A Survey of Hallucination in "Large" Foundation Models

Vipula Rawte[1]*, **Amit Sheth**[1], **Amitava Das**[1]
[1]AI Institute, University of South Carolina, USA
{vrawte}@mailbox.sc.edu

Abstract

Hallucination in a foundation model (FM) refers to the generation of content that strays from factual reality or includes fabricated information. This survey paper provides an extensive overview of recent efforts that aim to identify, elucidate, and tackle the problem of hallucination, with a particular focus on "Large" Foundation Models (LFMs). The paper classifies various types of hallucination phenomena that are specific to LFMs and establishes evaluation criteria for assessing the extent of hallucination. It also examines existing strategies for mitigating hallucination in LFMs and discusses potential directions for future research in this area. Essentially, the paper offers a comprehensive examination of the challenges and solutions related to hallucination in LFMs.

1 Introduction

Foundation Models (FMs), exemplified by GPT-3 (Brown et al., 2020) and Stable Diffusion (Rombach et al., 2022), marks the commencement of a novel era in the realm of machine learning and generative artificial intelligence. Researchers introduced the term **"foundation model"** to describe machine learning models that are trained on extensive, diverse, and unlabeled data, enabling them to proficiently handle a wide array of general tasks. These tasks encompass language comprehension, text and image generation, and natural language conversation.

1.1 What is a Foundation Model

Foundation models refer to massive AI models trained on extensive volumes of unlabeled data, typically through self-supervised learning. This training approach yields versatile models capable of excelling in a diverse range of tasks, including image classification, natural language processing, and question-answering, achieving remarkable levels of accuracy.

These models excel in tasks involving generative abilities and human interaction, such as generating marketing content or producing intricate artwork based on minimal prompts. However, adapting and implementing these models for enterprise applications can present certain difficulties (Bommasani et al., 2021).

1.2 What is Hallucination in Foundation Model?

Hallucination in the context of a foundation model refers to a situation where the model generates content that is not based on factual or accurate information. Hallucination can occur when the model produces text that includes details, facts, or claims that are fictional, misleading, or entirely fabricated, rather than providing reliable and truthful information.

This issue arises due to the model's ability to generate plausible-sounding text based on patterns it has learned from its training data, even if the generated content does not align with reality. Hallucination can be unintentional and may result from various factors, including biases in the training data, the model's lack of access to real-time or up-to-date information, or the inherent limitations of the model in comprehending and generating contextually accurate responses.

Addressing hallucination in foundation models and LLMs is crucial, especially in applications where factual accuracy is paramount, such as journalism, healthcare, and legal contexts. Researchers and developers are actively working on techniques to mitigate hallucinations and improve the reliability and trustworthiness of these models. With the recent rise in this problem Fig. 2, it has become even more critical to address them.

*corresponding author

1.3 Why this survey?

In recent times, there has been a significant surge of interest in LFMs within both academic and industrial sectors. Additionally, one of their main challenges is *hallucination*. The survey in (Ji et al., 2023) describes hallucination in natural language generation. In the era of **large** models, (Zhang et al., 2023c) have done another great timely survey studying hallucination in LLMs. However, besides not only in LLMs, the problem of hallucination also exists in other foundation models such as image, video, and audio as well. Thus, in this paper, we do the first comprehensive survey of hallucination across all major modalities of foundation models.

1.3.1 Our contributions

The contributions of this survey paper are as follows:

1. We succinctly categorize the existing works in the area of hallucination in LFMs, as shown in Fig. 1.

2. We offer an extensive examination of large foundation models (LFMs) in Sections 2 to 5.

3. We cover all the important aspects such as i. detection, ii. mitigation, iii. tasks, iv. datasets, and v. evaluation metrics, given in Table 1.

4. We finally also provide our views and possible future direction in this area. We will regularly update the associated open-source resources, available for access at `https://github.com/vr25/hallucination-foundation-model-survey`

1.3.2 Classification of Hallucination

As shown in Fig. 1, we broadly classify the LFMs into **four** types as follows: i. Text, ii. Image, iii. video, and iv. Audio.

The paper follows the following structure. Based on the above classification, we describe the hallucination and mitigation techniques for all four modalities in: i. text (Section 2), ii. image (Section 3), iii. video (Section 4), and iv. audio (Section 5). In Section 6, we briefly discuss how hallucinations are NOT always bad, and hence, in the creative domain, they can be well-suited to producing artwork. Finally, we give some possible future directions for addressing this issue along with a conclusion in Section 7.

2 Hallucination in Large Language Models

As shown in Fig. 4, hallucination occurs when the LLM produces fabricated responses.

