American Journal of Artificial Intelligence
2026, Vol. 10, No. 1, pp. 61–70
https://doi.org/10.11648/j.ajai.20261001.16

Research Article

Understanding Model Hallucinations: Causes, Mitigation Strategies, and Evaluation Metrics for Detection

Diarah Reuben Samuel[1, *], Adekunel Adefemi Aderemi[2], Osueke Christian Okechukwu[1], Onu Peter[3], Diarah Ifeyinwa Sandra[4], Ozichi Emuoyibofarhe[5], Olaomi Bimpe Agnes[6], Evoh Edwin Emeng[2]

[1]Mechatronics Engineering Programme, Bowen University, Iwo, Nigeria

[2]Mechatronics Engineering Department, Federal University, Oye Ekiti, Nigeria

[3]Mechanical/Mechatronics Engineering Department, Pan Atinatic University, Lagos, Nigeria

[4]Accounting and Finance Programme, Bowen University, Iwo, Nigeria

[5]Computer Science Programme, Bowen University, Iwo, Nigeria

[6]Ladoke Akintola Universityof Technology, Ogbomoso, Nigeria

Abstracts

Foundation models (FMs) have the potential to revolutionize various fields, but their reliability is often compromised by hallucinations. This paper delves into the intricate nature of model hallucinations, exploring their root causes, mitigation strategies, and evaluation metrics. We provide a comprehensive overview of the challenges posed by hallucinations, including factual inaccuracies, logical inconsistencies, and the generation of fabricated content. To address these issues, we discuss a range of techniques, such as improving data quality, refining model architectures, and employing advanced prompting techniques. We also highlight the importance of developing robust evaluation metrics to detect and quantify hallucinations. By understanding the underlying mechanisms and implementing effective mitigation strategies, we can unlock the full potential of FMs and ensure their reliable and trustworthy operation. Foundation Models (FMs), such as large language models and multimodal transformers, have demonstrated transformative capabilities across a wide range of applications in artificial intelligence, including natural language processing, computer vision, and decision support systems. Despite their remarkable success, the reliability and trustworthiness of these models are frequently undermined by a phenomenon known as hallucination, the generation of outputs that are factually incorrect, logically inconsistent, or entirely fabricated. This study presents a comprehensive examination of model hallucinations, focusing on their underlying causes, mitigation approaches, and evaluation metrics for systematic detection. We begin by analyzing the root causes of hallucination, which span data-related factors such as bias, noise, and imbalance, as well as architectural and training issues like over-parameterization, poor generalization, and the lack of grounded reasoning. The paper categorizes hallucinations into factual, logical, and contextual types, illustrating how each arises in different stages of model inference and decision-making. We further discuss how prompt engineering, attention misalignment, and inadequate fine-tuning contribute to the persistence of erroneous model outputs. To mitigate these challenges, we explore a range of strategies, including improving data curation and preprocessing pipelines, integrating factual verification and retrieval-augmented mechanisms, and refining model architectures to enhance interpretability and context awareness. Techniques such as reinforcement learning with human feedback (RLHF), chain-of-thought prompting, and hybrid symbolic-neural approaches are highlighted for their potential in reducing hallucination rates while maintaining model fluency and adaptability. Furthermore,

*Correspondence: Diarah Reuben Samuel (diarah.samuel@bowen.edu.ng)

Received: 21 October 2025; **Accepted:** 3 November 2025; **Published:** 2 February 2026

this work emphasizes the critical need for rigorous and standardized evaluation metrics capable of quantifying the severity, frequency, and impact of hallucinations. Metrics such as factual consistency scores, semantic similarity indices, and hallucination detection benchmarks are discussed as essential tools for assessing model reliability. Ultimately, this paper provides a structured understanding of model hallucinations as both a technical and ethical challenge in the deployment of Foundation Models. By elucidating their origins and presenting practical mitigation frameworks, we aim to advance the development of more transparent, accountable, and trustworthy AI systems. The insights presented herein contribute to ongoing efforts to ensure that Foundation Models not only achieve high performance but also uphold factual integrity and user trust across real-world applications.

Keywords

Hallucinations, Misleading Information, AI Model, Foundation Models

1. Introduction

Model hallucination occurs when an AI model generates incorrect or misleading information, often diverging from real-world facts or the given prompt. This phenomenon is particularly prevalent in large language models (LLMs)." AI hallucination, a phenomenon where AI models produce incorrect or misleading information, can arise from several factors. These include inadequate training data, flawed assumptions, biases inherent in the training data, and limitations in the model's understanding of context and nuance.

Foundation models, exemplified by GPT-3 and Stable Diffusion, have ushered in a new era of machine learning and generative AI. These models, trained on massive, diverse datasets, are capable of handling a wide range of tasks, including language understanding, text and image generation, and natural language conversation. This paradigm shift has the potential to revolutionize various industries and applications [18].

1.1. Causes of Hallucination

Hallucinations occur when foundation models (FMs) generate plausible yet factually incorrect or nonsensical content. These outputs, ranging from minor inaccuracies to entirely fabricated information, can manifest across various modalities, including text, images, videos, and audio.

Several factors contribute to hallucinations, such as biases in training data, limited access to current information, and inherent model limitations in comprehending context and generating coherent responses. Deploying such powerful models without addressing hallucinations can lead to the dissemination of misinformation, incorrect conclusions, and potentially harmful consequences in critical applications.

