A Comprehensive Literature Review with Self-Reflection

Literature Review

October 7, 2025

Abstract

This literature review provides a comprehensive analysis of recent research in the field. The review synthesizes findings from 277 research papers, identifying key themes, methodological approaches, and future research directions.

Contents

1 Introduction

1.1 Defining Hallucination and Its Immediate Impact

Hallucination in Large Language Models (LLMs) fundamentally challenges their reliability and trustworthiness, manifesting as the generation of factually inaccurate, nonsensical, or unfaithful information presented as truthful or coherent **??**. This phenomenon is a critical concern for the secure and accountable deployment of AI, as it directly undermines the dependability and credibility of LLMs across diverse applications. Establishing a clear conceptual foundation and common terminology is therefore paramount for understanding and addressing this pervasive issue.

The conceptualization of hallucination has evolved from task-specific observations to comprehensive frameworks tailored for general-purpose LLMs. Pioneering work in abstractive summarization by **?** provided an early systematic characterization, categorizing hallucinations into two primary forms based on their relationship to the source document: *intrinsic hallucinations*, which misrepresent or contradict information explicitly present in the input, and *extrinsic hallucinations*, which introduce information not inferable from the source but presented as fact. Their human evaluation revealed a significant prevalence of these errors, with over 70

As LLMs grew in scale and versatility, the scope of hallucination expanded beyond simple source document fidelity. **?** extended this taxonomy specifically for LLMs, recognizing that models often generate content without a single, explicit source document. They distinguished between *input-conflicting* errors (similar to intrinsic, but generalized to user prompts), *context-conflicting* errors (contradicting previously generated turns in a dialogue or a longer generated text), and *fact-conflicting* errors (contradicting established world knowledge). This framework highlights the multifaceted nature of LLM hallucinations, where errors can arise from misinterpreting instructions, losing coherence over extended generations, or fabricating facts entirely. Further refining this understanding, **?** proposed a more granular framework, profiling hallucination by its *degree* (mild, moder-

ate, alarming), *orientation* (Factual Mirage and Silver Lining, each with intrinsic/extrinsic sub-categories), and six specific *types* (e.g., Numeric Nuisance, Generated Golem). While **?**'s categories offer a broad, practical classification for LLM behavior, **?**'s detailed taxonomy provides a finer-grained lens for diagnostic analysis, allowing researchers to pinpoint specific error patterns and quantify hallucination vulnerability with metrics like their proposed Hallucination Vulnerability Index (HVI). This progression from broad categories to more specific, measurable types reflects the field's increasing sophistication in dissecting the problem.

The immediate impact of hallucination is profound and far-reaching, directly affecting the utility and societal acceptance of LLMs. The generation of factually incorrect or nonsensical content erodes user trust and credibility, making LLMs unreliable sources of information **?**. This is particularly critical in high-stakes domains such as healthcare, legal advice, or financial consulting, where misinformation can lead to severe consequences **?**. For instance, fact-conflicting errors can propagate misinformation at an unprecedented scale, while context-conflicting errors can render dialogue systems incoherent and frustrating. Beyond textual outputs, the problem extends to multimodal LLMs, where hallucinations can manifest as descriptions that contradict visual or audio inputs, posing significant safety and reliability concerns in applications like autonomous driving or medical image analysis **?**. The pervasive nature of hallucination, therefore, necessitates an urgent focus on robust solutions to ensure the secure, dependable, and accountable integration of AI into critical societal functions.

1.2 Scope and Motivation of the Review

The proliferation of Large Language Models (LLMs) across diverse domains, from content generation to critical decision support, has underscored a fundamental challenge: hallucination. Defined as the generation of plausible yet factually incorrect, inconsistent, or unfaithful information, hallucination poses a significant barrier to the dependable, transparent, and safe deployment of these powerful AI systems **?**. This pervasive issue can lead to the dissemination of misinformation, erode user trust, and introduce substantial

risks in sensitive applications, as highlighted by concerns in fields like medical education where incorrect information can lead to patient harm ?. Consequently, a comprehensive understanding and robust management of hallucination are paramount for the continued advancement and responsible integration of LLMs into society, ultimately contributing to the development of trustworthy artificial intelligence.

This literature review aims to provide a structured and critical synthesis of the evolving research landscape surrounding LLM hallucination. Our scope encompasses a holistic examination, tracing the intellectual trajectory from the foundational understanding of hallucination to advanced mitigation techniques and the emerging complexities in multimodal contexts. Specifically, we will delve into the historical characterization and evolving taxonomies of hallucination, exploring its root causes and the profound theoretical insights into its inherent inevitability. Subsequently, the review will scrutinize the diverse methodologies developed for benchmarking and detection, ranging from reference-free and consistency-based approaches to fine-grained, rationale-based evaluations and specialized diagnostic frameworks for Retrieval-Augmented Generation (RAG) systems. A significant portion of this analysis is dedicated to a critical appraisal of mitigation strategies, categorizing them into external grounding techniques (such as RAG and Knowledge Graphs) and internal model interventions (including decoding-time strategies, training-based approaches, and causal interventions). Finally, we will address the unique challenges and solutions pertinent to multimodal hallucination, alongside advanced topics like adversarial robustness, safety-critical applications, and the pursuit of unified theoretical frameworks.

The primary motivation for this comprehensive analysis stems from the urgent need to consolidate the rapidly expanding body of knowledge in this critical area. The field has witnessed an explosion of research, driven by the increasing scale and versatility of LLMs, which often exhibit subtle yet impactful errors that are difficult for both humans and models to detect ?. This review seeks to synthesize the current state-of-the-art, identify critical gaps in existing approaches, and highlight promising future directions. By systematically mapping the progress in understanding, detecting, and mitigating hallucination, we aim to provide researchers and practitioners with a clear roadmap for developing more

reliable and accountable AI systems. The imperative to ensure verifiability and accountability, for instance, has led to innovative approaches like enabling LLMs to generate text with explicit citations, crucial for building trustworthy information-seeking applications **?**. Furthermore, the unique challenges posed by multimodal models, such as the "modality gap" and inherited LLM hallucination tendencies, necessitate dedicated attention **?**. This synthesis is crucial for moving beyond reactive problem-solving to proactive robustness engineering, fostering a deeper scientific understanding of AI reliability, and ensuring that LLMs can be deployed responsibly and ethically in an increasingly AI-driven world.

To achieve this, the review is structured as follows: Section 2 delves into the foundational understanding and theoretical limits of hallucination. Section 3 surveys the landscape of benchmarking and detection methodologies. Sections 4 and 5 then explore the two major families of mitigation strategies: external grounding and reasoning, and internal model interventions and training, respectively. Section 6 specifically addresses applications, specialized domains, and the complexities of multimodal hallucination. Finally, Section 7 discusses advanced topics and future directions, before Section 8 concludes with a summary of key developments and persistent challenges.

2 Foundational Understanding and Theoretical Limits

2.1 Historical Characterization and Evolving Taxonomies

The challenge of hallucination in natural language generation (NLG) has evolved from an observed anomaly to a systematically categorized and theoretically bounded problem. Initial empirical studies laid the groundwork for understanding and classifying these errors, providing the intellectual lineage for subsequent research and guiding early detection efforts across diverse tasks.

Pioneering work in abstractive summarization first systematically identified and distinguished different types of unfaithful content. **?** conducted a large-scale human evaluation, categorizing hallucinations into "intrinsic" (misrepresenting source information) and "extrinsic" (adding uninferable information). This study not only quantified the high

prevalence of hallucinations (over 70

As LLMs became more prevalent, the scope of hallucination expanded beyond summarization. ? introduced a broader taxonomy for LLMs, distinguishing between "input-conflicting" (diverging from user input), "context-conflicting" (inconsistent with prior generated text), and "fact-conflicting" (misaligning with world knowledge) hallucinations. This framework emphasized the unique challenges posed by LLMs' scale and versatility. Further deepening the understanding of hallucination sources, ? proposed an association analysis framework that attributed hallucinations to specific model capability deficiencies, such as commonsense memorization, relational reasoning, and instruction following, offering a more nuanced, capability-based characterization.