2.1 LLMs

SELFCHECKGPT (Manakul et al., 2023), is a method for zero-resource black-box hallucination detection in generative LLMs. This technique focuses on identifying instances where these models generate inaccurate or unverified information without relying on additional resources or labeled data. It aims to enhance the trustworthiness and reliability of LLMs by providing a mechanism to detect and address hallucinations without external guidance or datasets. Self-contradictory hallucinations in LLMs are explored in (Mündler et al., 2023). and addresses them through evaluation, detection, and mitigation techniques. It refers to situations where LLMs generate text that contradicts itself, leading to unreliable or nonsensical outputs. This work presents methods to evaluate the occurrence of such hallucinations, detect them in LLM-generated text, and mitigate their impact to improve the overall quality and trustworthiness of LLM-generated content.

PURR (Chen et al., 2023) is a method designed to efficiently edit and correct hallucinations in language models. PURR leverages denoising language model corruptions to identify and rectify these hallucinations effectively. This approach aims to enhance the quality and accuracy of language model outputs by reducing the prevalence of hallucinated content.

Hallucination datasets: Hallucinations are commonly linked to knowledge gaps in language models (LMs). However, (Zhang et al., 2023a) proposed a hypothesis that in certain instances when language models attempt to rationalize previously generated hallucinations, they may produce false statements that they can independently identify as inaccurate. Thus, they created three question-answering datasets where ChatGPT and GPT-4 frequently provide incorrect answers and accompany them with explanations that contain at least one false assertion.

HaluEval (Li et al., 2023b), is a comprehensive benchmark designed for evaluating hallucination in LLMs. It serves as a tool to systematically assess LLMs' performance in terms of hallucination

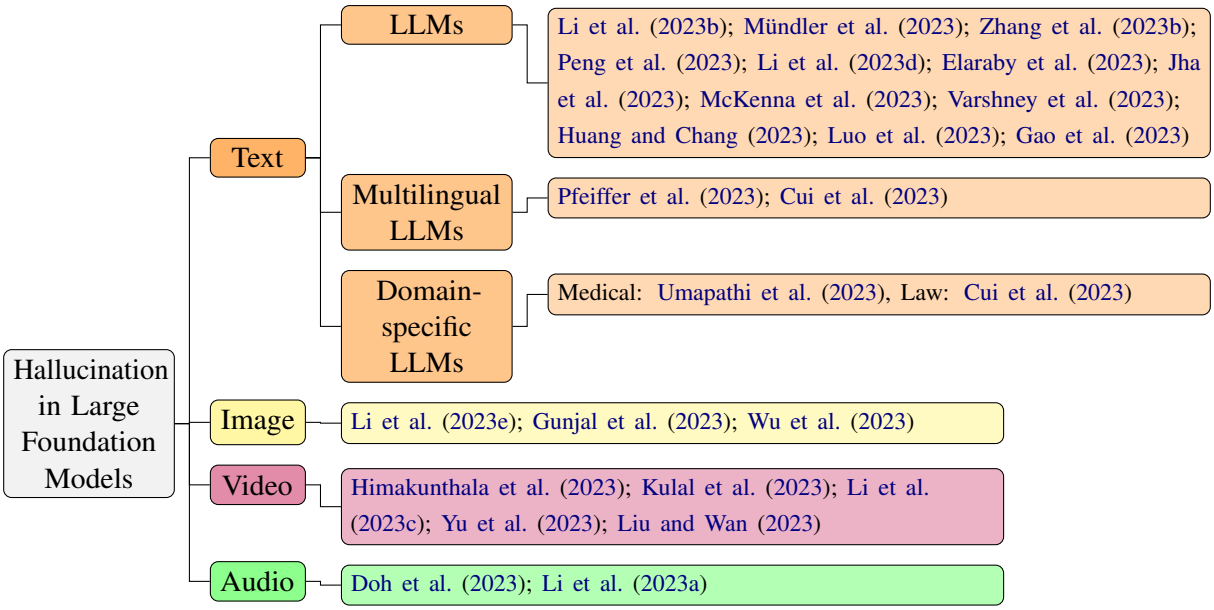

Figure 1: Taxonomy for Hallucination in Large Foundation Models

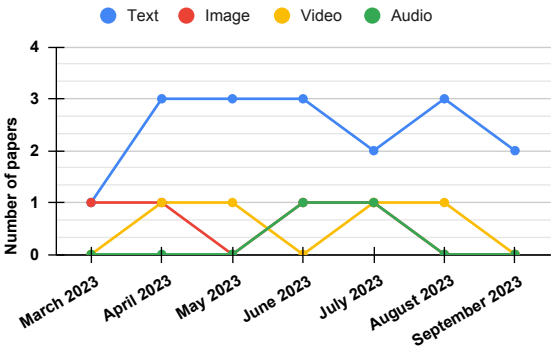

Figure 2: The evolution of "hallucination" papers for Large Foundation Models (LFMs) from March 2023 to September 2023.