To mitigate hallucinations, researchers are actively exploring strategies like fine-tuning models with domain-specific data, using diverse and robust training data, and developing improved evaluation metrics to identify and reduce hallucination tendencies [26].

1.2. Importance of Hallucination

Although model hallucinations pose a significant challenge to AI, they also offer valuable insights into the limitations of current AI technology. By understanding the root causes of hallucinations, researchers can identify areas for improvement and develop more robust and reliable AI systems.

Recent years have witnessed a surge of interest in foundation models (FMs) across both academia and industry. One of the primary challenges associated with FMs is hallucination, the generation of plausible but factually incorrect or nonsensical content.

While previous surveys have explored hallucination in natural language generation [6] and large language models (LLMs) specifically [30]-the issue extends to other modalities, including image, video, and audio. This paper aims to provide a comprehensive survey of hallucination across all major modalities of foundation models.

This article seeks to deliver an in-depth exploration of model hallucinations within artificial intelligence systems, a phenomenon that significantly undermines the reliability and credibility of machine learning applications. We will delve into the various factors that lead to these hallucinations, such as biases present in the training data, limitations inherent in model architectures, and the methodologies employed during training. By examining these elements, we aim to clarify the contexts in which hallucinations are more likely to manifest. Furthermore, we will conduct a thorough review of existing strategies aimed at mitigating hallucinations, evaluating the efficacy of approaches like enhanced data curation practices, modifications in model design, and effective post-processing techniques, thereby providing practical recommendations for industry practitioners. In addition, we will introduce a framework for establishing robust evaluation metrics that can effectively detect and measure the extent of hallucination in different models, highlighting the necessity for standardized metrics that enable meaningful comparisons across various applications. Ultimately, this comprehensive survey aspires to shed light on the intricate nature of model hallucinations, furnishing researchers and practitioners with essential insights that can foster the development of more reliable and trustworthy artificial intelligence technologies.

1.3. Hallucination Categories

In Figure 1, we categorize foundation models (FMs) into four primary modalities: text, image, video, and audio. Foundation models, despite their remarkable capabilities, can sometimes produce hallucinations. This phenomenon occurs when the model generates text that deviates from factual accuracy, presenting fictional or misleading information. This can happen because the model, trained on vast datasets, learns patterns and generates text that sounds plausible but may not align with reality.

Several factors can contribute to hallucinations, including biases present in the training data, the model's limited access to real-time information, and inherent constraints in its ability to comprehend and generate contextually accurate responses. It's crucial to recognize that these instances are often unintentional and stem from the model's limitations rather than a deliberate attempt to deceive.

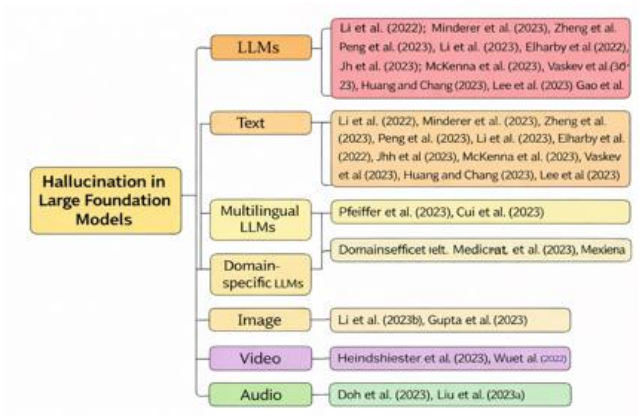

Figure 1. *Taxonomy for Hallucination in large foundation models. (Vipula et al. (2023).*

1.4. Mitigation of Hallucination

The lack of a standardized metric for evaluating object hallucination has hindered progress in understanding and addressing this issue. To fill this gap, [4] introduced NOPE (Negative Object Presence Evaluation), a novel benchmark designed to assess object hallucination in vision-language (VL) models through visual question answering (VQA). Leveraging LLMs, the study generated 29.5k synthetic negative pronoun (NegP) data for NOPE. This dataset was used to extensively evaluate the performance of 10 VL models in discerning the absence of objects in visual questions, alongside their standard performance on visual questions across nine other VQA datasets.

1.5. Importance of Hallucination

Model hallucinations, while potentially frustrating, can be harnessed for creative innovation, problem-solving, and research advancement when used responsibly and critically evaluated. [21] offers an intriguing perspective on the potential of hallucinating models as collaborative creative partners. By generating outputs that may not be strictly factual, these models can spark innovative thinking and unconventional solutions. This creative use of hallucination can lead to unexpected and valuable outcomes.

However, it's essential to distinguish between constructive and harmful hallucinations. Factually inaccurate or norm-violating outputs can be problematic, especially when individuals rely on LLMs for expert knowledge. Conversely, in creative domains, the ability to generate surprising and unconventional responses can stimulate novel ideas and connections.

2. Related Work

Researchers have explored self-contradictory hallucinations in LLMs, where the model generates text that contradicts itself, leading to unreliable or nonsensical outputs [15]. Their work presents methods to evaluate the occurrence of such hallucinations, detect them in LLM-generated text, and mitigate their impact to enhance the overall quality and trustworthiness of LLM-generated content.

