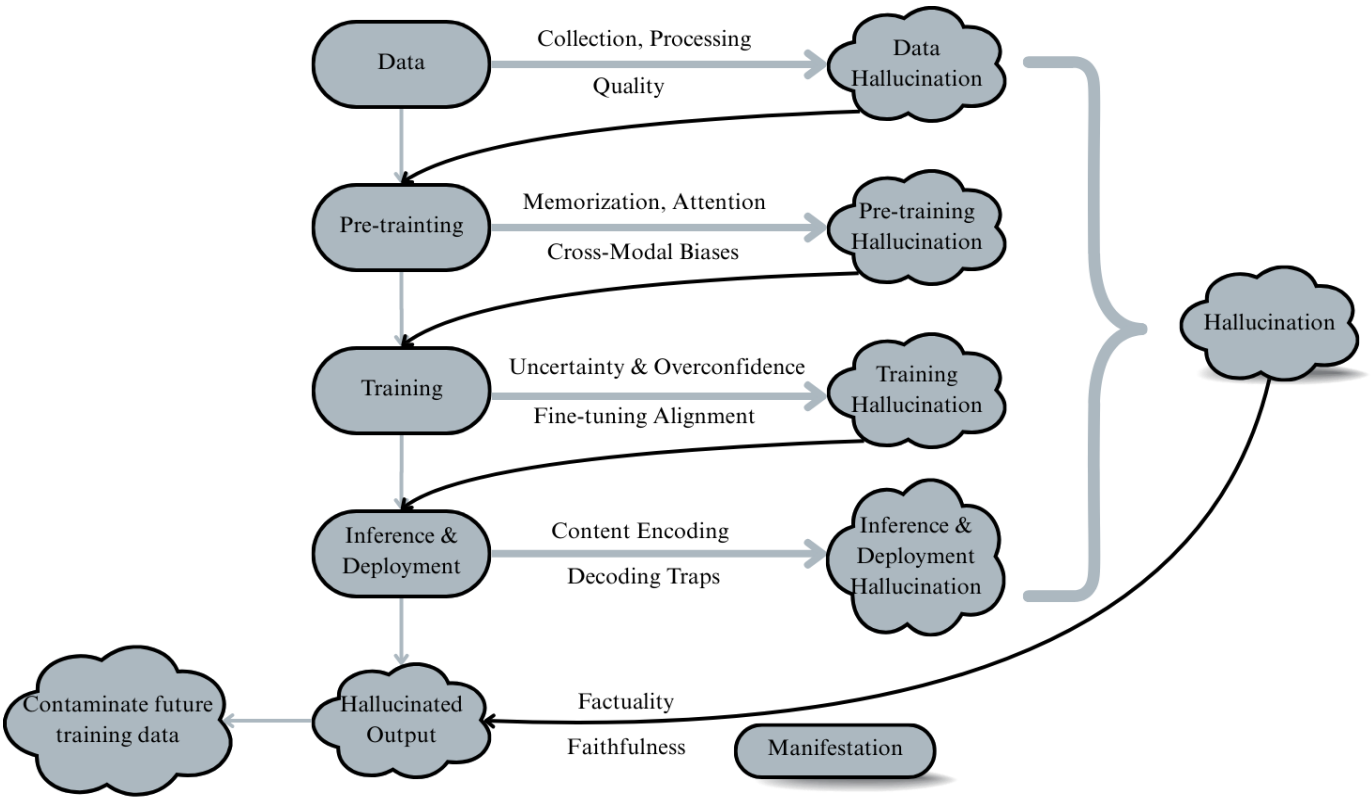

Fig. 5. Mechanistic View of Hallucination

TABLE I. MAPPING BETWEEN MECHANISMS AND OBSERVABLE HALLUCINATION PHENOMENA

Mechanism Stage	Specific Mechanisms	Mapped Phenomena (Observable Hallucination Types)
Data	Collection errors, Processing flaws, Quality issues	**Factual hallucination** (wrong entities, dates, numeric/time errors, geographic/time mix-ups)
Pre-training	Memorization bias, Attention pathologies, Cross-modal biases	**Factual hallucination** (knowledge boundary errors) **Narrative elaboration** (intrinsic/extrinsic embellishment, factual mirage) **Instruction/context inconsistency** (context ignored due to poor attention)
Training	Fine-tuning misalignment, Exposure bias	**Unverifiability / Fabrication** (overclaims, unverifiable statements, outdated facts) **Logical inconsistency** (internal contradictions, semantic distortions)
Inference & Deployment	Content encoding errors, Decoding traps, Long context limitations	**Instruction & context inconsistency** (prompt violations, ignored context, misinterpretation) **Incompleteness** (truncated lists, missing details) **Logical inconsistency** (reasoning limitations, semantic breakdowns)
Feedback Loop	Contaminated future training data	**Amplified factual errors / fabrication** (propagation of prior hallucinations into future model behavior)

By systematically mapping hallucination sources to different stages of the LLM lifecycle, our framework provides the basis for stage-specific and mechanism-sensitive mitigation. Such mapping avoids the pitfalls of black-box treatment and offers an interpretable foundation for mitigation design.

Second, our review highlights the importance of developing detection and evaluation methods that move beyond the phenomenal level of output, which only captures observable manifestations. Most existing methods rely on post-hoc judgments of generated text [17], [120], constraining detection to manifestations that are already visible. Therefore, a mechanism-informed definition opens the way for detecting internal signals of hallucination before outputs are finalized.

For instance, Zhang et al. [60] forecast hallucinations by identifying their internal triggering conditions, while Jiang et al. [64] design a classifier built on logit lens representations that reflects internal state variations and achieves a predictive detection rate with 88% accuracy. These studies illustrate how dynamic internal mechanisms can serve as early warning signals. Building on this understanding, a mechanism-level

definition can further guide the development of proactive detection methods and more generalizable benchmarks.