User: I am writing a paper about the drug Skytrofa.

GPT-3.5: That sounds interesting! Skytrofa (molgradex) is a medication developed for the treatment of autoimmune pulmonary alveolar proteinosis (aPAP). It's a rare lung disorder where a type of protein called surfactant accumulates in the air sacs of the lungs, leading to breathing difficulties.

Figure 3: An illustration of hallucination (Luo et al., 2023). Incorrect information is highlighted in Red.

across various domains and languages, helping researchers and developers gauge and improve the reliability of these models.

Hallucination mitigation using external knowledge: Using interactive question-knowledge alignment, (Zhang et al., 2023b) presents a method for mitigating language model hallucination Their proposed approach focuses on aligning generated text with relevant factual knowledge, enabling users to interactively guide the model's responses to produce more accurate and reliable information. This technique aims to improve the quality and factuality of language model outputs by involving users in the alignment process. LLM-AUGMENTER (Peng et al., 2023) improves LLMs using external knowledge and automated feedback. It highlights the need to address the limitations and potential factual errors in LLM-generated content. This method involves incorporating external knowledge sources and automated feedback mechanisms to enhance the accuracy and reliability of LLM outputs. By doing so, the paper aims to mitigate factual inaccuracies and improve the overall quality of LLM-generated text. Similarly, (Li et al., 2023d) introduces a framework called "Chain of Knowledge" for grounding LLMs with structured knowledge bases. Grounding refers to the process of connecting LLM-generated text with structured knowledge to improve factual accuracy and reliability. The framework utilizes a hierarchical approach, chaining multiple knowledge sources together to provide context and enhance the understanding of LLMs. This approach aims to improve the alignment of LLM-generated content with structured knowledge, reducing the risk of generating inaccurate or hallucinated information.

Smaller, open-source LLMs with fewer param-

eters often experience significant hallucination issues compared to their larger counterparts (Elaraby et al., 2023). This work focuses on evaluating and mitigating hallucinations in BLOOM 7B, which represents weaker open-source LLMs used in research and commercial applications. They introduce HALOCHECK, a lightweight knowledge-free framework designed to assess the extent of hallucinations in LLMs. Additionally, it explores methods like knowledge injection and teacher-student approaches to reduce hallucination problems in low-parameter LLMs.

Moreover, the risks associated with LLMs can be mitigated by drawing parallels with web systems (Huang and Chang, 2023). It highlights the absence of a critical element, "citation," in LLMs, which could improve content transparency, and verifiability, and address intellectual property and ethical concerns.

Hallucination mitigation using prompting techniques: "Dehallucinating" refers to reducing the generation of inaccurate or hallucinated information by LLMs. Dehallucinating LLMs using formal methods guided by iterative prompting is presented in (Jha et al., 2023). They employ formal methods to guide the generation process through iterative prompts, aiming to improve the accuracy and reliability of LLM outputs. This method is designed to mitigate the issues of hallucination and enhance the trustworthiness of LLM-generated content.

2.2 Multilingual LLMs

Large-scale multilingual machine translation systems have shown impressive capabilities in directly translating between numerous languages, making them attractive for real-world applications. However, these models can generate hallucinated translations, which pose trust and safety issues when deployed. Existing research on hallucinations has mainly focused on small bilingual models for high-resource languages, leaving a gap in understanding hallucinations in massively multilingual models across diverse translation scenarios.

To address this gap, (Pfeiffer et al., 2023) conducted a comprehensive analysis on both the M2M family of conventional neural machine translation models and ChatGPT, a versatile LLM that can be prompted for translation. The investigation covers a wide range of conditions, including over 100 translation directions, various resource levels, and languages beyond English-centric pairs.

2.3 Domain-specific LLMs

Hallucinations in mission-critical areas such as medicine, banking, finance, law, and clinical settings refer to instances where false or inaccurate information is generated or perceived, potentially leading to serious consequences. In these sectors, reliability and accuracy are paramount, and any form of hallucination, whether in data, analysis, or decision-making, can have significant and detrimental effects on outcomes and operations. Consequently, robust measures and systems are essential to minimize and prevent hallucinations in these high-stakes domains.