SELFCHECKGPT (Potsawee Manakul 2023) is a zero-resource, black-box technique designed to detect hallucinations in generative LLMs. This method identifies instances where these models generate inaccurate or unverified information without requiring additional resources or labelled data. By detecting and addressing hallucinations, SELFCHECKGPT aims to improve the trustworthiness and reliability of LLMs.

PURR [2] is a method designed to efficiently edit and correct hallucinations in language models. By leveraging denoising language model corruptions, PURR effectively identifies and rectifies these hallucinations. This approach aims to improve the quality and accuracy of language model outputs by reducing the incidence of hallucinated content.

Literature offers a zero-resource, black-box approach to detect hallucinations in any LLM without requiring external resources. This method leverages the principle that an LLM with domain expertise will generate consistent and coherent responses. Conversely, randomly sampled responses from unfamiliar topics are more likely to contain contradictory and hallucinated information [16].

Shiping Yang (2023) introduced a novel self-check approach based on reverse validation to automatically identify factual errors in LLMs without relying on external resources. They also created a benchmark, Passage-level Hallucination Detection (PHD), using ChatGPT and human expert annotations to evaluate different methods. Assessing the accuracy of long-form text generated by LLMs is challenging due to the frequent intermingling of accurate and inaccurate information, making simple quality judgments insufficient.

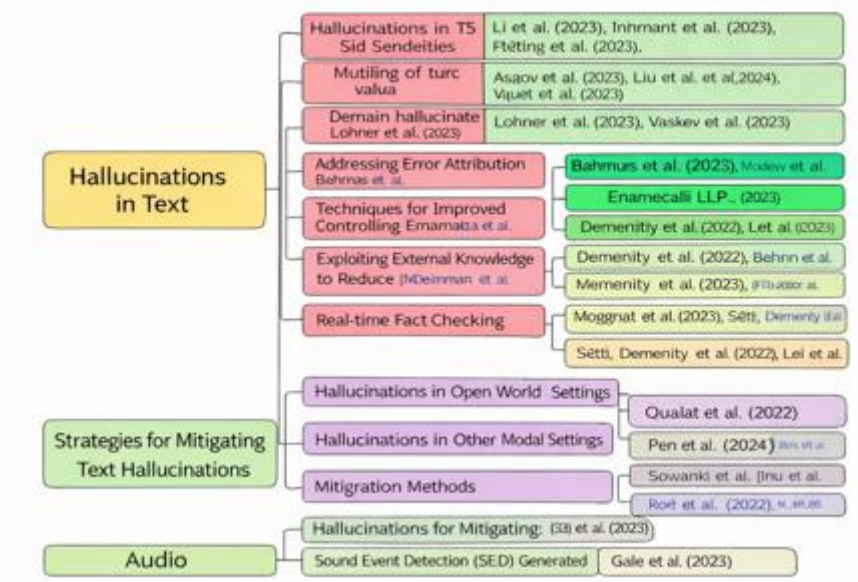

Figure 2. *Taxonomy of hallucination in large foundation models, organised around detection and mitigation techniques [21].*

To address this challenge, [14] introduced FACTSCORE, a novel evaluation metric that assesses the factual accuracy of text by breaking it down into individual claims and evaluating their reliability. [8] proposed a novel approach to mitigate hallucination in LLMs by drawing inspiration from web systems. They highlighted the lack of a "citation" mechanism in LLMs, which involves acknowledging or referencing sources of information, as a critical gap.

To address the challenge of identifying factual inaccuracies in LLM-generated content, [27] developed a multi-task learning (MTL) framework. This framework leverages advanced long text embeddings like e5-mistral-7b-instruct, along with models such as GPT-3, SpanBERT, and RoFormer. The MTL approach demonstrated a 40% average improvement in accuracy on the FACTOID benchmark compared to leading textual entailment methods.

To address this challenge, [31] introduced a formal framework defining hallucination as inconsistencies between computable LLMs and a ground truth function. The study examines existing hallucination mitigation strategies and their practical implications for real-world LLM deployment through this framework.

Introduced the Sorry, Come Again (SCA) prompting technique to mitigate hallucination in contemporary LLMs. SCA enhances comprehension by optimally paraphrasing prompts and injecting [PAUSE] tokens to delay LLM generation. The study analyzes linguistic nuances in prompts and their impact on hallucinated generation, highlighting the challenges posed by prompts with lower readability, formality, or concreteness.