The need for fine-grained classification to enable effective detection and mitigation led to increasingly detailed taxonomies. ? provided a comprehensive review, organizing hallucinations by various text generation tasks (e.g., QA, dialogue, cross-modal systems) and analyzing their origins from data collection, knowledge gaps, and optimization processes. Expanding on this, ? introduced a highly granular framework, profiling hallucination by its degree (mild, moderate, alarming), orientation (Factual Mirage, Silver Lining, each with intrinsic/extrinsic sub-categories), and six specific types (e.g., Numeric Nuisance, Geographic Erratum). This work also proposed the Hallucination Vulnerability Index (HVI) for quantitative assessment, pushing towards standardized measurement. Similarly, ? developed HaluEval 2.0, a benchmark featuring a detailed taxonomy of factuality hallucinations including Entity-error, Relation-error, Incompleteness, Outdatedness, Overclaim, and Unverifiability, which is crucial for fine-grained detection and analysis.

The evolution of taxonomies also extended to the granularity of detection. ? pioneered the concept of token-level, reference-free hallucination detection for free-form text generation, introducing the HADES dataset. This marked a significant step towards real-time, fine-grained identification of errors, moving beyond coarse sentence or document-level assessments. Furthermore, as LLMs became multimodal, new forms of hallucination emerged, necessitating specialized characterization. ? conducted the first

systematic empirical study on "object hallucination" in Large Vision-Language Models (LVLMs), where generated descriptions contradict visual input, and proposed POPE as a more stable evaluation method. This was further refined by **?**, who characterized LMM hallucination by the types of negative instructions (e.g., nonexistent object manipulation, knowledge manipulation) that models fail to follow.

Collectively, these studies illustrate a clear intellectual progression from initial empirical observation to sophisticated, multi-dimensional classification schemes. This systematic structuring of the problem space has been instrumental in guiding the development of targeted detection benchmarks, such as those by **?** for automated dataset generation and **?** for entity-relationship based verification, and laying the groundwork for effective mitigation strategies. While these evolving taxonomies provide powerful tools for analysis and intervention, the theoretical proof by **?** that hallucination is an inherent and inevitable limitation for all computable LLMs sets a fundamental ceiling on the problem, suggesting that complete elimination is impossible. This theoretical understanding reinforces the critical importance of robust detection and mitigation, which heavily rely on these sophisticated characterization frameworks.

2.2 Root Causes and Mechanistic Insights

Understanding the underlying factors contributing to hallucination in Large Language Models (LLMs) is paramount for developing effective and targeted mitigation strategies, moving beyond superficial observations to delve into their mechanistic origins. Hallucinations are not merely random errors but stem from a complex interplay of issues spanning the entire LLM lifecycle, from data acquisition to inference.

A primary root cause lies in the **inherent biases and limitations of vast training data**. LLMs are trained on web-scale corpora that inevitably contain noisy, outdated, or even fabricated information **??**. This noisy data can directly lead to unfaithful generations, as demonstrated by **?**, who addressed this by proposing reference revision to improve the quality of noisy training data for summarization. Empirical studies reveal that low-frequency knowledge in pre-training data correlates with a higher incidence of

hallucinations, while specialized domain data can significantly alleviate domain-specific errors ?. Furthermore, the very process of tokenization and the lack of data diversity can introduce information loss and biases, undermining the factual integrity of generated content ?. This can lead to phenomena like "knowledge overshadowing," where dominant or over-represented facts in the training data cause LLMs to over-generalize, potentially resulting in "amalgamated hallucinations" where disparate but frequently co-occurring facts are incorrectly combined.

Beyond data, **limitations in model architecture and inference-time errors** significantly contribute to hallucination. During the optimization process, LLMs can exhibit "stochastic parroting" due to maximum likelihood estimation, or suffer from "exposure bias" where errors compound over sequential token generation ?. ? empirically showed that certain decoding strategies (e.g., diversity-oriented decoding in professional domains) and even model quantization can elicit higher hallucination rates. A critical architectural challenge arises in processing long contexts; Retrieval-Augmented Language Models (RALMs) often "get lost in long contexts," where irrelevant information distracts the model and exacerbates hallucination ?. This suggests deficiencies in attention mechanisms or contextual filtering. Moreover, LLMs struggle with fundamental capabilities in Retrieval-Augmented Generation (RAG) scenarios, such as negative rejection (failing to abstain from answering when no relevant information is available) and counterfactual robustness (prioritizing incorrect retrieved information over their own correct internal knowledge, even when warned) ?. This highlights a mechanistic flaw in how LLMs weigh and integrate information. The tendency for LLMs to repeat their own hallucinations when conditioned on prior incorrect generations, a form of error propagation, is a significant inference-time issue addressed by methods like Chain-of-Verification ?. Mechanistically, hallucinations can also be linked to the model's internal uncertainty; ? found that entity-level hallucinations correlate with low predictive probability and high entropy, providing a real-time signal for detection. The very nature of maximum-likelihood training can lead LLMs to assign probabilities to non-factual information, making them prone to fabrication ?.

Deficiencies in knowledge representation and reasoning also play a crucial role. **?** attributed hallucinations to specific model capability deficiencies, including commonsense memorization, relational reasoning, and instruction following, drawing parallels to cognitive psychology. LLMs often struggle to effectively integrate and reason over structured knowledge, treating Knowledge Graphs (KGs) as plain text rather than leveraging their inherent graphical structure for robust reasoning **??**. This leads to difficulties in multi-hop reasoning and maintaining factual consistency. The problem of error propagation, where mistakes in early reasoning steps cascade into subsequent generations, further compounds these issues **?**. In dialogue systems, hallucinations frequently manifest as the injection of erroneous entities, indicating a failure in precise knowledge grounding **?**. Furthermore, dialogue-level hallucinations extend beyond mere factual errors to include incoherence (conflicting with input, context, or self), irrelevance, overreliance on specific information, and general reasoning errors, pointing to complex mechanistic failures in maintaining conversational state and consistency **?**. Attributing these diverse hallucinations to specific internal model behaviors or layers remains a significant challenge, often requiring sophisticated probing techniques.

Ultimately, a groundbreaking theoretical perspective suggests that hallucination is not merely an engineering bug but an **innate limitation of LLMs**. **?** provided a formal proof demonstrating the inevitability of hallucination for all computable LLMs, regardless of their architecture or training data. This fundamental insight shifts the paradigm from attempting to eliminate hallucination entirely to focusing on robust detection, quantification, and management strategies.

In conclusion, the root causes of hallucination are multifaceted, encompassing biases and limitations in training data, architectural constraints that hinder robust knowledge processing, and inference-time errors that propagate inaccuracies. These mechanistic insights, from the granular level of token generation uncertainty to the theoretical inevitability of errors, are indispensable for developing targeted, efficient, and truly impactful mitigation strategies that address the underlying mechanisms rather than just the symptoms. Future research must continue to unravel these complex interactions to build

more trustworthy and reliable LLM systems.

2.3 The Inevitability of Hallucination

The pervasive phenomenon of hallucination in Large Language Models (LLMs), where models generate plausible but factually incorrect or unfaithful information, has long been approached as a complex engineering problem. Initial research focused on empirically characterizing these errors, developing detection mechanisms, and devising mitigation strategies, with the implicit goal of eventually eradicating them. For instance, early work by ? systematically defined and categorized intrinsic and extrinsic hallucinations in abstractive summarization, highlighting their high prevalence and the inadequacy of traditional evaluation metrics. This foundational empirical understanding was extended to multi-modal contexts by ?, who studied "object hallucination" in Large Vision-Language Models (LVLMs) and introduced the `POPE` evaluation method to address the instability of prior metrics. Such efforts aimed to precisely identify the problem's manifestations, laying the groundwork for targeted solutions.

The field subsequently witnessed a proliferation of sophisticated detection and mitigation techniques, reflecting a concerted effort to engineer solutions to this perceived problem. Methods like `SelfCheckGPT` ? emerged, offering zero-resource, black-box hallucination detection by leveraging the consistency of stochastically sampled LLM outputs. Researchers also delved into the root causes, with ? quantifying and attributing hallucination to specific model capability deficiencies through association analysis, still within a problem-solving paradigm. Mitigation strategies became increasingly elaborate, ranging from retrieval-augmented generation (RAG) approaches like `ReAct` ? and `IRCoT` ? that ground LLMs in external knowledge, to methods enabling verifiable generation with citations ?. Other techniques involved self-correction mechanisms such as Chain-of-Verification (`CoVe`) ?, knowledge graph prompting (`MindMap`) ?, and dynamic retrieval augmentation (`DRAD`) ? that trigger retrieval based on real-time hallucination detection. Even novel decoding strategies like Induce-then-Contrast Decoding (`ICD`) ? were developed, aiming to penalize induced hallucinations to enhance factuality. Comprehensive

surveys by ?, ?, ?, ?, and ? meticulously cataloged these diverse efforts, underscoring the community's dedication to overcoming hallucination as an engineering challenge. Benchmarks like HaluEval 2.0 ?, HADES ?, AutoHall ?, ERBench ?, UHGEval ?, and DiaHalu ? were developed to rigorously evaluate progress in this fight.