Medicine: The issue of hallucinations in LLMs, particularly in the medical field, where generating plausible yet inaccurate information can be detrimental. To tackle this problem, (Umapathi et al., 2023) introduces a new benchmark and dataset called Med-HALT (Medical Domain Hallucination Test). It is specifically designed to evaluate and mitigate hallucinations in LLMs. It comprises a diverse multinational dataset sourced from medical examinations across different countries and includes innovative testing methods. Med-HALT consists of two categories of tests: reasoning and memory-based hallucination tests, aimed at assessing LLMs' problem-solving and information retrieval capabilities in medical contexts.

Law: ChatLaw (Cui et al., 2023), is an open-source LLM specialized for the legal domain. To ensure high-quality data, the authors created a meticulously designed legal domain fine-tuning dataset. To address the issue of model hallucinations during legal data screening, they propose a method that combines vector database retrieval with keyword retrieval. This approach effectively reduces inaccuracies that may arise when solely relying on vector database retrieval for reference data retrieval in legal contexts.

3 Hallucination in Large Image Models

Contrastive learning models, employing a Siamese structure (Wu et al., 2023), have displayed impressive performance in self-supervised learning. Their success hinges on two crucial conditions: the presence of a sufficient number of positive pairs and the existence of ample variations among them. Without meeting these conditions, these frameworks may lack meaningful semantic distinctions and become susceptible to overfitting. To tackle these

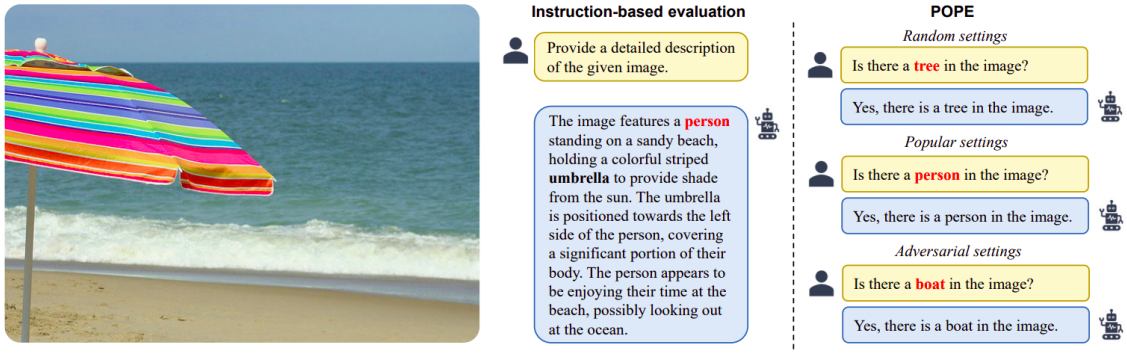

Figure 4: Instances of object hallucination within LVLMs (Li et al., 2023e). Ground-truth objects in annotations are indicated in **bold**, while red objects represent hallucinated objects by LVLMs. The left case occurs in the conventional instruction-based evaluation approach, while the right cases occur in three variations of POPE.

challenges, we introduce the Hallucinator, which efficiently generates additional positive samples to enhance contrast. The Hallucinator is differentiable, operating in the feature space, making it amenable to direct optimization within the pre-training task and incurring minimal computational overhead.

Efforts to enhance LVLMs for complex multimodal tasks, inspired by LLMs, face a significant challenge: object hallucination, where LVLMs generate inconsistent objects in descriptions. This study (Li et al., 2023e) systematically investigates object hallucination in LVLMs and finds it's a common issue. Visual instructions, especially frequently occurring or co-occurring objects, influence this problem. Existing evaluation methods are also affected by input instructions and LVLM generation styles. To address this, the study introduces an improved evaluation method called POPE, providing a more stable and flexible assessment of object hallucination in LVLMs.

Instruction-tuned Large Vision Language Models (LVLMs) have made significant progress in handling various multimodal tasks, including Visual Question Answering (VQA). However, generating detailed and visually accurate responses remains a challenge for these models. Even state-of-the-art LVLMs like InstructBLIP exhibit a high rate of hallucinatory text, comprising 30 percent of non-existent objects, inaccurate descriptions, and erroneous relationships. To tackle this issue, the study (Gunjal et al., 2023)introduces MHalDetect1, a Multimodal Hallucination Detection Dataset designed for training and evaluating models aimed at detecting and preventing hallucinations. M-HalDetect contains 16,000 finely detailed annotations on VQA examples, making it the first comprehensive dataset for detecting hallucinations in detailed image descriptions.

4 Hallucination in Large Video Models

Hallucinations can occur when the model makes incorrect or imaginative assumptions about the video frames, leading to the creation of artificial or erroneous visual information Fig. 5.