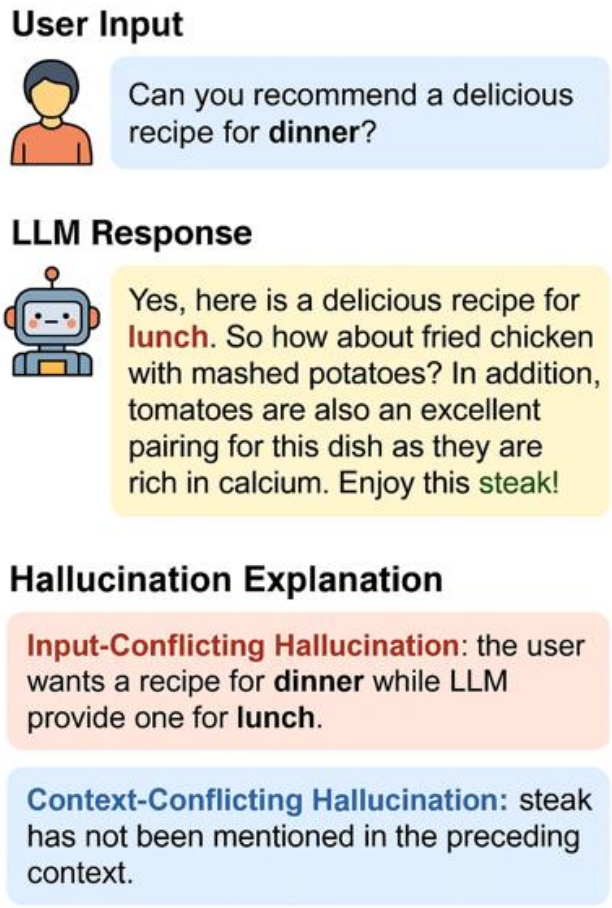

Figure 3. *LLM responses showing the types of hallucinations, highlighted in red, green, and blue [21].*

3. Methodology

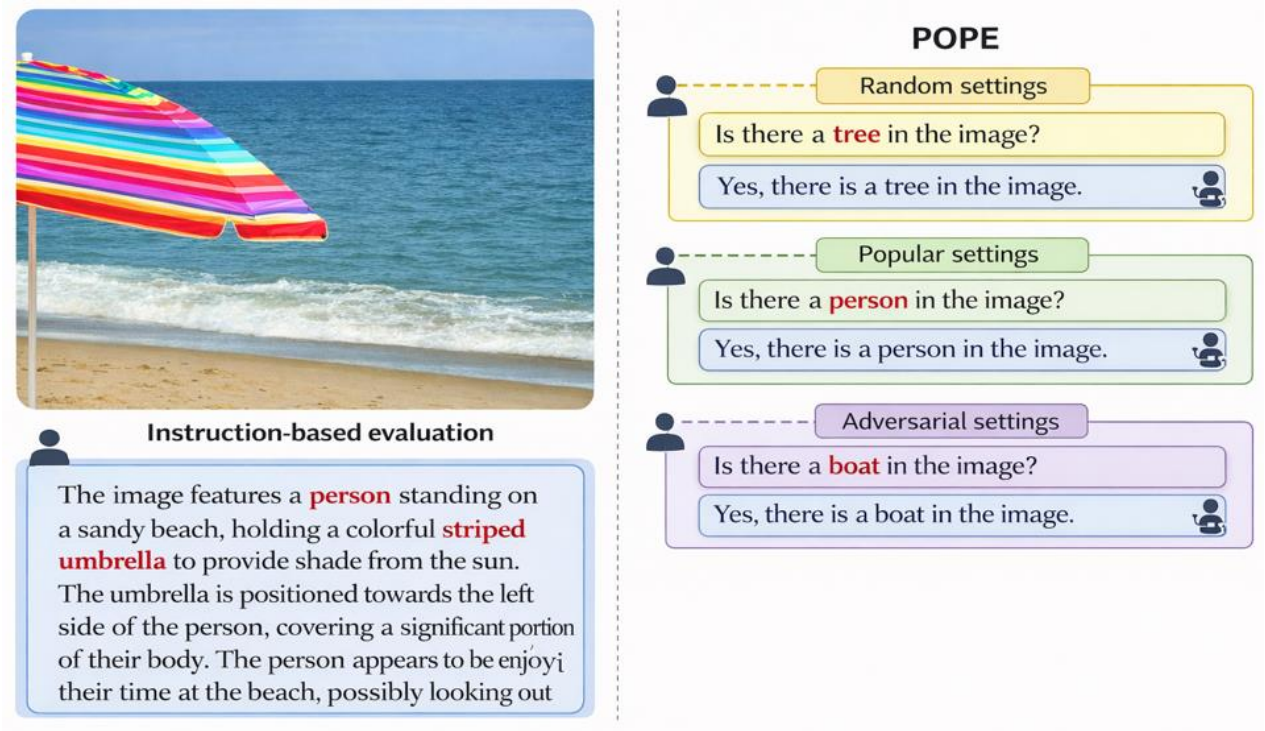

Figure 4. *Instances of object hallucination within LVLMs. [28]. Ground-truth objects in annotations are indicated in bold, while red objects represent hallucinated objects by LVLMs. The left case occurs in the conventional instruction-based evaluation approach, while the right case occur in three variations of POPE [21].*

3.1. Hallucination in Large Image and Video Model

Dense video captioning, the task of generating descriptions for multiple events within a continuous video, demands a deep understanding of video content and strong contextual reasoning. However, this complex task is fraught with challenges, often leading to inaccuracies and hallucinations [22] and [13].

Large video models (LVMs) represent a significant advancement, enabling large-scale processing of video data. Despite their potential for diverse applications like video understanding and generation, LVMs are prone to hallucinations, where misinterpretations of video frames can lead to the generation of artificial or inaccurate visual data. Figure 5 showcases examples of these hallucinations observed in LVMs.

Contrastive learning models, often employing a Siamese architecture [23, 24] have demonstrated significant potential in self-supervised learning. However, their effectiveness hinges on the availability of a sufficient number of diverse positive pairs. Without this, these models may struggle to learn meaningful semantic distinctions and succumb to overfitting.

To address this limitation, the Hallucinator was introduced, which is a novel approach that efficiently generates additional positive samples to enhance contrast. The Hallucinator operates directly in the feature space, making it differentiable and seamlessly integrable into the pretraining task with minimal computational overhead. (Jing Wu n.d.)