However, a groundbreaking theoretical insight has fundamentally shifted this perspective, moving hallucination from a solvable engineering problem to an inherent, unavoidable characteristic of LLMs. ? provided a formal proof demonstrating that hallucination is an *inevitable* limitation for any computable LLM, regardless of architectural advancements or training data improvements. Employing diagonalization arguments, akin to those used to prove the undecidability of the Halting Problem, their work formalizes LLMs as total computable functions and defines hallucination as inconsistencies with a computable ground truth function. The proof establishes that no computable LLM can perfectly learn all computable functions, implying that there will always exist inputs for which any given LLM will generate factually incorrect or nonsensical outputs. This means that complete elimination of hallucination is mathematically impossible, setting a fundamental theoretical ceiling on the problem.

This pivotal realization necessitates a profound paradigm shift in how the AI community approaches dependability and credibility. Instead of pursuing the elusive goal of eradication, the research agenda is re-framed towards robust detection, effective mitigation, and responsible deployment of LLMs. While techniques like RGB ? continue to benchmark the limitations of even advanced mitigation strategies, confirming that LLMs still struggle with noise robustness, negative rejection, and information integration, the new theoretical understanding provides context. It suggests that these challenges are not merely temporary hurdles but reflections of an innate limitation. Future work must therefore focus on building systems that can reliably identify when an LLM is likely to hallucinate, actively correct or prevent such instances through external grounding ?????, and integrate self-reflection capabilities ? to manage, rather than eliminate, this inherent flaw. This paradigm shift acknowledges that LLMs are powerful but inherently fallible tools, demanding a new emphasis on designing trustworthy human-AI collaboration

frameworks where external verification and critical assessment are paramount **???**.

3 Benchmarking and Detection Methodologies

3.1 Reference-Free and Consistency-Based Detection

Detecting hallucinations in Large Language Models (LLMs) without relying on external ground truth or human-annotated references is crucial for evaluating proprietary models and scaling detection efforts where human annotation is impractical. These methods offer efficient and flexible assessment of factual correctness and internal coherence, focusing on the model's self-consistency or its ability to self-verify.

Early efforts laid the groundwork for fine-grained, reference-free detection. **?** introduced the first token-level, reference-free hallucination detection task, along with the HADES dataset, which was created by perturbing Wikipedia text and using an iterative model-in-the-loop annotation strategy. This pioneering work provided a benchmark for pinpointing hallucinations at a granular level without external human references, though its dataset was based on perturbed existing text rather than purely generative outputs.

Building on the need for black-box and zero-resource solutions, **?** proposed *SelfCheckGPT*, a seminal method that leverages the inherent stochasticity of LLMs. The core idea is that if an LLM genuinely "knows" a fact, multiple stochastically sampled responses to the same prompt will be consistent; conversely, hallucinated facts will likely lead to divergent or contradictory samples. *SelfCheckGPT* offers five variants for measuring informational consistency, including BERTScore, Question Answering, n-gram overlap, Natural Language Inference (NLI), and LLM Prompting, making it applicable to proprietary models without access to internal probabilities or external knowledge bases. However, *SelfCheckGPT* can sometimes struggle when LLMs tend to repeat their own hallucinations across samples, reinforcing incorrect information.

To address this limitation, **?** introduced *MetaQA*, a self-contained hallucination detection approach that significantly enhances consistency-based methods through the use of metamorphic relations (MRs) and prompt mutation. *MetaQA* generates diverse mu-

tations (synonymous and antonymous) from an LLM's initial response and then uses the LLM itself as a "test oracle" to verify the factual consistency of these mutations. This innovative application of MRs allows for more robust exposure of factual inconsistencies, with *MetaQA* empirically outperforming *SelfCheckGPT* across various LLMs and datasets by effectively preventing the model from reinforcing its own errors.

Beyond consistency-based sampling, other reference-free approaches focus on automated dataset generation and self-contained verification mechanisms. **?** developed *AutoHall*, a pipeline for automatically generating model-specific hallucination datasets from existing fact-checking data, eliminating the need for costly manual annotation. *AutoHall* also proposes a zero-resource, black-box detection method based on LLM-driven self-contradiction, where an LLM's initial response is compared against multiple independently generated references to identify inconsistencies. This approach offers a scalable solution for creating dynamic benchmarks that adapt to evolving LLMs.

Another category of "reference-free" methods leverages structured data or powerful LLMs as implicit verification oracles, removing the need for human ground truth annotation for every instance. **?** introduced *ERBench*, a novel benchmark construction framework that utilizes existing relational databases (RDBs) and their integrity constraints (e.g., Functional Dependencies, Foreign Key Constraints) to automatically generate complex, verifiable questions and rationales. This allows for automated verification of both the LLM's answer and its step-by-step reasoning against the database's inherent factual knowledge. Similarly, **?** proposed an LLM-based detection framework that uses advanced LLMs, such as GPT-4, to extract factual statements from a target LLM's responses and then judge their truthfulness, even considering interrelations between statements. While relying on a powerful external LLM, this method is "reference-free" from human annotation, offering a scalable way to evaluate factual accuracy.

In conclusion, reference-free and consistency-based detection methods represent a vital frontier in hallucination research, offering scalable and flexible solutions for evaluating LLMs, especially proprietary models. Techniques like *SelfCheckGPT* and *MetaQA* leverage the internal consistency of LLM outputs, while others like *AutoHall* and *ERBench*

automate dataset generation and verification against structured knowledge. Despite their promise, these methods still face challenges related to computational cost, the quality of stochastic sampling, the inherent capabilities of the LLMs used for self-verification, and the potential for biases even in automated or LLM-as-a-judge systems. Future research will likely focus on enhancing the robustness and efficiency of these self-contained detection mechanisms, further reducing their reliance on external resources, and expanding their applicability to more complex and nuanced forms of hallucination.

3.2 Fine-Grained and Rationale-Based Evaluation

Traditional evaluations of Large Language Model (LLM) outputs often rely on coarse-grained metrics that only assess overall answer correctness, failing to pinpoint the precise nature or location of factual errors and reasoning flaws. To foster greater confidence and enable targeted improvements, advanced evaluation methodologies have emerged, focusing on fine-grained analysis and the verifiability of an LLM's underlying reasoning process, or "rationales." These approaches move beyond simple correctness to provide a deeper, more granular understanding of LLM hallucinations, identifying errors at token or sentence levels and often leveraging structured data to automatically check logical steps.

Early efforts to characterize hallucination, such as the distinction between intrinsic and extrinsic types in abstractive summarization by **?**, laid the groundwork for more detailed analysis. A significant step towards fine-grained error localization was the introduction of HADES, the first token-level, reference-free hallucination detection benchmark for free-form text generation by **?**. This benchmark, created through contextual perturbation and iterative human annotation, enabled the identification of hallucinated tokens without relying on a ground-truth reference, addressing a critical limitation for real-time applications. Building on this, **?** developed the Neural Path Hunter, which employs a token-level hallucination critic to identify and mask erroneous entity mentions in dialogue systems, subsequently correcting them by retrieving facts from a Knowledge Graph (KG).

The need to verify an LLM's reasoning process directly led to the development of rationale-based evaluation. **?** introduced ALCE, a benchmark that enables LLMs to

generate text with explicit citations, using NLI-based metrics to automatically evaluate whether generated statements are supported by their cited passages. This marked an important shift towards making LLM claims verifiable. Further advancing this, ? proposed ERBench, a novel benchmark that leverages existing relational databases (RDBs) to generate complex, automatically verifiable questions and, crucially, to check the LLM's *rationales*. By utilizing database integrity constraints like Functional Dependencies (FDs) and Foreign Key Constraints (FKCs), ERBench can automatically verify if an LLM's step-by-step reasoning contains FD-inferred critical keywords, providing a robust mechanism to assess factual consistency and logical coherence. Similarly, ? introduced MindMap, a prompting pipeline that guides LLMs to reason over structured Knowledge Graph inputs and explicitly generate a "Graph of Thoughts" (or mind map), thereby making their reasoning pathways transparent and verifiable against KG facts. ? extended this concept with Chain-of-Knowledge, a framework that dynamically adapts knowledge from heterogeneous sources (structured and unstructured) and performs progressive rationale correction, preventing error propagation in multi-step reasoning. The OKGQA benchmark by ? further evaluates LLM+KG integration in open-ended question answering, employing metrics like FActScore and SAFE to measure factual precision and assess factual support, demonstrating how KGs can enhance trustworthiness.