Video content:

Caption 1:
A *woman* is throwing darts at a board.
She throws them at a board.
She jumps off into the distance and smiles.

Caption 2:
A man is seen standing in a room and leads into a man *speaking to the camera*.
The man is throwing darts at a dart board .
The man then throws the dart *board* and then *goes back to the camera*.

Caption 3:
A man in a *white* shirt is standing *at a dart board*.
He throws a dart at the end.

Figure 5: A video featuring three captions generated by various captioning models (Liu and Wan, 2023), with factual errors highlighted in red italics.

The challenge of understanding scene affordances is tackled by introducing a method for inserting people into scenes in a lifelike manner (Kulal et al., 2023). Using an image of a scene with a marked area and an image of a person, the model seamlessly integrates the person into the

scene while considering the scene's characteristics. The model is capable of deducing realistic poses based on the scene context, adjusting the person's pose accordingly, and ensuring a visually pleasing composition. The self-supervised training enables the model to generate a variety of plausible poses while respecting the scene's context. Additionally, the model can also generate lifelike people and scenes on its own, allowing for interactive editing.

VideoChat (Li et al., 2023c), is a comprehensive system for understanding videos with a chat-oriented approach. VideoChat combines foundational video models with LLMs using an adaptable neural interface, showcasing exceptional abilities in understanding space, time, event localization, and inferring cause-and-effect relationships. To fine-tune this system effectively, they introduced a dataset specifically designed for video-based instruction, comprising thousands of videos paired with detailed descriptions and conversations. This dataset places emphasis on skills like spatiotemporal reasoning and causal relationships, making it a valuable resource for training chat-oriented video understanding systems.

Recent advances in video inpainting have been notable (Yu et al., 2023), particularly in cases where explicit guidance like optical flow can help propagate missing pixels across frames. However, challenges arise when cross-frame information is lacking, leading to shortcomings. So, instead of borrowing pixels from other frames, the model focuses on addressing the reverse problem. This work introduces a dual-modality-compatible inpainting framework called Deficiency-aware Masked Transformer (DMT). Pretraining an image inpainting model to serve as a prior for training the video model has an advantage in improving the handling of situations where information is deficient.

Video captioning aims to describe video events using natural language, but it often introduces factual errors that degrade text quality. While factuality consistency has been studied extensively in text-to-text tasks, it received less attention in vision-based text generation. In this research (Liu and Wan, 2023), the authors conducted a thorough human evaluation of factuality in video captioning, revealing that 57.0% of model-generated sentences contain factual errors. Existing evaluation metrics, mainly based on n-gram matching, do not align well with human assessments. To address this issue, they introduced a model-based factuality metric called FactVC, which outperforms previous metrics in assessing factuality in video captioning.

5 Hallucination in Large Audio Models

Automatic music captioning, which generates text descriptions for music tracks, has the potential to enhance the organization of vast musical data. However, researchers encounter challenges due to the limited size and expensive collection process of existing music-language datasets. To address this scarcity, (Doh et al., 2023) used LLMs to generate descriptions from extensive tag datasets. They created a dataset known as LP-MusicCaps, comprising around 2.2 million captions paired with 0.5 million audio clips. They also conducted a comprehensive evaluation of this large-scale music captioning dataset using various quantitative natural language processing metrics and human assessment. They trained a transformer-based music captioning model on this dataset and evaluated its performance in zero-shot and transfer-learning scenarios.

Ideally, the video should enhance the audio, and in (Li et al., 2023a), they have used an advanced language model for data augmentation without human labeling. Additionally, they utilized an audio encoding model to efficiently adapt a pre-trained text-to-image generation model for text-to-audio generation.

6 Hallucination is *not* always harmful: A different perspective

Suggesting an alternative viewpoint, (Wiggers, 2023) discusses how hallucinating models could serve as "collaborative creative partners," offering outputs that may not be entirely grounded in fact but still provide valuable threads to explore. Leveraging hallucination creatively can lead to results or novel combinations of ideas that might not readily occur to most individuals.

"Hallucinations" become problematic when the statements generated are factually inaccurate or contravene universal human, societal, or particular cultural norms. This is especially critical in situations where an individual relies on the LLM to provide expert knowledge. However, in the context of creative or artistic endeavors, the capacity to generate unforeseen outcomes can be quite advantageous. Unexpected responses to queries can surprise humans and stimulate the discovery of novel idea connections.