To address these challenges, researchers have proposed various innovative approaches. [10] introduced a novel framework that leverages event sequence generation and sequential video captioning, trained with reinforcement learning and two-level rewards. This approach effectively captures contextual information, resulting in more coherent and accurate captions.

A weakly supervised, model-based factuality metric termed *FactVC* was introduced, outperforming existing methods for factuality evaluation [3]. To further support research in this area, two annotated datasets for assessing the factuality of video captions were also released. Literature proposed a context-aware model that incorporates information from past and future events to influence the description of the current event. By leveraging a robust pre-trained context encoder and a gate-attention mechanism, this model significantly outperforms existing context-aware and pre-trained models on the YouCookII and ActivityNet datasets (Weilun Wu and Yang Gao. n.d.).

A streaming model incorporating a memory module for

processing long video sequences, together with a streaming decoding algorithm for early prediction, was introduced, demonstrating significant performance improvements on dense video captioning benchmarks such as ActivityNet, YouCook2, and ViTT [25, 31].

3.2. Detecting and Preventing AI Hallucinations

Researchers have highlighted the issue of object hallucinations in Vision-Language Pre-training (VLP) models, where generated descriptions can contain non-existent or inaccurate objects. [28, 29] further underscored the severity of this problem, suggesting that visual instructions can influence hallucination. They introduced POPE, a polling-based query method, to improve the evaluation of object hallucination.

To address the lack of a standardized metric for assessing object hallucination, [4] introduced NOPE, a novel benchmark for evaluating object hallucination in vision-language (VL) models through visual question answering (VQA). Leveraging LLMs, they generated 29.5k synthetic negative pronoun (NegP) data for NOPE. This dataset was used to extensively evaluate the performance of 10 VL models in discerning the absence of objects in visual questions, alongside their standard performance on visual questions across nine other VQA datasets.

Leveraging LLMs, the study generated 29.5k synthetic negative pronoun (NegP) data for NOPE. This dataset was used to extensively evaluate the performance of 10 VL models in discerning the absence of objects in visual questions, alongside their standard performance on visual questions across nine other VQA datasets.

While most existing research has primarily focused on object hallucinations, [6] delved deeper into Intrinsic Vision-Language Hallucination (IVL-Hallu). They proposed several novel IVL-Hallu tasks, categorizing them into four types: attribute, object, multi-modal conflicting, and counter-common-sense hallucinations. To assess and explore IVL-Hallu, they introduced a challenging benchmark dataset and conducted experiments on five LVLMs, revealing their limitations in effectively addressing these tasks.

To mitigate object hallucination in LVLMs without relying on costly training or APIs, [25], and [26] introduced MARINE. This training-free and API-free approach enhances the visual understanding of LVLMs by integrating existing open-source vision models and utilizing guidance without classifiers to integrate object grounding features, thereby improving the precision of the generated outputs. Evaluations across six LVLMs reveal MARINE's effectiveness in reducing hallucinations and enhancing output detail, validated through assessments using GPT-4V.

[1] introduced a CLIP-Guided Decoding (CGD) training-free approach to reduce object hallucination at decoding time.

HalluciDoctor [17] tackled hallucinations in Multi-modal Large Language Models (MLLMs) by using human error detection to identify and eliminate various types of hallucinations. By rebalancing data distribution via counterfactual visual instruction expansion, they successfully mitigated 44.6% of hallucinations while maintaining competitive performance.