Finally, this framework reiterates the urgent need for a unified definition of hallucination. Current definitional fragmentation reflects the absence of a common reference point, hindering research efforts by introducing inconsistent terminology and evaluation standards. A mechanism-informed definition could work as the necessary reference point, harmonizing diverse perspectives under a single organizing schema. Future work should therefore focus on operationalizing this unified definition, developing shared benchmarks and metrics grounded in lifecycle mechanisms to ensure consistency across domains and applications.

B. Hallucination from the Perspective of Universal Village

Universal Village (UV) is a new proposed concept exemplifying an ideal future society that pursues harmony between humans and nature through the wise use of technologies. The framework of UV contains UV Elements, UV Design Process, and UV Connectivity at different levels [121]. In this framework, the UV Elements include four basic elements and eight system components. The four basic elements (Development of new energy source, Development of new material, Development of effective microbial technology and environmental protection technologies, UV Lifestyle enabled by Information technology) provide the conceptual foundation [121], while the eight subsystems (Smart Home and Community, Smart Medicine and Healthcare, ITS, Urban Planning and Crowd Management, Smart Energy Management, Smart City Infrastructure, Smart Response System for City Emergency, Smart Environmental Protection, Smart Humanity) translate them into practice across critical domains. Unlike current smart city solutions, which often remain fragmented and rely merely on data collection without embedding closed feedback control loops, these subsystems are incorporated as UV Elements to function as active and feedback-driven units. Their integration highlights that they are not peripheral add-ons but essential components of a unified ecosystem, ensuring robustness, safety, and adaptability through a closed feedback control loop, including data acquisition, communication, decision-making, and action [121]. Moreover, these subsystems not only have close interactions with each other but also are affected by four major impacting factors of smart cities: information flow, material cycle, lifestyle, and community.

Within this context, hallucination can be conceptualized as a data flaw originating from information flow, one of the four factors that impact subsystems. At the subsystem level, hallucination could disrupt the closed feedback loop, leading to erroneous data acquisition, distorted communication, compromised decision-making, and misguided action, which may in turn reinforce the errors. At the system level, these flaws could further spread through mutual interactions, while affecting material cycles, lifestyle, and community, transforming hallucination from an isolated malfunction into a structural vulnerability of the UV ecosystem.