A parallel development involves leveraging LLMs themselves as evaluators for fine-grained and rationale-based checks. ? quantified and attributed hallucination by analyzing its association with specific model capability deficiencies, moving towards understanding the *reasons* behind errors. ? introduced AutoHall, an automated dataset generation method, and a zero-resource hallucination detection technique based on LLM self-contradiction, where the model queries itself multiple times to check for consistency. This concept was significantly refined by ? with Chain-of-Verification (CoVe), a multi-step process where an LLM generates specific verification questions about its own initial response and then answers them independently to self-correct factual errors. This "factored" approach effectively reduces the LLM's tendency to repeat its own hallucinations. Extending this, ? proposed an iterative self-reflection methodology for medical gener-

ative question-answering, where the LLM systematically generates, scores, and refines both its background knowledge and answers for factuality and consistency, demonstrating strong performance in high-stakes domains. More recently, ? presented HaluEval 2.0, a benchmark that uses powerful LLMs like GPT-4 to extract factual statements from responses and judge their truthfulness, considering interrelations between statements for a fine-grained assessment. ? introduced MetaQA, a self-contained detection method that employs metamorphic relations and prompt mutation to expose factual inconsistencies by having the LLM verify transformed versions of its own output.

Beyond direct verification, comprehensive taxonomies and diagnostic benchmarks also contribute to fine-grained understanding. ? and ? provided extensive taxonomies of hallucination, categorizing error types and their underlying mechanisms. ? further refined this with a fine-grained taxonomy encompassing degree, orientation, and specific types of hallucination, alongside a Hallucination Vulnerability Index (HVI) for quantitative assessment. For Retrieval-Augmented Generation (RAG) systems, ? developed the RGB benchmark to diagnose specific LLM capabilities crucial for RAG, such as Noise Robustness, Negative Rejection, and Information Integration, offering a granular view of how LLMs process external knowledge. ? introduced DiaHalu, a dialogue-level benchmark with a comprehensive taxonomy (e.g., Non-factual, Incoherence, Reasoning Error) to evaluate hallucinations in multi-turn conversations, highlighting the need for context-aware, fine-grained analysis in dynamic interactions. Furthermore, mitigation strategies like the Coarse-to-Fine Highlighting by ? implicitly rely on fine-grained identification of key lexical units within retrieved contexts to improve factual grounding, demonstrating the interplay between fine-grained analysis and intervention. Similarly, ? developed DRAD, which uses Real-time Hallucination Detection (RHD) by analyzing the uncertainty of output entities to trigger targeted retrieval, showcasing fine-grained, real-time detection based on internal model signals.

In conclusion, the evolution of LLM evaluation has clearly moved towards more sophisticated, fine-grained, and rationale-based approaches. While significant progress has been made in identifying specific error locations, verifying reasoning steps through struc-

tured data, and leveraging LLMs for self-correction, challenges remain. The scalability of human annotation, the computational cost of multi-step verification, and the generalizability of detection methods across diverse hallucination types and domains are ongoing research areas. Future directions will likely focus on developing more efficient, robust, and universally applicable methods for fine-grained error diagnosis and transparent rationale verification, crucial for building truly trustworthy and reliable AI systems.

3.3 Benchmarking Retrieval-Augmented Generation (RAG)

The effective mitigation of Large Language Model (LLM) hallucination through Retrieval-Augmented Generation (RAG) systems critically depends on the development and application of rigorous, specialized evaluation benchmarks. While foundational work established the nature of hallucination and the necessity of external knowledge grounding **?**, the proliferation of RAG methods necessitated dedicated frameworks to systematically assess their performance, diagnose specific failure modes, and ensure the reliable utilization of external information. These benchmarks move beyond general LLM evaluation to scrutinize the complex interplay between retrieval and generation, directly addressing the challenges RAG introduces.

A pioneering effort in this domain is the **Retrieval-Augmented Generation Benchmark (RGB)** introduced by **?**. RGB is a multi-lingual corpus designed to systematically diagnose four fundamental RAG capabilities crucial for robust hallucination mitigation:

1. **Noise Robustness**: The ability of an LLM to generate accurate answers even when irrelevant or distracting information is present in the retrieved context.

2. **Negative Rejection**: The capacity of an LLM to correctly abstain from answering when no relevant information is available in the retrieved documents, preventing the generation of ungrounded content.

3. **Information Integration**: The skill of synthesizing facts from multiple retrieved documents to formulate a comprehensive and accurate answer.

4. **Counterfactual Robustness**: The resilience of an LLM to resist factually incorrect information present in the retrieved context, especially when it contradicts the model's internal knowledge.

? constructed RGB instances using recent news to minimize bias from LLMs' pre-trained knowledge, leveraging LLMs for QA generation and search APIs for document retrieval. Their empirical analysis revealed significant shortcomings in current state-of-the-art LLMs across these dimensions. For instance, models frequently failed to reject answers when only noisy documents were provided (Negative Rejection) and often prioritized factually incorrect retrieved information over their own correct internal knowledge (Counterfactual Robustness), even when explicitly warned. This highlights that while RAG offers a promising mitigation strategy, the effective utilization of external knowledge remains a substantial challenge, with LLMs often struggling to discern utility and truthfulness within the provided context.

Complementing RGB's diagnostic focus on core RAG capabilities, other specialized benchmarks address critical aspects of RAG output quality, verifiability, and the often-overlooked retrieval step. **?** developed **ALCE (Automatic LLMs' Citation Evaluation)**, which stands as the first reproducible benchmark for evaluating end-to-end RAG systems that not only generate answers but also provide verifiable citations. ALCE introduces novel NLI-based metrics for citation recall and precision, demonstrating that even advanced LLMs often lack complete citation support for their generated statements. This benchmark is crucial for assessing the trustworthiness and verifiability of RAG outputs, moving beyond mere factual correctness to accountability. While ALCE focuses on the *attribution* of generated content, **?** proposed **ERBench**, an Entity-Relationship based benchmark that leverages relational databases to construct complex, automatically verifiable questions. Crucially, ERBench assesses the correctness of an LLM's *rationale* alongside its final answer, providing a fine-grained approach to pinpoint where RAG-augmented LLMs falter in their reasoning process given the retrieved context. This contrasts with ALCE by scrutinizing the internal logical steps rather than just the external citation links.

A critical limitation of many RAG benchmarks, including RGB, ALCE, and ERBench,

is their primary focus on the *generator's* ability to utilize *provided* context, often assuming perfect or near-perfect retrieval. However, the success of RAG heavily relies on the quality of the retrieved passages themselves. Addressing this gap, **?** conducted a comprehensive study on the capabilities of LLMs in **utility evaluation for open-domain question answering**. They introduced a benchmarking procedure and a collection of candidate passages with varying characteristics to scrutinize how LLMs judge the *utility* of retrieved information in supporting question answering, distinguishing it from mere relevance. Their findings reveal that while well-instructed LLMs can differentiate between relevance and utility, they are highly susceptible to newly generated counterfactual passages. This work is vital as it directly evaluates the LLM's role in discerning the quality of the retrieval step, highlighting that a flawed retriever or an LLM's inability to correctly judge passage utility can independently lead to hallucinations, even if the generation component is otherwise robust.

Furthermore, for evaluating RAG in more realistic, unconstrained generation scenarios, **?** introduced **UHGEval** for Chinese LLMs, which prompts unconstrained continuations using "beginning text" from real news articles. Its novel `kwPrec` (keyword precision) metric, leveraging an LLM for factual keyword extraction, is particularly valuable for assessing the factual relevance of RAG outputs in open-ended contexts, where structured verification might be challenging.