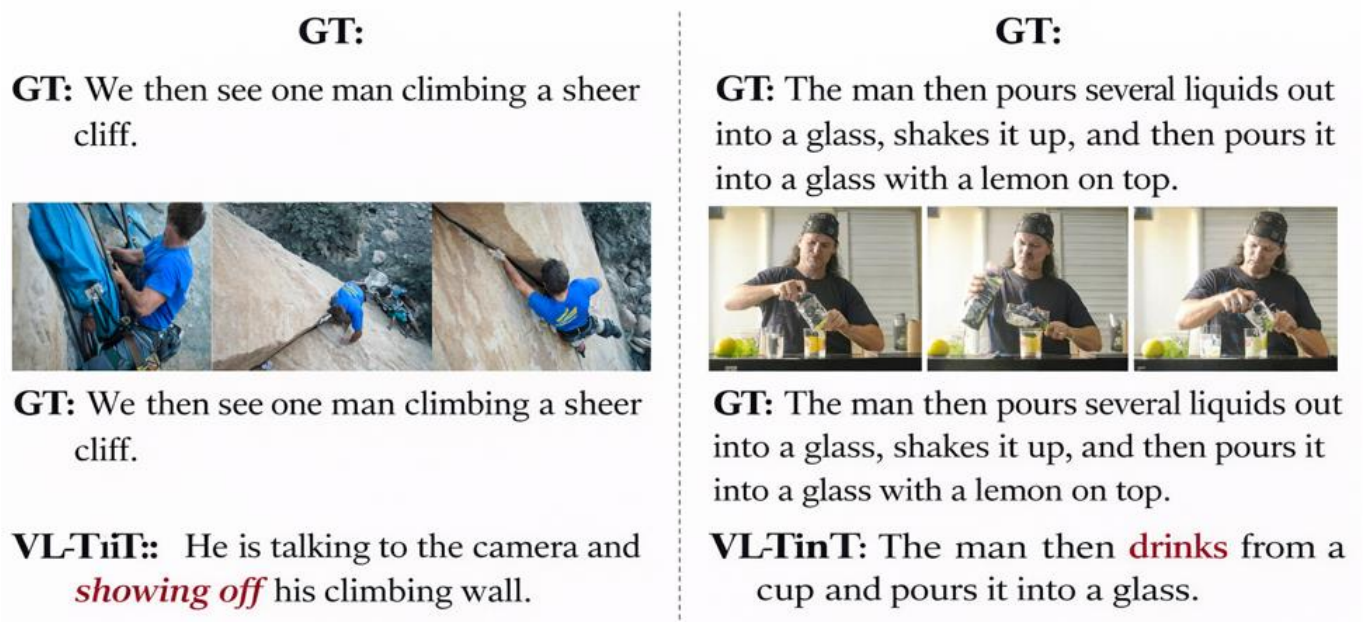

Figure 5. *A video featuring descriptions generated by VLTinT model and ground truth (GT) with description errors highlighted in red. (Chuang and Fazli, 2023).*

Despite their proficiency in visual semantic comprehension and meme humor, MLLMs struggle with chart analysis and understanding. To address this, [19, 30] proposed ChartBench, a benchmark assessing chart comprehension.

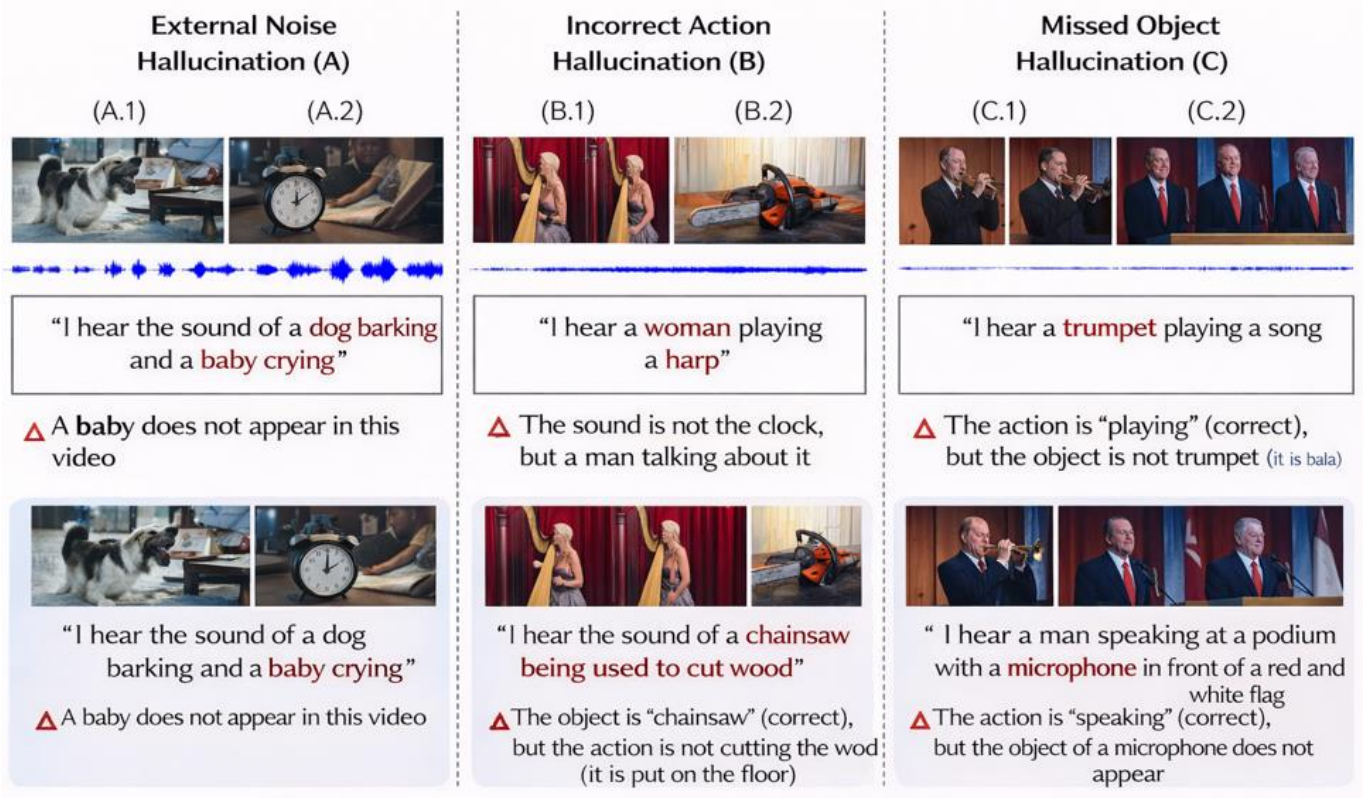

Figure 6. *Audio hallucination examples for each class- Type A: Involving hallucinations of both objects and actions: Type B: featuring accurate objects but hallucinated actions; Type C: Displaying correct actions but hallucinated objects. (Nishimura et al., 2024).*

ChartBench exposes MLLMs' limited reasoning with complex charts, prompting the need for novel evaluation metrics like Acc+ and a handcrafted prompt, ChartCoT.