	Title	Detect	Mitigate	Task(s)	Dataset	Evaluation Metric
TEXT	SELFCHECKGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models (Manakul et al., 2023)	✓	✗	QA	Manual (WikiBio)	Token probability or entropy
	HaluEval: A Large-Scale Hallucination Evaluation Benchmark for Large Language Models (Li et al., 2023b)	✓	✓	QA, Dialogue Summarization, General	HaluEval	Automatic
	Self-contradictory Hallucinations of Large Language Models: Evaluation, Detection and Mitigation (Mündler et al., 2023)	✓	✓	Text generation	Manual	F1 score
	PURR: Efficiently Editing Language Model Hallucinations by Denoising Language Model Corruptions (Chen et al., 2023)	✗	✓	Editing for Attribution	Multiple question answering, Dialog datasets	Attribution, Preservation
	Mitigating Language Model Hallucination with Interactive Question-Knowledge Alignment (Zhang et al., 2023b)	✗	✓	Question-knowledge alignment	FuzzyQA	Attributable to Identified Sources (Castaldo and Yang, 2007)
	How Language Model Hallucinations Can Snowball (Zhang et al., 2023a)	✓	✗	QA	Manual	Accuracy
	Check Your Facts and Try Again: Improving Large Language Models with External Knowledge and Automated Feedback (Peng et al., 2023)	✗	✓	Task oriented dialog and open-domain question answering	News Chat, Customer Service	Knowledge F1 (KF1) and BLEU-4
	ChatLawLLM (Cui et al., 2023)	✗	✓	QA	Manual	ELO model ranking
	The Internal State of an LLM Knows When its Lying (Azaria and Mitchell, 2023)	✓	✗	Classification	Manual	Accuracy
	Chain of Knowledge: A Framework for Grounding Large Language Models with Structured Knowledge Bases (Li et al., 2023d)	✓	✓	Knowledge intensive tasks	FEVER, AdvHotpotQA	Accuracy
	HALO: Estimation and Reduction of Hallucinations in Open-Source Weak Large Language Models (Elaraby et al., 2023)	✓	✓	Consistency, Factuality, BS, QA, NLI	Manual on NBA domain	Pearson and Kendall tau correlation coefficients
	A Stitch in Time Saves Nine: Detecting and Mitigating Hallucinations of LLMs by Validating Low-Confidence Generation (Varshney et al., 2023)	✓	✓	Article generation	WikiBio	Percentage of mitigated hallucinations
	Dehallucinating Large Language Models Using Formal Methods Guided Iterative Prompting (Jha et al., 2023)	✓	✗	Dialog	-	-
	Med-HALT: Medical Domain Hallucination Test for Large Language Models (Umapathi et al., 2023)	✗	✗	Reasoning Hallucination Test (RHT), Memory Hallucination Test (MHT)	Med-HALT	Accuracy, Pointwise score
	Sources of Hallucination by Large Language Models on Inference Tasks (McKenna et al., 2023)	✓	✗	Textual entailment	Altered directional inference adatset	Enatilment probability
	Hallucinations in Large Multilingual Translation Models (Pfeiffer et al., 2023)	✓	✓	MT	FLORES-101, WMT, and TICO	spBLEU

Table 1 continued from previous page

	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 (Luo 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, SQA, QReCC	Attributable to Identified Sources (Castaldo and Yang, 2007)
IMAGE	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
VIDEO	Let's Think Frame by Frame: Evaluating Video Chain of Thought with Video Infilling and Prediction (Himakunthala et al., 2023)	✗	✓	Video infilling, Scene prediction	Manual	N/A
	Putting People in Their Place: Affordance-Aware Human Insertion into Scenes (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 (Krishna et al., 2017)	Factual consistency for Video Captioning (FactVC)
AUDIO	LP-MusicCaps: LLM-based pseudo music captioning (Doh et al., 2023)	✗	✓	Audio Captioning	LP-MusicCaps	BLEU1 to 4 (B1, B2, B3, B4), METEOR (M), and ROUGE-L (R-L)
	Audio-Journey: Efficient Visual+LLM-aided Audio Encodec Diffusion (Li et al., 2023a)	✗	✓	Classification	Manual	Mean average precision (mAP)

Table 1: Summary of all the works related to hallucination in all four modalities of the large foundation models. Here, we have divided each work by the following factors: 1. Detection, 2. Mitigation, 3. Tasks, 4. Datasets, and 5. Evaluation metrics. ✓ indicates that it is present in the paper whereas ✗ indicates it is *not* present.

7 Conclusion and Future Directions

We concisely classify the existing research in the field of hallucination within LFMs. We provide an in-depth analysis of these LFMs, encompassing critical aspects including 1. Detection, 2. Mitigation, 3. Tasks, 4. Datasets, and 5. Evaluation metrics.