Ultimately, these specialized benchmarks are indispensable for identifying specific failure modes within RAG systems, guiding improvements in both retrieval mechanisms and the LLM's ability to process and integrate external knowledge. They collectively push the field towards a more nuanced understanding of RAG performance, encompassing core capabilities, verifiability, rationale correctness, and the critical assessment of retrieval utility. However, current benchmarks often rely on synthetic data or specific QA formats, and may not fully capture the complexities of dynamic, multi-turn conversational RAG interactions or scenarios with conflicting information from multiple diverse sources. While these advancements provide powerful tools for evaluation, the theoretical understanding that hallucination is an inherent and inevitable limitation for all computable LLMs **?** un-

derscores that RAG systems, despite their sophistication, can only mitigate, not entirely eliminate, factual errors. Future research must therefore focus on developing more dynamic, adaptive, and fine-grained benchmarks that can simulate complex, real-world RAG interactions, proactively identify subtle failure modes in the retriever-generator interplay, and rigorously assess the system's ability to handle conflicting or ambiguous retrieved information, pushing towards more dependable AI applications.

4 Mitigation Strategies: External Grounding and Reasoning

4.1 Retrieval-Augmented Generation (RAG) and Knowledge Graphs

Large Language Models (LLMs) are prone to generating plausible but factually incorrect information, a phenomenon known as hallucination, primarily due to their reliance on potentially outdated or incomplete internal knowledge acquired during pre-training **???**. To mitigate this, Retrieval-Augmented Generation (RAG) and Knowledge Graphs (KGs) have emerged as pivotal strategies, providing LLMs with dynamic access to external, up-to-date, and verifiable information **??**. These techniques aim to ground LLM responses in factual evidence, thereby significantly enhancing factual consistency, credibility, and transparency.

The evolution of RAG, which augments LLMs with external knowledge retrieval, began by establishing foundational concepts for integrating external tools. Early paradigms demonstrated the power of allowing LLMs to interact with environments to gather information, thereby reducing hallucinations that arise from purely internal, ungrounded reasoning. Building on this, **?** introduced Interleaving Retrieval with Chain-of-Thought (IRCoT), a dynamic RAG approach that leverages intermediate Chain-of-Thought steps as queries for iterative knowledge retrieval. This method significantly reduced factual errors in multi-step question answering by up to 50

As RAG matured, research shifted towards more advanced, modular architectures that dynamically integrate knowledge and address the limitations of basic retrieval, such as noise, inefficiency, and context length. ? proposed Dynamic Retrieval Augmentation based on hallucination Detection (DRAD), a framework that synchronizes retrieval with real-time hallucination detection. DRAD's Real-time Hallucination Detection (RHD) component identifies potential hallucinations by analyzing entity-level uncertainty (low probability, high entropy) in the LLM's output, triggering targeted retrieval only when necessary. This conditional retrieval mechanism makes the RAG process more efficient and precise. To combat the challenge of LLMs "getting lost in long contexts" from retrieved documents, ? introduced Coarse-to-Fine highlighTing (COFT). COFT dynamically identifies and emphasizes key information at varying granularities (word, sentence, paragraph) within lengthy retrieved contexts, leveraging external Knowledge Graphs (KGs) like Wikidata for entity extraction and a novel contextual weighting mechanism, leading to over 30

Beyond unstructured text retrieval, Knowledge Graphs (KGs) offer a structured, explicit, and verifiable source of information, proving invaluable for enhancing factual consistency. Early integration efforts, such as Neural Path Hunter by ?, focused on reducing hallucination in dialogue systems. This generate-then-refine strategy employed a "Token-level Hallucination Critic" to identify erroneous entities and an "Entity Mention Retriever" to query a local k-hop subgraph of a KG for factual correction, resulting in a 39.98

However, integrating KGs presents unique challenges, including the knowledge acquisition bottleneck, inherent incompleteness, and the fidelity of verbalizing structured data. To address KG incompleteness and sparsity, ? proposed Logic Query of Thoughts (LGOT), which combines LLM reasoning with KG-based logic query reasoning to break down complex queries into subquestions, leveraging both sources to mitigate the limitations of each. For verifiable commonsense reasoning, particularly concerning long-tail

entities, ? introduced Right for Right Reasons (R^3). R^3 axiomatically surfaces LLM's intrinsic commonsense knowledge while grounding every factual reasoning step on KG triples, significantly reducing hallucination and reasoning errors in commonsense KGQA. Furthermore, to combat noise and irrelevant data often introduced when retrieving from extensive KGs, ? developed the Adaptive Multi-Aspect Retrieval-augmented over KGs (Amar) framework. Amar retrieves knowledge across entities, relations, and subgraphs, using a self-alignment module to enhance retrieved text and a relevance gating module to filter irrelevant information, achieving state-of-the-art performance on benchmarks like WebQSP and CWQ. In abstractive summarization, ? demonstrated how entity-linked KGs could mitigate hallucinations by providing external world knowledge, redefining faithfulness beyond mere document extractiveness through a generate-and-revise pipeline with a Fact Injected Language Model (FILM).

The synergy between RAG and KGs is increasingly explored to create more robust hybrid systems. KGs can guide the retrieval process, structure information, or serve as a verification layer for facts retrieved from unstructured text. For instance, as noted, COFT ? leverages KGs for entity extraction to improve highlighting in long contexts. ? explored hybrid RAG approaches, combining search engine and KG results with fine-tuning, demonstrating that a hybrid strategy achieved the highest score in the CRAG benchmark by allowing LLMs to say "I don't know" when uncertain, thereby reducing hallucination. Empirically, ? investigated the trustworthiness of KG-augmented LLMs in open-ended question answering, introducing the OKGQA benchmark. Their findings confirmed that KG integration generally reduces factual errors, particularly for queries requiring deeper reasoning, and demonstrated robustness even when KGs were partially contaminated, using a prize-cost strategy for efficient knowledge extraction.

Despite these significant advancements, RAG and KG integration still face considerable challenges. Benchmarks like ALCE by ?, designed for evaluating LLMs' ability to generate text with verifiable citations, revealed that even advanced models like GPT-4 struggle with complete citation support and multi-document synthesis, with approximately 50

In conclusion, RAG and Knowledge Graphs are indispensable tools for grounding LLMs in external, verifiable knowledge, significantly reducing hallucinations and improving factual accuracy. However, ongoing research is crucial to address the complexities of dynamic context management, multi-source synthesis, robust error handling, and the inherent limitations of KGs themselves. Future directions must focus on developing more sophisticated hybrid mechanisms for LLMs to discern reliable information, integrate diverse knowledge sources seamlessly, and provide transparent, verifiable outputs, ultimately fostering more trustworthy and reliable AI applications.

4.2 Synergizing Reasoning and Acting (ReAct)

The pervasive challenge of ungrounded hallucinations in Large Language Models (LLMs) necessitates a fundamental shift from purely internal, potentially fallacious, reasoning to paradigms that dynamically intertwine an LLM's internal cognitive processes ("thoughts") with external actions (e.g., API calls, tool use). This synergistic approach aims to dynamically gather information, interact with environments, and verify facts, thereby directly addressing the generation of plausible but factually incorrect content. By grounding LLM behavior in external validation, these frameworks enhance robustness, factual accuracy, and interpretability in complex, interactive tasks, moving beyond the limitations of relying solely on internal knowledge **?**.

A seminal contribution in this direction is the `ReAct` framework, introduced by **?**. `ReAct` addresses the inherent limitations of reasoning-only approaches, such as Chain-of-Thought (CoT), which are susceptible to fact hallucination and error propagation due to their exclusive reliance on internal representations. By prompting LLMs to generate both verbal reasoning traces ("thoughts") and task-specific actions in an interleaved manner, `ReAct` enables a dynamic loop: "reason to act" (formulating plans based on current reasoning) and "act to reason" (incorporating observations from external interactions to refine subsequent reasoning). This grounding in external, verifiable information, such as through a Wikipedia API, significantly mitigates hallucination and enhances interpretability, as

demonstrated across knowledge-intensive reasoning (e.g., HotpotQA, FEVER) and interactive decision-making benchmarks (e.g., ALFWorld, WebShop). The explicit thoughts provide a human-aligned and diagnosable decision-making process, allowing for inspection of factual correctness and reasoning paths ?.

Building upon this principle of dynamic external grounding, ? further refined the interaction between reasoning and retrieval with `IRCoT` (Interleaved Retrieval with Chain-of-Thought Reasoning). While `ReAct` established the general synergy, `IRCoT` specifically leverages intermediate Chain-of-Thought steps as dynamic queries for iterative knowledge retrieval. This adaptive information-seeking process, where each reasoning step informs subsequent retrieval and newly retrieved facts ground further reasoning, significantly reduces factual errors and improves the accuracy of multi-step question answering, particularly in few-shot settings. This demonstrates a more fine-grained integration of external knowledge within the reasoning process, directly extending the `ReAct` paradigm's core mechanism.