VII. Conclusion

Hallucination remains one of the most persistent challenges in the deployment of generative AI systems. Even if hallucination starts to receive growing attention in public discourse and industry narratives, our analysis demonstrates that the field continues to rely on fragmented and inconsistent definitions that hinder both evaluation and mitigation. By systematically reviewing 76 definitions of hallucination across AI-related literature, we reveal the conceptual ambiguities and practical shortcomings that arise from vocabulary inconsistencies, shifting evaluation baselines, and insufficient grounding in mechanisms. To address these gaps, we introduce a Lifecycle-Based, Mechanism-Aligned, and Phenomenon-Consistent definition of hallucination. This reframed perspective captures not only the observable manifestations of hallucination, such as factuality and faithfulness errors, but also the underlying generative processes across the stages of data collection, model development, and deployment. Our definition also provides a conceptually coherent and practically actionable foundation for subsequent evaluation, diagnosis, and mitigation of hallucination.

References

[1] OpenAI. "Introducing gpt-5." openai.com. https://openai.com/index/introducing-gpt-5/ (accessed Augest 14, 2025).

[2] K. S. Jason Weston. "Blender bot 2.0: An open source chatbot that builds long-term memory and searches the internet." ai.meta.com. https://ai.meta.com/blog/blender-bot-2-an-open-source-chatbot-that-builds-long-term-memory-and-searches-the-internet/ (accessed Augest 14, 2025).

[3] Cambridge. "'hallucinate' is cambridge dictionary's word of the year 2023." www.cambridge.org. https://www.cambridge.org/news-and-insights/hallucinate-is-cambridge-word-of-the-year-2023 (accessed Augest 14, 2025).

[4] Cambridge University Press. "hallucination." dictionary.cambridge.org. https://dictionary.cambridge.org/us/dictionary/english/hallucination (accessed Augest 14, 2025).

[5] W. Davis. "Hospitals use a transcription tool powered by an error-prone openai model." www.theverge.com. https://www.theverge.com/2024/10/27/24281170/open-ai-whisper-hospitals-transcription-hallucinations-studies?utm_source=chatgpt.com (accessed Augest 14, 2025).

[6] V. Magesh *et al.*, "Hallucination-free? assessing the reliability of leading ai legal research tools," *Journal of Empirical Legal Studies*, vol. 22, no. 2, pp. 216–242, 2025.

[7] L. NEUMEISTER. "Lawyers submitted bogus case law created by chatgpt. a judge fined them $5,000." apnews.com. https://apnews.com/article/artificial-intelligence-chatgpt-fake-case-lawyers-d6ae9fa79d0542db9e1455397aef381c (accessed Augest 14, 2025).

[8] S. Merken. "Ai 'hallucinations' in court papers spell trouble for lawyers." www.reuters.com. https://www.reuters.com/technology/artificial-intelligence/ai-hallucinations-court-papers-spell-trouble-lawyers-2025-02-18/ (accessed Augest 14, 2025).

[9] H. Field. "Openai is pursuing a new way to fight a.i. 'hallucinations'." www.cnbc.com. https://www.cnbc.com/2023/05/31/openai-is-pursuing-a-new-way-to-fight-ai-hallucinations.html (accessed Augest 14, 2025).

[10] OpenAI. "Gpt-4o system card." openai.com. https://openai.com/index/gpt-4o-system-card/ (accessed Augest 14, 2025).

[11] C. Fang *et al.*, "Ingest-and-ground: Dispelling hallucinations from continually-pretrained llms with rag," *arXiv preprint arXiv:2410.02825*, 2024.

[12] N. Maleki, B. Padmanabhan, and K. Dutta, "Ai hallucinations: a misnomer worth clarifying," in *2024 IEEE conference on artificial intelligence (CAI)*. IEEE, 2024, pp. 133–138.

[13] A. Ioste, "Hallucinations or attention misdirection? the path to strategic value extraction in business using large language models," *arXiv preprint arXiv:2402.14002*, 2024.