3.3. Evaluation Matrices to Detect Degree of Hallucination

Benchmarking is a critical tool for evaluating and mitigating model hallucinations. By establishing standardized metrics and datasets, researchers can objectively assess model performance and track progress in addressing this challenge.

A two-level hierarchical fusion method was pioneered to synthesize facial expression sequences from a single neutral face image [5]. To enable effective training of the system, a dedicated dataset comprising 112 video sequences representing four facial expressions (happy, angry, surprise, and fear) collected from 28 individuals was developed.. This approach generated realistic facial expression sequences with minimal artifacts.

In the domain of video understanding, end-to-end chat-centric systems have gained significant attention. [12] introduced the YouCook2 dataset, a comprehensive collection of cooking videos with temporally localized and described procedural segments, to facilitate procedure learning tasks.

A novel framework termed *VideoChat* was introduced to integrate video foundation models with large language models, thereby enhancing spatiotemporal reasoning, event localization, and causal relationship inference in video understanding [7, 9, 11]. They constructed a video-centric instruction dataset with detailed descriptions and conversations, emphasizing spatiotemporal reasoning and causal relationships. To mitigate model hallucinations, they employed a multi-step process involving GPT-4 to condense video descriptions into coherent narratives.

To delve into the challenge of inferring scene affordances, [20] curated a dataset of 2.4 million video clips, showcasing various plausible poses aligned with scene context

Table 1. *Comprehensive Overview of Hallucination Research in Large Foundation Models: Detection, Mitigation, Tasks, Datasets, and Evaluation Metrics [21].*

Title	Detect?	Mitigate?	Task(s)	Dataset	Evaluation Metric
Citation: A Key to Building Responsible and Accountable Large Language Models (Huang and Chang, 2023)	✓	✓	N/A	N/A	N/A
Zero-resource hallucination prevention for large language models (Lao et al., 2023)	✓	✓	Concept extraction, guessing, aggregation	Concept-7	AUC, ACC, F1, PEA
RARR: Researching and Revising What Language Models Say, Using Language Models (Gao et al., 2023)	✓	✓	Editing for Attribution	NQ, SQuAD, QReCC	Attributable to Idea-filled Sources (Castaño and Yang, 2007)
Evaluating Object Hallucination in Large Vision-Language Models (Li et al., 2023e)	✗	✓	Image captioning	MSCOCO (Lin et al., 2014)	Caption Hallucination Assessment with Image Relevance (CHAIR) (Rohrbach et al., 2018)
Detecting and Preventing Hallucinations in Large Vision Language Models (Gunjal et al., 2023)	✓	✓	Visual Question Answering (VQA)	M-HalDetect	Accuracy
Plausible May Not Be Faithful: Probing Object Hallucination in Vision-Language Pre-training (Dai et al., 2022)	✗	✓	Image captioning	CHAIR (Rohrbach et al., 2018)	CIDEr
Let's Think Frame by Frame: Evaluating Video Chain of Thought with Video Inference and Prediction (Himankahtala et al., 2023)	✓	✗	Video infilling, Scene prediction	Manual	N/A
Putting People in Their Place: Affordance-Aware Human Insertion into Videos (Kulal et al., 2023)	✗	✓	Affordance prediction	Manual (2.4M video clips)	FID, PCKh
VideoChat: Chat-Centric Video Understanding (Li et al., 2023c)	✓	✓	Visual dialogue	Manual	N/A
Models See Hallucinations: Evaluating the Factuality in Video Captioning (Liu and Wan, 2023)	✗	✓	Video captioning	ActivityNet Captions (Krishna et al., 2017), YouCook2 (Zhou et al., 2017)	Factual consistency for Video Captioning (FactVC)
LP-MusicCaps: LLM-based pseudo music captioning (Doh et al., 2023)	✓	✓	Audio captioning	LP-MusicCaps	BLEU-1 to 4 (B1–B4), METEOR (M), ROUGE-L
Audio-Journey: Efficient Visual-LLM-aided Audio Encode Diffusion (Li et al., 2023a)	✓	✓	Classification	Manual	Mean Average Precision (mAP)

Table 1 above summarises various literature related to hullicination in all four modalities of the large foundation models. This study has secmented each work by the following criterias: a) Detection b) Mitigation c) Task d) data set e) Evalautioon Metrics. The '√' indicate it is present in the paper whereas the 'X' indicates it is absent in the paper.

4. Future Outlook

Hallucinations in LLMs pose a significant challenge to their widespread adoption. While significant progress has been made in detection and mitigation techniques, further research is needed to develop robust and effective solutions. By addressing this issue, we can unlock the full potential of LLMs and ensure their responsible and beneficial use.

Researchers are actively investigating techniques to mitigate hallucinations, a crucial step for sensitive applications. Key future directions include improving data quality, refining evaluation metrics, enhancing model reasoning, and developing robust ethical guidelines.

5. Conclusions

This survey paper provides a comprehensive overview of research on model hallucinations in foundation models (FMs), covering critical aspects such as detection, mitigation, tasks, datasets, and evaluation metrics. The paper highlights the pervasive nature of hallucinations in FMs and the urgent need to address this issue due to their increasing importance in various domains. A key contribution of this paper is the development of a structured taxonomy for classifying hallucinations in text, image, video, and audio domains.