The `ReAct` paradigm extends beyond simple API calls to encompass more complex tool use and agentic behaviors, proving critical for ensuring the reliability of LLM-generated content in practical applications. For instance, in automated code development, LLMs can generate code (an action) but require external verification to mitigate hallucination. ? describe Meta's `TestGen-LLM` tool, which uses LLMs to improve existing unit tests. Crucially, `TestGen-LLM` verifies its generated test classes against a set of filters to ensure measurable improvement and eliminate problems due to LLM hallucination, exemplifying the "act to reason" principle where external execution and verification ground the LLM's creative output. Similarly, a survey by ? highlights that agent planning and iterative refinement, core components of the `ReAct` paradigm, are essential for hallucination mitigation in various code development tasks, from generation to testing and documentation.

Furthermore, the concept of LLM agents interacting within environments and managing internal states to avoid hallucination is explored in multi-agent contexts. ? investigate LLM-based agents in cooperative text games, observing emergent collaborative behaviors but also systematic failures in planning optimization and "hallucination about the task

state." They demonstrate that using explicit belief state representations can mitigate these issues, enhancing task performance and the accuracy of Theory of Mind inferences. This underscores the importance of grounding not just facts, but also the agent's understanding of its own state and the environment, a direct extension of `ReAct`'s interactive grounding principle.

Despite the significant advantages, the effectiveness of tool-augmented and agentic LLMs is not without specific challenges related to their synergistic nature. Diagnosing hallucinations in these complex systems requires specialized benchmarks that go beyond general factual accuracy. **?** introduced `ToolBeHonest`, a multi-level diagnostic benchmark specifically designed for tool-augmented LLMs. `ToolBeHonest` assesses hallucinations through both depth (solvability detection, solution planning, missing-tool analysis) and breadth (scenarios with missing, potential, or limited functionality tools). Their findings reveal that advanced models like Gemini-1.5-Pro and GPT-4o still face significant challenges, particularly in accurately assessing task solvability. This suggests that the LLM's internal reasoning about *when and how* to use tools, and its ability to correctly interpret tool outputs, remains a critical bottleneck. The benchmark also indicates that verbose replies from open-weight models can degrade performance, while proprietary models excel with longer reasoning, highlighting the nuanced interplay between reasoning verbosity and tool-use reliability.

In summary, the `ReAct` paradigm and its successors represent a powerful shift towards more robust, factually grounded, and interpretable LLMs by dynamically integrating internal reasoning with external actions and observations. While these systems offer substantial improvements in mitigating ungrounded hallucinations through external validation and iterative refinement, challenges persist in the LLM's ability to optimally plan tool use, interpret complex environmental feedback, and maintain coherent internal states in interactive settings. Future research in this area will likely focus on enhancing the LLM's meta-reasoning capabilities regarding tool selection and application, improving robustness to noisy or ambiguous external information, and developing more sophisticated mechanisms for self-correction within these interactive loops.

4.3 Self-Correction and Verification Mechanisms

Large Language Models (LLMs) frequently generate outputs that, despite their fluency, can contain factual inaccuracies or logical inconsistencies, a phenomenon known as hallucination **??**. To enhance the intrinsic dependability and logical coherence of LLM outputs, a critical area of research focuses on developing mechanisms that enable these models to critically evaluate their own generated content, identify potential errors, and perform revisions. Comprehensive surveys by **?** and **?** highlight automated self-correction as a pivotal strategy, categorizing diverse techniques, while **?** provides a broader overview of mitigation strategies, including self-reflection and prompt engineering.

One major category of self-correction mechanisms relies on the LLM's internal reasoning and consistency checks to identify and rectify errors without direct external grounding. A prominent approach is *Chain-of-Verification (CoVe)* **?**, which enables LLMs to systematically self-critique factual claims through a multi-step process: generating a baseline response, planning specific verification questions, executing these verifications, and finally generating a revised, fact-checked response. The "factored" variant of CoVe is particularly notable, as it ensures verification questions are answered independently of the initial potentially hallucinated response, minimizing the risk of repeating errors. Building on such multi-step reasoning, **?** proposes an iterative self-reflection methodology, particularly effective in high-stakes medical question-answering, where the LLM iteratively generates, scores, and refines both background knowledge and answers for factuality and consistency. For more complex reasoning tasks, **?** introduced Multi-Aspect Feedback (MAF), an iterative refinement framework that integrates multiple feedback modules (including frozen LMs) to address diverse error types within reasoning chains, such as logical inconsistencies and mathematical errors. A purely self-contained method, MetaQA **?**, leverages metamorphic relations and prompt mutation, allowing the LLM to act as its own "test oracle" by generating and verifying diverse mutations of its response to expose factual inconsistencies. These internal methods offer the advantage of not requiring external tools, making them broadly applicable and efficient when external resources are unavailable. However, their effectiveness is inherently limited by the LLM's internal knowledge and

reasoning capabilities, and they can incur substantial computational costs due to multiple inference steps, especially for complex or long-form generations.

Beyond internal consistency, a crucial aspect of self-correction involves LLMs recognizing and acting upon their uncertainty, either by triggering further verification or by appropriately abstaining from answering. Researchers have explored quantifying LLM uncertainty through various internal signals, such as low predictive probability of output tokens, high semantic entropy **?**, or analyzing inference dynamics across model layers **???**. For instance, **?** proposes a token-level uncertainty quantification method, Claim Conditioned Probability (CCP), to fact-check atomic claims in LLM outputs, demonstrating strong improvements in hallucination detection by focusing on the uncertainty of a particular claim value. Building on this, **?** introduced Dynamic Retrieval Augmentation based on Hallucination Detection (DRAD), which uses Real-time Hallucination Detection (RHD) to identify potential hallucinations by analyzing the uncertainty of output entities. This allows for targeted and efficient self-correction by conditionally retrieving and incorporating external knowledge precisely when needed. More directly addressing the strategy of abstention, **?** proposes CoKE, a method to teach LLMs to recognize and express their "knowledge boundary." CoKE probes LLMs' internal confidence to elicit explicit expressions of ignorance, enabling them to answer known questions while declining unknown ones, thereby reducing hallucinations caused by fabrication. Similarly, **?**'s Real-time Verification and Rectification (EVER) framework, which performs step-wise validation during generation, explicitly handles extrinsic hallucinations by flagging unverified content with a "not sure" warning or by abstaining from answering, significantly enhancing trustworthiness. While these uncertainty-driven approaches offer a powerful mechanism for improving reliability and accountability, accurately quantifying and interpreting LLM uncertainty remains a complex challenge, particularly in diverse and open-ended generation tasks where the model's "knowledge" is implicit and dynamic.

While Section 4.1 discussed Retrieval-Augmented Generation (RAG) and Knowledge Graphs (KGs) as primary external grounding strategies, these external resources can also be integral to an LLM's self-correction *process*. Here, the focus shifts from merely re-

trieving information to how the LLM *uses* that information to *critique and revise its own output*. For example, **?** introduces MindMap, a prompting pipeline where LLMs comprehend and reason over structured KGs, aggregating evidence from KGs and their implicit knowledge to build a "mind map" for transparent reasoning. This effectively enables self-evaluation and correction by grounding the LLM's thoughts in explicit facts. Similarly, Chain-of-Knowledge (CoK) **?** dynamically adapts knowledge from heterogeneous sources (structured and unstructured) and employs a progressive rationale correction mechanism, rectifying preceding reasoning steps before generating subsequent ones to prevent error propagation. In dialogue systems, **?**'s Neural Path Hunter uses a generate-then-refine strategy, where a hallucination critic identifies and masks erroneous entities in a generated response, which are then corrected by retrieving facts from a KG. The empirical study by **?** further reinforces the value of KGs in reducing hallucinations and improving trustworthiness in open-ended question answering, even with partially contaminated knowledge, by providing a verifiable external source for self-correction. These externally-guided self-correction methods often achieve superior factual accuracy by leveraging up-to-date and verifiable information, but they introduce dependencies on external infrastructure, increase computational overhead, and require robust mechanisms for integrating and reasoning over diverse knowledge sources, posing a trade-off with the self-contained nature of purely internal methods.

Despite these advancements, significant challenges remain in self-correction and verification. The computational cost of multi-step reasoning and iterative refinement can be substantial, especially for complex tasks or real-time applications. Generalizability across diverse hallucination types and domains, beyond factual errors, still requires further research, particularly for logical inconsistencies or creative generation. While mechanisms for self-correction and verification significantly improve LLM dependability, the inherent limitations of LLMs mean that complete elimination of hallucination might be theoretically impossible, as suggested by **?**. Future work must focus on developing more efficient, robust, and universally applicable self-correction strategies that seamlessly integrate internal reasoning with external knowledge, enhance the accuracy and interpretability of

uncertainty quantification, and ensure greater accountability and trustworthiness in autonomous LLM operations.