[14] K. v. Deemter, "Hallucination-lm-mechanism:fxy:deemter2024pitfalls," *Computational Linguistics*, vol. 50, no. 2, pp. 807–816, 2024.

[15] F. Leiser *et al.*, "From chatgpt to factgpt: A participatory design study to mitigate the effects of large language model hallucinations on users," in *Proceedings of Mensch Und Computer 2023*, ser. MuC '23. New York, NY, USA: Association for Computing Machinery, 2023, p. 81–90. [Online]. Available: https://doi.org/10.1145/3603555.3603565

[16] J. Li *et al.*, "The dawn after the dark: An empirical study on factuality hallucination in large language models. arxiv, article," *arXiv preprint arXiv:2401.03205*, 2024.

[17] M. Elaraby *et al.*, "Halo: Estimation and reduction of hallucinations in open-source weak large language models," *arXiv preprint arXiv:2308.11764*, 2023.

[18] D. Lei *et al.*, "Chain of natural language inference for reducing large language model ungrounded hallucinations," *arXiv preprint arXiv:2310.03951*, 2023.

[19] X. Shi, Z. Zhu, Z. Zhang, and C. Li, "Hallucination mitigation in natural language generation from large-scale open-domain knowledge graphs," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023, pp. 12 506–12 521.

[20] Z. Ji *et al.*, "Towards mitigating hallucination in large language models via self-reflection," *arXiv preprint arXiv:2310.06271*, 2023.

[21] S. Dhuliawala *et al.*, "Chain-of-verification reduces hallucination in large language models," 2023. [Online]. Available: https://arxiv.org/abs/2309.11495

[22] O. H. Hamid, "Beyond probabilities: Unveiling the delicate dance of large language models (llms) and ai-hallucination," in *2024 IEEE Conference on Cognitive and Computational Aspects of Situation Management (CogSIMA)*. IEEE, 2024, pp. 85–90.

[23] W. Lan *et al.*, "A survey of hallucination in large visual language models," *arXiv preprint arXiv:2410.15359*, 2024.

[24] P. Tupsakhare and N. D. Kulkarni, "Strategies for avoiding gpt hallucinations," *International Research Journal of Modernization in Engineering, Technology and Science*, vol. 6, no. 7, pp. 4131–4136, 2024. [Online]. Available: https://www.researchgate.net/publication/384260920

[25] F. Harrington, E. Rosenthal, and M. Swinburne, "Mitigating hallucinations in large language models with sliding generation and self-checks," *Authorea Preprints*, 2024.

[26] L. Huang *et al.*, "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions," *ACM Transactions on Information Systems*, vol. 43, no. 2, pp. 1–55, 2025.

[27] S. Qi, Y. He, and Z. Yuan, "Can we catch the elephant? the evolvement of hallucination evaluation on natural language generation: A survey," *arXiv e-prints*, pp. arXiv–2404, 2024.

[28] S. Fairburn and J. Ainsworth, "Mitigate large language model hallucinations with probabilistic inference in graph neural networks," 2024.

[29] H. Kang, J. Ni, and H. Yao, "Ever: Mitigating hallucination in large language models through real-time verification and rectification," 2024. [Online]. Available: https://arxiv.org/abs/2311.09114

[30] D. Gosmar and D. A. Dahl, "Hallucination mitigation using agentic ai natural language-based frameworks," *arXiv preprint arXiv:2501.13946*, 2025.

[31] N. Nananukul and M. Kejriwal, "Halo: An ontology for representing hallucinations in generative models," *CoRR*, 2023.

[32] N. Mündler, J. He, S. Jenko, and M. Vechev, "Self-contradictory hallucinations of large language models: Evaluation, detection and mitigation," *arXiv preprint arXiv:2305.15852*, 2023.

[33] M. A. M. Abdelghafour, M. Mabrouk, and Z. Taha, "Hallucination mitigation techniques in large language models," *International Journal of Intelligent Computing and Information Sciences*, vol. 24, no. 4, pp. 73–81, 2024.

[34] H. Ye, T. Liu, A. Zhang, W. Hua, and W. Jia, "Cognitive mirage: A review of hallucinations in large language models," *arXiv preprint arXiv:2309.06794*, 2023.