Some possible future directions to address the hallucination challenge in the LFMs are given below.

7.1 Automated Evaluation of Hallucination

In the context of natural language processing and machine learning, hallucination refers to the generation of incorrect or fabricated information by AI models. This can be a significant problem, especially in applications like text generation, where the goal is to provide accurate and reliable information. Here are some potential future directions in the automated evaluation of hallucination:

Development of Evaluation Metrics: Researchers can work on creating specialized evaluation metrics that are capable of detecting hallucination in generated content. These metrics may consider factors such as factual accuracy, coherence, and consistency. Advanced machine learning models could be trained to assess generated text against these metrics.

Human-AI Collaboration: Combining human judgment with automated evaluation systems can be a promising direction. Crowdsourcing platforms can be used to gather human assessments of AI-generated content, which can then be used to train models for automated evaluation. This hybrid approach can help in capturing nuances that are challenging for automated systems alone.

Adversarial Testing: Researchers can develop adversarial testing methodologies where AI systems are exposed to specially crafted inputs designed to trigger hallucination. This can help in identifying weaknesses in AI models and improving their robustness against hallucination.

Fine-Tuning Strategies: Fine-tuning pre-trained language models specifically to reduce hallucination is another potential direction. Models can be fine-tuned on datasets that emphasize fact-checking and accuracy to encourage the generation of more reliable content.

7.2 Improving Detection and Mitigation Strategies with Curated Sources of Knowledge

Detecting and mitigating issues like bias, misinformation, and low-quality content in AI-generated text is crucial for responsible AI development. Curated sources of knowledge can play a significant role in achieving this. Here are some future directions:

Knowledge Graph Integration: Incorporating knowledge graphs and curated knowledge bases into AI models can enhance their understanding of factual information and relationships between concepts. This can aid in both content generation and fact-checking.

Fact-Checking and Verification Models: Develop specialized models that focus on fact-checking and content verification. These models can use curated sources of knowledge to cross-reference generated content and identify inaccuracies or inconsistencies.

Bias Detection and Mitigation: Curated sources of knowledge can be used to train AI models to recognize and reduce biases in generated content. AI systems can be programmed to check content for potential biases and suggest more balanced alternatives.

Active Learning: Continuously update and refine curated knowledge sources through active learning. AI systems can be designed to seek human input and validation for ambiguous or new information, thus improving the quality of curated knowledge.

Ethical Guidelines and Regulation: Future directions may also involve the development of ethical guidelines and regulatory frameworks for the use of curated knowledge sources in AI development. This could ensure responsible and transparent use of curated knowledge to mitigate potential risks.

In summary, these future directions aim to address the challenges of hallucination detection and mitigation, as well as the responsible use of curated knowledge to enhance the quality and reliability of AI-generated content. They involve a combination of advanced machine learning techniques, human-AI collaboration, and ethical considerations to ensure AI systems produce accurate and trustworthy information.

References

Amos Azaria and Tom Mitchell. 2023. The internal state of an llm knows when its lying. *arXiv preprint arXiv:2304.13734*.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Eric T Castaldo and Edmund Y Yang. 2007. Severe sepsis attributable to community-associated methicillin-resistant staphylococcus aureus: an emerging fatal problem. *The American Surgeon*, 73(7):684–687.

Anthony Chen, Panupong Pasupat, Sameer Singh, Hongrae Lee, and Kelvin Guu. 2023. Purr: Efficiently editing language model hallucinations by denoising language model corruptions.

Jiaxi Cui, Zongjian Li, Yang Yan, Bohua Chen, and Li Yuan. 2023. Chatlaw: Open-source legal large language model with integrated external knowledge bases. *arXiv preprint arXiv:2306.16092*.

Wenliang Dai, Zihan Liu, Ziwei Ji, Dan Su, and Pascale Fung. 2022. Plausible may not be faithful: Probing object hallucination in vision-language pre-training. *arXiv preprint arXiv:2210.07688*.

SeungHeon Doh, Keunwoo Choi, Jongpil Lee, and Juhan Nam. 2023. Lp-musiccaps: Llm-based pseudo music captioning. *arXiv preprint arXiv:2307.16372*.

Mohamed Elaraby, Mengyin Lu, Jacob Dunn, Xueying Zhang, Yu Wang, and Shizhu Liu. 2023. Halo: Estimation and reduction of hallucinations in open-source weak large language models. *arXiv preprint arXiv:2308.11764*.

Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, et al. 2023. Rarr: Researching and revising what language models say, using language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16477–16508.