5 Mitigation Strategies: Internal Model Interventions and Training

5.1 Decoding-Time Strategies and Logit Manipulation

Decoding-time strategies represent a crucial and efficient paradigm for mitigating hallucinations in Large Language Models (LLMs) and Multimodal Large Language Models (MLLMs). These techniques intervene directly during the token generation process, offering a training-free and flexible alternative to costly model retraining. The core principle involves dynamically shaping the model's output probability distribution (logits) to steer generation towards more factual and coherent content, often by leveraging internal uncertainty signals or external factual checks.

Traditional decoding methods, such as greedy decoding or beam search, prioritize high-probability sequences but can inadvertently amplify statistical biases or language priors, leading to hallucinations **?**. To counter this, advanced decoding strategies primarily fall into two categories: **contrastive decoding**, which penalizes tokens inconsistent with a reference or a subtly altered input, and **direct logit steering**, which explicitly modifies token probabilities based on various signals.

5.1.1 Contrastive Decoding for Factual Consistency

Contrastive decoding methods operate by comparing the output distribution of the primary model (or a subtly altered version of the input) with a "negative" or "reference" signal to suppress hallucinated content. This typically involves multiple forward passes, where the difference in log probabilities guides the token selection.

A prominent approach, **Visual Contrastive Decoding (VCD) ?**, addresses object hallucinations in MLLMs by introducing visual uncertainty. VCD applies a Gaussian noise

mask to the original image, creating a distorted input. It then contrasts the log probabilities from the original and distorted inputs, penalizing tokens that are highly probable under the distorted (hallucination-prone) input. This mechanism effectively calibrates the MLLM's over-reliance on language priors and statistical biases, demonstrating significant improvements on benchmarks like POPE and MME. To prevent VCD from promoting implausible tokens, an adaptive plausibility constraint is applied, pruning candidate tokens based on the confidence of the original output distribution.

Building on this, *ConVis* **?** introduces a novel training-free contrastive decoding method for MLLMs that leverages a Text-to-Image (T2I) model. By reconstructing an image from the MLLM's initial caption and using this T2I-generated image as a contrastive signal, *ConVis* penalizes tokens corresponding to visually represented hallucinations. This innovative use of an external generative model as a "visual critic" achieves state-of-the-art performance on benchmarks like CHAIR and POPE.

Beyond visual modalities, contrastive decoding has been adapted for unimodal LLMs and specialized domains. *Induce-then-Contrast Decoding (ICD)* **?** proposes creating a "factually weak LLM" by fine-tuning the original model on non-factual samples. During inference, the log probabilities of these induced hallucinations from the weak model are subtracted from the original model's predictions, thereby amplifying truthful outputs. ICD also incorporates an adaptive plausibility constraint to selectively penalize only potentially untruthful tokens, preserving overall generation quality. This method has shown remarkable improvements in factuality on TruthfulQA and FACTSCORE, enabling smaller models to rival larger ones. Similarly, *ALternate Contrastive Decoding (ALCD)* **?** addresses hallucinations in medical information extraction (MIE) by separating identification and classification functions into sub-task models. During inference, ALCD alternately contrasts output distributions from these sub-task models, selectively enhancing specific capabilities while minimizing the influence of other inherent LLM abilities, further demonstrating the versatility of contrastive signals.

A more advanced form of contrastive decoding, **Hallucination-Induced Optimization (HIO) ?**, refines the concept of a "negative" model. HIO involves a training stage to

create an "Evil LVLM" by intentionally inducing hallucinations using a *reversed* Bradley-Terry model. This "Evil LVLM" is trained to *prioritize* hallucinatory content, amplifying the logits of incorrect tokens. During inference, logits from this "Evil LVLM" are contrasted with the original LVLM to precisely reduce hallucinations. HIO's theoretical analysis emphasizes the need for consistent logit differences between potential hallucinated and correct tokens, and its Amplification of Multiple Targeted Hallucination (AMTH) technique allows for simultaneous training against diverse hallucination types. While HIO involves a pre-training step for the "Evil LVLM," its core mechanism is a decoding-time contrast, making it a powerful extension of this paradigm.

A common trade-off in contrastive decoding is the potential for increased inference time due to multiple forward passes and the risk of degrading the quality or coherence of generated content by overly penalizing tokens. Adaptive constraints and careful tuning of contrastive strength (e.g., and parameters in VCD and ICD) are crucial to balance hallucination reduction with maintaining linguistic quality.

5.1.2 Direct Logit Steering and Prompt-Derived Influence

Beyond explicit contrastive signals, other decoding-time strategies directly manipulate the logit distribution or influence it through prompt-derived information without requiring deep architectural interventions.

OPERA (Over-trust Penalty and Retrospection-Allocation) **?** is a novel MLLM decoding method applied during beam-search. It is grounded in the observation that MLLM hallucinations often stem from "over-trust" in a few "summary tokens" within the self-attention matrix, leading to neglect of visual information. OPERA introduces a penalty term on the model logits during beam-search decoding, derived from a column-wise metric on the self-attention map that indicates the "over-trust degree." Furthermore, its Retrospection-Allocation Strategy allows the decoding process to "roll back" and re-select better candidates if strong over-trust patterns are detected, dynamically correcting the generation path. This method effectively mitigates hallucinations without additional training, data, or external knowledge.

Similarly, *MVP* (Multi-View Multi-Path Reasoning) ? enhances image comprehension through multi-view information seeking and guides decoding via a "certainty score" (based on logit differences) across multiple paths. This training-free and tool-free framework directly leverages logit-based certainty to alleviate hallucinations by addressing both incomplete visual understanding and low certainty during token generation.

Prompt-based steering methods, while not directly manipulating logits in a mathematical sense, influence the model's internal state such that subsequent token probabilities are altered. *Counterfactual Inception* ? is a training-free method that prompts MLLMs to self-generate "counterfactual keywords" (deliberately deviating from visual content). These keywords are then filtered and incorporated into a prompt that instructs the MLLM to avoid them, thereby guiding the model towards more factual outputs by implicitly steering the logit space. *Order Matters in Hallucination* ? highlights how the sequence of reasoning in prompts can significantly impact an LLM's consistency and factual accuracy. By generating reasoning first and then the conclusion, models are less prone to fabricating answers, indicating that prompt structure can indirectly steer the internal generation process and thus the logit probabilities towards more reliable outcomes.

It is crucial to distinguish these methods from those that involve deeper manipulation of internal model states or causal interventions on attention mechanisms (covered in Subsection 5.3), or those that rely heavily on external tools and iterative self-correction (covered in Section 4.3). The techniques discussed here primarily focus on shaping the final token probabilities through contrastive signals or direct logit adjustments, often with minimal changes to the model's internal architecture or complex external interactions.

In conclusion, decoding-time strategies represent a vital and evolving area for mitigating hallucinations, offering efficient, training-free solutions. Approaches range from innovative contrastive decoding methods like VCD ?, ConVis ?, ICD ?, ALCD ?, and HIO ?, which leverage diverse "negative" signals, to sophisticated logit-based steering techniques such as OPERA ? and MVP ?, which directly adjust token probabilities based on internal observations or certainty. Furthermore, prompt-derived influences ?? offer flexible means of guiding model outputs. While significant progress has been made in im-

proving factual consistency and inference efficiency, ongoing challenges include ensuring generalizability across diverse hallucination types, reducing reliance on external proprietary models, and developing more robust internal uncertainty quantification mechanisms to optimize logit adjustments without compromising overall generation quality. Future directions will likely focus on more sophisticated negative signal generation, adaptive logit scaling, and hybrid approaches that seamlessly integrate these decoding-time interventions with other mitigation paradigms.

5.2 Training-Based Approaches and Preference Optimization

Embedding hallucination resistance directly into the learning phases of Large Language Models (LLMs) and Multimodal Large Language Models (MLLMs) represents a proactive strategy to build robustness from the ground up, reducing reliance on post-hoc corrections. These training-based approaches encompass modifications to pre-training objectives, robust instruction tuning with carefully curated datasets, and sophisticated preference optimization techniques that align model behavior with desired factual accuracy.