[35] Z. Ji *et al.*, "Survey of hallucination in natural language generation," *ACM Computing Surveys*, vol. 55, no. 12, pp. 1–38, 2023.

[36] R. S. El-Mallakh and K. L. Walker, "Hallucinations, psuedohallucinations, and parahallucinations," *Psychiatry: Interpersonal and Biological Processes*, vol. 73, no. 1, pp. 34–42, 2010.

[37] T. Stening. "What are ai chatbots actually doing when they 'hallucinate'? here's why experts don't like the term." news.northeastern.edu. https://news.northeastern.edu/2023/11/10/ai-chatbot-hallucinations/ (accessed July 29, 2025).

[38] A. Gundogmusler, F. Bayindiroglu, and M. Karakucukoglu, "Mathematical foundations of hallucination in transformer-based large language models for improvisation," *Authorea Preprints*, 2024.

[39] Z. Xu, S. Jain, and M. Kankanhalli, "Hallucination is inevitable: An innate limitation of large language models," *arXiv preprint arXiv:2401.11817*, 2024.

[40] Q. Chen and B. Wang, "Valuable hallucinations: Realizable non-realistic propositions," *arXiv preprint arXiv:2502.11113*, 2025. [Online]. Available: https://arxiv.org/abs/2502.11113

[41] P. Sui, E. Duede, S. Wu, and R. So, "Confabulation: The surprising value of large language model hallucinations," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Bangkok, Thailand: Association for Computational Linguistics, August 2024, pp. 14 274–14 284. [Online]. Available: https://aclanthology.org/2024.acl-long.770/

[42] W. Witkowski. "Openai's sam altman tells salesforce's marc benioff that ai 'hallucinations' are more feature than bu." www.marketwatch.com. https://www.marketwatch.com/story/openais-sam-altman-tells-salesforces-marc-benioff-that-ai-hallucinations-are-more-feature-than-bug-1c035c52 (accessed July 29, 2025).

[43] Z. He, B. Zhang, and L. Cheng, "Shakespearean sparks: The dance of hallucination and creativity in llms' decoding layers," *arXiv preprint arXiv:2503.02851*, 2025.

[44] M. Lee, "A mathematical investigation of hallucination and creativity in gpt models," *Mathematics*, vol. 11, no. 10, p. 2320, 2023.

[45] X. Jiang *et al.*, "A survey on large language model hallucination via a creativity perspective," *arXiv preprint arXiv:2402.06647*, 2024.

[46] B. A. Huberman and S. Mukherjee, "Hallucinations and emergence in large language models," *Available at SSRN 4676180*, 2023.

[47] S. Lin, J. Hilton, and O. Evans, "Truthfulqa: Measuring how models mimic human falsehoods," *arXiv preprint arXiv:2109.07958*, 2021.

[48] D. Muhlgay *et al.*, "Generating benchmarks for factuality evaluation of language models," *arXiv preprint arXiv:2307.06908*, 2023.

[49] A. M. Rush, S. Chopra, and J. Weston, "A neural attention model for abstractive sentence summarization," *arXiv preprint arXiv:1509.00685*, 2015.

[50] A. R. Fabbri, I. Li, T. She, S. Li, and D. R. Radev, "Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model," *arXiv preprint arXiv:1906.01749*, 2019.

[51] H. Zhou, C. Zheng, K. Huang, M. Huang, and X. Zhu, "Kdconv: A chinese multi-domain dialogue dataset towards multi-turn knowledge-driven conversation," *arXiv preprint arXiv:2004.04100*, 2020.

[52] P. Kaul *et al.*, "Throne: An object-based hallucination benchmark for the free-form generations of large vision-language models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 27 228–27 238.

[53] K. Andriopoulos and J. Pouwelse, "Augmenting llms with knowledge: A survey on hallucination prevention," *arXiv preprint arXiv:2309.16459*, 2023.

[54] Y. Li *et al.*, "Evaluating object hallucination in large vision-language models," *arXiv preprint arXiv:2305.10355*, 2023.