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] Ailin Deng, ZC and BH 2024,. 'Seeing is believing: Mitigating hallucination in large visionlanguage models via clip-guided decoding.'

[2] Anthony Chen, PPSSHLKG 2023, 'PURR: Efficiently Editing Language Model Hallucinations by Denoising Language Model Corruptions', *Computation and Language*.

[3] Fuxiao Liu, KLLLJWYY and LWang 2023, 'Mitigating hallucination in large multi-modal models via robust instruction tuning. ', *In The Twelfth International Conference on Learning Representations*.

[4] Holy Lovenia, WDSCZJ and PF 2023, 'Negative object presence evaluation (nope) to measure object hallucination in vision-language models', *arXiv preprint arXiv: 2310.05338*.

[5] Jian Zhang, YZ and FWu 2006, 'Videobased facial expression hallucination: A two-level hierarchical fusion approach', *In International Conference on Advanced Concepts for Intelligent Vision Systems*.

[6] Jiaxi Cui, ZLYYBCLY 2023, 'ChatLaw: Open-Source Legal Large Language Model with Integrated External Knowledge Bases', *Computation and Language*.

[7] Jiazhen Liu, and XLi 2024, 'Phd: A prompted visual hallucination evaluation dataset', *arXiv preprint arXiv: 2403.11116*.

[8] Jie Huang, KC-CC 2024, 'Citation: A Key to Building Responsible and Accountable Large Language Models', *Computation and Language*.

[9] Jing Wu, JH and NH n.d., 'Hallucination improves the performance of unsupervised visual representation learning.', *Computer Vision and Pattern Recognition*.

[10] Jonghwan Mun, LYZRNX and BH 2019, ' Streamlined dense video captioning', *In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6588-6597.

[11] KunChang Li, and YQ 2023, 'Videochat: Chat-centric video understanding', *arXiv preprint arXiv: 2305.06355*.

[12] Luowei Zhou, CX and JC 2018, 'Towards automatic learning of procedures from web instructional videos. ', *In Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32.

[13] Maitreya Suin and AN Rajagopalan. 2020, 'An efficient framework for dense video captioning.', *In Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 12039-12046.

[14] Min, S, Krishna, K, Lyu, X, Lewis, M, Yih, W, Koh, P, Iyyer, M, Zettlemoyer, L & Hajishirzi, H 2023, 'FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation', in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 12076-12100.

[15] Niels Mündler, JHSJMV 2023, 'Self-contradictory Hallucinations of Large Language Models: Evaluation, Detection and Mitigation', *Computation and Language*.

[16] Potsawee Manakul, ALMJFG 2023, 'SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models', *Computation and Language*.

[17] Qifan Yu, and YZhuang 2023, 'Hallucidoctor: Mitigating hallucinatory toxicity in visual instruction data.', *arXiv preprint arXiv: 2311.13614*.

[18] Rombach, R, Blattmann, A, Lorenz, D, Esser, P & Ommer, B 2022, 'High-Resolution Image Synthesis with Latent Diffusion Models', in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, pp. 10674-10685.

[19] Shiping Yang, RSXW 2023, 'A New Benchmark and Reverse Validation Method for Passage-level Hallucination Detection'.

[20] Sumith Kulal, and KKS 2023, 'Putting people in their place: Affordance-aware human insertion into scenes.', *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17089-17099.

[21] Vipula Rawte, ASAD 2023, 'A Survey of Hallucination in "Large" Foundation Models', *AI Institute, University of South Carolina, USA*.

[22] Vladimir Iashin and Esa Rahtu 2020, 'Multi-modal dense video captioning.', *In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp. 958-959.

[23] Weilun Wu and Yang Gao. n.d., 'A context-aware model with a pre-trained context encoder for dense video captioning. ', *In International Conference on Cyber Security, Artificial Intelligence, and Digital Economy (CSAIDE 2023)*, vol. 12718, pp. 387-396.

[24] Wenliang Dai, ZLZJDS and PFung n.d., ‘ Plausible may not be faithful: Probing object hallucination in vision-language pre-training’, *arXiv preprint arXiv: 2210.07688*.

[25] Xingyi Zhou, and CS 2024, ‘Streaming dense video captioning. ’, *arXiv preprint arXiv: 2404.01297*.

[26] Xintong Wang, JPLDCB 2024a, ‘Mitigating Hallucinations in Large Vision-Language Models with Instruction Contrastive Decoding’, *Computer Vision and Pattern Recognition*.

[27] Xintong Wang, Jingheng Pan, Liang Ding, and Chris Biemann 2024b, ‘Mitigating Hallucinations in Large Vision-Language Models with Instruction Contrastive Decoding’.

[28] Yifan Li, and J-RW 2023, ‘Evaluating object hallucination in large vision-language models’, *arXiv preprint arXiv: 2305.10355*.

[29] Yue Zhang, 2023, ‘Siren’s Song in the AI Ocean: A Survey on Hallucination in Large Language Models’, *Computation and Language*.

[30] Zhengzhuo Xu, and JGuo 2023, ‘Chartbench: A benchmark for complex visual reasoning in charts. ’, *arXiv preprint arXiv: 2312.15915*.

[31] Ziwei Xu, SJ and MKankanhalli 2024. ‘Hallucination is inevitable: An innate limitation of large language models’. *In Proceedings of the 31st ACM International Conference on Multimedia*.