Anisha Gunjal, Jihan Yin, and Erhan Bas. 2023. Detecting and preventing hallucinations in large vision language models. *arXiv preprint arXiv:2308.06394*.

Vaishnavi Himakunthala, Andy Ouyang, Daniel Rose, Ryan He, Alex Mei, Yujie Lu, Chinmay Sonar, Michael Saxon, and William Yang Wang. 2023. Let's think frame by frame: Evaluating video chain of thought with video infilling and prediction. *arXiv preprint arXiv:2305.13903*.

Jie Huang and Kevin Chen-Chuan Chang. 2023. Citation: A key to building responsible and accountable large language models. *arXiv preprint arXiv:2307.02185*.

Susmit Jha, Sumit Kumar Jha, Patrick Lincoln, Nathaniel D Bastian, Alvaro Velasquez, and Sandeep Neema. 2023. Dehallucinating large language models using formal methods guided iterative prompting. In *2023 IEEE International Conference on Assured Autonomy (ICAA)*, pages 149–152. IEEE.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, pages 706–715.

Sumith Kulal, Tim Brooks, Alex Aiken, Jiajun Wu, Jimei Yang, Jingwan Lu, Alexei A Efros, and Krishna Kumar Singh. 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*, pages 17089–17099.

Juncheng B Li, Jackson Sam Michaels, Laura Yao, Lijun Yu, Zach Wood-Doughty, and Florian Metze. 2023a. Audio-journey: Efficient visual+ llm-aided audio encodec diffusion. In *Workshop on Efficient Systems for Foundation Models@ ICML2023*.

Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023b. Helma: A large-scale hallucination evaluation benchmark for large language models. *arXiv preprint arXiv:2305.11747*.

KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. 2023c. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*.

Xingxuan Li, Ruochen Zhao, Yew Ken Chia, Bosheng Ding, Lidong Bing, Shafiq Joty, and Soujanya Poria. 2023d. Chain of knowledge: A framework for grounding large language models with structured knowledge bases. *arXiv preprint arXiv:2305.13269*.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023e. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco:

Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.

Hui Liu and Xiaojun Wan. 2023. Models see hallucinations: Evaluating the factuality in video captioning. *arXiv preprint arXiv:2303.02961*.

Junyu Luo, Cao Xiao, and Fenglong Ma. 2023. Zero-resource hallucination prevention for large language models. *arXiv preprint arXiv:2309.02654*.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models.

Nick McKenna, Tianyi Li, Liang Cheng, Mohammad Javad Hosseini, Mark Johnson, and Mark Steedman. 2023. Sources of hallucination by large language models on inference tasks. *arXiv preprint arXiv:2305.14552*.

Niels Mündler, Jingxuan He, Slobodan Jenko, and Martin Vechev. 2023. Self-contradictory hallucinations of large language models: Evaluation, detection and mitigation.

Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou Yu, Weizhu Chen, et al. 2023. Check your facts and try again: Improving large language models with external knowledge and automated feedback. *arXiv preprint arXiv:2302.12813*.

Jonas Pfeiffer, Francesco Piccinno, Massimo Nicosia, Xinyi Wang, Machel Reid, and Sebastian Ruder. 2023. mmt5: Modular multilingual pre-training solves source language hallucinations.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. *arXiv preprint arXiv:1809.02156*.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695.

Logesh Kumar Umapathi, Ankit Pal, and Malaikannan Sankarasubbu. 2023. Med-halt: Medical domain hallucination test for large language models. *arXiv preprint arXiv:2307.15343*.

Neeraj Varshney, Wenlin Yao, Hongming Zhang, Jianshu Chen, and Dong Yu. 2023. A stitch in time saves nine: Detecting and mitigating hallucinations of llms by validating low-confidence generation. *arXiv preprint arXiv:2307.03987*.

Kyle Wiggers. 2023. Are ai models doomed to always hallucinate?

Jing Wu, Jennifer Hobbs, and Naira Hovakimyan. 2023. Hallucination improves the performance of unsupervised visual representation learning. *arXiv preprint arXiv:2307.12168*.

Yongsheng Yu, Heng Fan, and Libo Zhang. 2023. Deficiency-aware masked transformer for video inpainting. *arXiv preprint arXiv:2307.08629*.

Muru Zhang, Ofir Press, William Merrill, Alisa Liu, and Noah A. Smith. 2023a. How language model hallucinations can snowball.

Shuo Zhang, Liangming Pan, Junzhou Zhao, and William Yang Wang. 2023b. Mitigating language model hallucination with interactive question-knowledge alignment.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. 2023c. Siren's song in the ai ocean: A survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.