[55] H. Davies, "From i-ching to ai: Interrogating digital divination," in *Proceedings of International Symposium on Electronic Art (ISEA)*, 2024.

[56] OpenAI. "Writing with ai." openai.com. https://openai.com/chatgpt/use-cases/writing-with-ai/ (accessed July 29, 2025).

[57] P. Mirowski, K. W. Mathewson, J. Pittman, and R. Evans, "Co-writing screenplays and theatre scripts with language models: Evaluation by industry professionals," in *Proceedings of the 2023 CHI conference on human factors in computing systems*, 2023, pp. 1–34.

[58] S. Yuan and M. Färber, "Hallucinations can improve large language models in drug discovery," *arXiv preprint arXiv:2501.13824*, 2025.

[59] L. Yu, M. Cao, J. C. K. Cheung, and Y. Dong, "Mechanistic understanding and mitigation of language model non-factual hallucinations," *arXiv preprint arXiv:2403.18167*, 2024.

[60] Y. Zhang *et al.*, "Knowledge overshadowing causes amalgamated hallucination in large language models," *arXiv preprint arXiv:2407.08039*, 2024.

[61] Y. Qin *et al.*, "Knowledge inheritance for pre-trained language models," *arXiv preprint arXiv:2105.13880*, 2021.

[62] M. Li *et al.*, "Knowledge boundary of large language models: A survey," *arXiv preprint arXiv:2412.12472*, 2024.

[63] E. Stringhi, "Hallucinating (or poorly fed) llms? the problem of data accuracy," *i-lex*, vol. 16, no. 2, pp. 54–63, 2023.

[64] C. Jiang *et al.*, "On large language models' hallucination with regard to known facts," *arXiv preprint arXiv:2403.20009*, 2024.

[65] X. Du, C. Xiao, and S. Li, "Haloscope: Harnessing unlabeled llm generations for hallucination detection," *Advances in Neural Information Processing Systems*, vol. 37, pp. 102 948–102 972, 2024.

[66] J. Oh *et al.*, "Erbench: An entity-relationship based automatically verifiable hallucination benchmark for large language models," *Advances in Neural Information Processing Systems*, vol. 37, pp. 53 064–53 101, 2024.

[67] H. Nguyen *et al.*, "Smoothing out hallucinations: Mitigating llm hallucination with smoothed knowledge distillation," 2025. [Online]. Available: https://arxiv.org/abs/2502.11306

[68] A. Ravichander, S. Ghela, D. Wadden, and Y. Choi, "Halogen: Fantastic llm hallucinations and where to find them," *arXiv preprint arXiv:2501.08292*, 2025.

[69] H. Cao *et al.*, "Mitigating knowledge conflicts in language model-driven question answering," *arXiv preprint arXiv:2411.11344*, 2024.

[70] Z. Yin, "A review of methods for alleviating hallucination issues in large language models," *Applied and Computational Engineering*, vol. 76, pp. 258–266, 2024.

[71] M. Yin *et al.*, "Entropy law: The story behind data compression and llm performance," *arXiv preprint arXiv:2407.06645*, 2024.

[72] J. Kim, Y. Kim, and Y. M. Ro, "What if...?: Counterfactual inception to mitigate hallucination effects in large multimodal models," *CoRR*, 2024.

[73] N. Varshney *et al.*, "Investigating and addressing hallucinations of llms in tasks involving negation," *arXiv preprint arXiv:2406.05494*, 2024.

[74] H. You *et al.*, "Ferret: Refer and ground anything anywhere at any granularity," *arXiv preprint arXiv:2310.07704*, 2023.

[75] J. Zhang *et al.*, "Eventhallusion: Diagnosing event hallucinations in video llms," *arXiv preprint arXiv:2409.16597*, 2024.

[76] N. McKenna *et al.*, "Sources of hallucination by large language models on inference tasks," *arXiv preprint arXiv:2305.14552*, 2023.

[77] U. Kamath, K. Keenan, G. Somers, and S. Sorenson, "Llm challenges and solutions," in *Large Language Models: A Deep Dive: Bridging Theory and Practice*. Springer, 2024, pp. 219–274.

[78] L. Du *et al.*, "Quantifying and attributing the hallucination of large language models via association analysis," *arXiv preprint arXiv:2309.05217*, 2023.

[79] M. Wang, M. Suzuki, H. Sakaji, and K. Izumi, "Interactive dualchecker for mitigating hallucinations in distilling large language models," *arXiv preprint arXiv:2408.12326*, 2024.

[80] Q. Luo *et al.*, "Self-attention and transformers: Driving the evolution of large language models," in *2023 IEEE 6th International conference on electronic information and communication technology (ICEICT)*. IEEE, 2023, pp. 401–405.

[81] S. Serrano and N. A. Smith, "Is attention interpretable?" *arXiv preprint arXiv:1906.03731*, 2019.

[82] X. Gong, T. Ming, X. Wang, and Z. Wei, "Damro: Dive into the attention mechanism of lvlm to reduce object hallucination," *arXiv preprint arXiv:2410.04514*, 2024.

[83] L. Kuhn, Y. Gal, and S. Farquhar, "Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation," *arXiv preprint arXiv:2302.09664*, 2023.

[84] H. Yuan *et al.*, "Whispers that shake foundations: Analyzing and mitigating false premise hallucinations in large language models," 2024. [Online]. Available: https://arxiv.org/abs/2402.19103

[85] M. Hahn, "Theoretical limitations of self-attention in neural sequence models," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 156–171, 2020.

[86] D. Chiang and P. Cholak, "Overcoming a theoretical limitation of self-attention," *arXiv preprint arXiv:2202.12172*, 2022.

[87] Z. Wei, S. Wang, X. Rong, X. Liu, and H. Li, "Shadows in the attention: Contextual perturbation and representation drift in the dynamics of hallucination in llms," *arXiv preprint arXiv:2505.16894*, 2025.

[88] V. Rawte *et al.*, "" sorry, come again?" prompting–enhancing comprehension and diminishing hallucination with [pause]-injected optimal paraphrasing," *arXiv preprint arXiv:2403.18976*, 2024.

[89] N. F. Liu *et al.*, "Lost in the middle: How language models use long contexts," *arXiv preprint arXiv:2307.03172*, 2023.

[90] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *International conference on machine learning*. PMLR, 2023, pp. 19 730–19 742.

[91] Z. Fan, Y. Wang, S. Polisetty, and Y. R. Fung, "Unveiling the lack of lvlm robustness to fundamental visual variations: Why and path forward," *arXiv preprint arXiv:2504.16727*, 2025.

[92] A. Radford *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PmLR, 2021, pp. 8748–8763.

[93] Y. Yuan, P. Gao, Q. Dai, J. Qin, and W. Xiang, "Uncertainty guided refinement for fine-grained salient object detection," *IEEE Transactions on Image Processing*, 2025.

[94] J. He *et al.*, "Cracking the code of hallucination in lvlms with vision-aware head divergence," *arXiv preprint arXiv:2412.13949*, 2024.

[95] P.-H. Huang, J.-L. Li, C.-P. Chen, M.-C. Chang, and W.-C. Chen, "Who brings the frisbee: Probing hidden hallucination factors in large vision-language model via causality analysis," in *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2025, pp. 6125–6135.

[96] A. Rohrbach, L. A. Hendricks, K. Burns, T. Darrell, and K. Saenko, "Object hallucination in image captioning," *arXiv preprint arXiv:1809.02156*, 2018.

[97] X. Wu *et al.*, "Autohallusion: Automatic generation of hallucination benchmarks for vision-language models," 2024.

[98] J. Li *et al.*, "Fine-grained semantically aligned vision-language pre-training," *Advances in neural information processing systems*, vol. 35, pp. 7290–7303, 2022.

[99] A. Kumar, R. Morabito, S. Umbet, J. Kabbara, and A. Emami, "Confidence under the hood: An investigation into the confidence-probability alignment in large language models," *arXiv preprint arXiv:2405.16282*, 2024.

[100] Y. Sun *et al.*, "Benchmarking hallucination in large language models based on unanswerable math word problem," *arXiv preprint arXiv:2403.03558*, 2024.

[101] X. Chen, L. Wang, W. Wu, Q. Tang, and Y. Liu, "Honest ai: Fine-tuning "small" language models to say "i don't know", and reducing hallucination in rag," 2024. [Online]. Available: https://arxiv.org/abs/2410.09699

[102] Y. A. Yadkori *et al.*, "Mitigating llm hallucinations via conformal abstention," *arXiv preprint arXiv:2405.01563*, 2024.

[103] M. Zhang, O. Press, W. Merrill, A. Liu, and N. A. Smith, "How language model hallucinations can snowball," *arXiv preprint arXiv:2305.13534*, 2023.

[104] Y. Xu, H. Dong, L. Wang, C. Xiong, and J. Li, "Reward models identify consistency, not causality," *arXiv preprint arXiv:2502.14619*, 2025.

[105] M. Hu *et al.*, "Mitigating large language model hallucination with faithful finetuning," *arXiv preprint arXiv:2406.11267*, 2024.

[106] C. Wang and R. Sennrich, "On exposure bias, hallucination and domain shift in neural machine translation," *arXiv preprint arXiv:2005.03642*, 2020.

[107] T. He, J. Zhang, Z. Zhou, and J. Glass, "Exposure bias versus self-recovery: Are distortions really incremental for autoregressive text generation?" *arXiv preprint arXiv:1905.10617*, 2019.

[108] Z. Gekhman *et al.*, "Does fine-tuning llms on new knowledge encourage hallucinations?" *arXiv preprint arXiv:2405.05904*, 2024.

[109] H. Wu *et al.*, "Iter-ahmcl: Alleviate hallucination for large language model via iterative model-level contrastive learning," *arXiv preprint arXiv:2410.12130*, 2024.

[110] W. Ren, X. Li, L. Wang, T. Zhao, and W. Qin, "Analyzing and reducing catastrophic forgetting in parameter efficient tuning," *arXiv preprint arXiv:2402.18865*, 2024.

[111] W. Liu, W. Zeng, K. He, Y. Jiang, and J. He, "What makes good data for alignment? a comprehensive study of automatic data selection in instruction tuning," *arXiv preprint arXiv:2312.15685*, 2023.

[112] H. Rashkin, D. Reitter, G. S. Tomar, and D. Das, "Increasing faithfulness in knowledge-grounded dialogue with controllable features," *arXiv preprint arXiv:2107.06963*, 2021.

[113] W. Shi *et al.*, "Trusting your evidence: Hallucinate less with context-aware decoding," in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, 2024, pp. 783–791.

[114] C. Shi *et al.*, "A thorough examination of decoding methods in the era of llms," *arXiv preprint arXiv:2402.06925*, 2024.

[115] H. Zhang, D. Duckworth, D. Ippolito, and A. Neelakantan, "Trading off diversity and quality in natural language generation," *arXiv preprint arXiv:2004.10450*, 2020.

[116] J. Maynez, S. Narayan, B. Bohnet, and R. McDonald, "On faithfulness and factuality in abstractive summarization," *arXiv preprint arXiv:2005.00661*, 2020.

[117] T. Liu *et al.*, "A token-level reference-free hallucination detection benchmark for free-form text generation," 2022. [Online]. Available: https://arxiv.org/abs/2104.08704

[118] G. Noarov *et al.*, "Foundations of top-k decoding for language models," *arXiv preprint arXiv:2505.19371*, 2025.

[119] W. Sun *et al.*, "Contrastive learning reduces hallucination in conversations," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 11, 2023, pp. 13 618–13 626.

[120] P. Manakul, A. Liusie, and M. J. Gales, "Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models," *arXiv preprint arXiv:2303.08896*, 2023.

[121] S. Cao *et al.*, "Preliminary study on evaluation of smart-cities technologies and proposed uv lifestyles," in *2018 4th International Conference on Universal Village (UV)*. Ieee, 2018, pp. 1–49.