Early efforts focused on refining pre-training objectives to enhance visual grounding. **?** systematically investigated object hallucination in Vision-Language Pre-training (VLP) models, revealing that optimizing for standard metrics like CIDEr could paradoxically increase hallucination. To counter this, they proposed ObjMLM (Object-Masked Language Modeling), a novel pre-training objective designed to improve token-level image-text alignment and controlled generation, thereby reducing object hallucination by up to 17.4

Beyond specific pre-training or fine-tuning data generation, some methods introduce generalizable training techniques. **?** proposed NoiseBoost, a simple yet effective method that injects Gaussian noise into projected visual tokens during Supervised Fine-tuning (SFT), Reinforcement Learning (RL), or Semi-Supervised Learning (SSL). This perturbation encourages MLLMs to distribute attention more evenly between visual and linguistic tokens, alleviating over-reliance on language priors and improving hallucination accuracy by 8.1

More advanced strategies leverage preference optimization, often incorporating AI-generated or human-annotated feedback to align model behavior. ? introduced Hallucination Severity-Aware Direct Preference Optimization (HSA-DPO), a novel preference learning approach for Large Vision Language Models (LVLMs). Their method generates fine-grained, sentence-level AI feedback (including hallucination type, severity, and rationale) using proprietary models, then trains an open-source detection model to enable a "detect-then-rewrite" pipeline for cost-effective preference dataset construction. HSA-DPO significantly reduced hallucination rates on benchmarks like AMBER and Object HalBench, outperforming state-of-the-art mitigation models by explicitly prioritizing the mitigation of critical hallucinations. Addressing the computational expense and potential biases of using large models as judges for preference data, ? proposed an "efficient self-improvement" framework for Multimodal Large Language Models (MLLMs) that is "model-level judge-free." This approach generates controllable negative samples by blending conditional and unconditional decoding paths, then uses a lightweight CLIP model for robust preference data inversion and quality control before applying Direct Preference Optimization (DPO). This judge-free method achieves superior precision and recall in hallucination control with significantly lower computational demands, offering a scalable pathway for MLLM self-improvement.

In summary, training-based approaches have evolved from modifying foundational pre-training objectives to sophisticated fine-tuning and preference optimization techniques. While early work highlighted the need for better visual grounding ?, subsequent research focused on enriching training data with fine-grained details ? or introducing generalizable regularization during training ?. The advent of preference optimization, particularly with fine-grained AI feedback ? and judge-free mechanisms ?, marks a significant step towards instilling intrinsic hallucination avoidance. However, challenges remain in ensuring the scalability and unbiased nature of synthetic data generation, managing the computational costs of iterative training, and preventing the introduction of new biases from reward models or lightweight verifiers. Future directions will likely involve developing more robust

and efficient methods for generating high-quality preference data, exploring hybrid training paradigms, and integrating diverse signals to further enhance factual accuracy and trustworthiness.

5.3 Internal State Manipulation and Causal Interventions

Advanced mitigation techniques for Large Language Models (LLMs) are increasingly moving beyond superficial prompt engineering or costly retraining, focusing instead on directly manipulating the model's internal states, attention mechanisms, or learned parameters. These fine-grained interventions offer mechanistic insights into model failures, enabling more precise and targeted corrections that enhance factual accuracy and reduce spurious correlations. While many prominent examples of these techniques arise in Multimodal Large Language Models (MLLMs) due to the explicit need to balance information across modalities, the underlying principles of internal state manipulation are broadly applicable to unimodal LLMs as well.

One significant thrust in this area involves directly enhancing or modifying specific modality features within the model's processing pipeline, particularly in MLLMs to combat modality bias or "visual amnesia" **?**. For instance, **?** introduced Visual Amplification Fusion (VAF), a training-free method that enhances attention to visual signals within the middle layers of MLLMs, where modality fusion predominantly occurs. VAF aims to counteract language bias by boosting visual feature processing without the inference speed penalties often associated with contrastive decoding methods. Building on this, **?** proposed Memory-space Visual Retracing (MemVR), which dynamically re-injects visual tokens as supplementary evidence into the MLLM's Feed Forward Network (FFN) at a "middle trigger layer" when the model exhibits high uncertainty. This adaptive re-injection refreshes visual memory, effectively combating "visual amnesia" without significant inference overhead. Further refining targeted interventions, **?** developed Image-Object Cross-Level Trusted Intervention (ICT), a plug-and-play method that calculates an "intervention direction" to shift the model's focus towards different levels of visual information. ICT operates during the forward pass by identifying and intervening on spe-

cific attention heads responsible for encoding overall image information and fine-grained object details. While these methods offer efficient, training-free solutions, their interventions are often heuristic, relying on empirical observations of where and how visual information degrades. The challenge lies in ensuring these boosts do not inadvertently over-emphasize visual cues, potentially suppressing valid linguistic context or introducing new biases.

Moving beyond direct feature boosting, another line of research employs causal inference and targeted manipulation of attention mechanisms to understand and balance modality priors or correct specific hallucination types. **?** presented CAUSAL MM, a principled causal inference framework that applies structural causal modeling (SCM), back-door adjustment, and counterfactual reasoning at both visual and language attention levels. By treating modality priors as confounding factors, CAUSAL MM deciphers their causal impact on MLLM output, enabling a more balanced integration of visual and linguistic information to mitigate hallucinations. This represents a more theory-driven approach compared to purely heuristic interventions. Similarly, in unimodal LLMs, **?** analyzed "false premise hallucinations" and identified a small subset of attention heads ("false premise heads") that disturb knowledge extraction. Their method, FAITH (False premise Attention head constraIning for miTigating Hallucinations), constrains these specific attention heads during inference, demonstrating a significant performance increase by manipulating only about 1

Beyond real-time inference interventions, internal state manipulation can also be applied to directly edit or update the factual knowledge embedded within a model's parameters or representations. This approach, often termed "model editing," aims to precisely alter specific behaviors without affecting unrelated knowledge. For instance, **?** proposed MedLaSA (Layer-wise Scalable Adapter strategy) for medical model editing. MedLaSA leverages causal tracing to identify the association of knowledge in neurons across different layers, generating a corresponding scale set. It then incorporates scalable adapters into the dense layers of LLMs, adjusting adapter weights and ranks based on the specific

knowledge. This allows for precise editing of semantically identical knowledge while minimizing impact on unrelated information, crucial for high-stakes domains like medicine. In a similar vein, **?** introduced DAFNet (Dynamic Auxiliary Fusion Network) for sequential model editing, which addresses the challenge of continuously rectifying mistakes and preventing catastrophic forgetting. DAFNet enhances semantic interaction among factual knowledge by aggregating intra-editing attention flow into auto-regressive self-attention and leveraging multi-layer diagonal inter-editing attention flow to update sequence-level representations. These methods demonstrate how direct manipulation of internal parameters and attention flow can be used to refine or update factual knowledge, offering a powerful alternative to full retraining. However, challenges remain in ensuring the scalability of these editing operations, preventing unintended side effects on other knowledge, and maintaining consistency over sequential edits.

Mechanistic insights into model failures can also be gleaned from adversarial attacks that specifically manipulate internal states to induce hallucinations. **?** demonstrated a novel hallucination attack that exploits the "attention sink" phenomenon in MLLMs. By manipulating attention scores and hidden embeddings to induce these sinks, the attack triggers hallucinated content with minimal image-text relevance. Understanding how such internal vulnerabilities can be exploited to *induce* hallucinations provides critical insights for developing more resilient internal state manipulation techniques, effectively turning offensive research into defensive strategies.

In conclusion, these advanced techniques represent a significant shift towards achieving fine-grained control and mechanistic understanding of LLM and MLLM behavior. By directly intervening in hidden states, attention mechanisms, or leveraging causal relationships and parameter modifications, researchers are developing powerful, often training-free, solutions to mitigate hallucinations. This allows for more precise and targeted corrections, moving beyond black-box approaches to address the root causes of factual inaccuracies. However, challenges remain in fully understanding the complex causal relationships within these models, ensuring generalizability across diverse tasks and architectures, and balancing the trade-offs between intervention strength and preserving beneficial model

capabilities. Future work will likely focus on more adaptive and intelligent internal interventions, potentially informed by a deeper causal understanding of multimodal and unimodal reasoning, and ensuring that these interventions are robust and scalable.

6 Applications, Specialized Domains, and Multimodal Hallucination

6.1 Characterizing Multimodal and Domain-Specific Hallucinations

Hallucinations in multimodal large language models (MLLMs) represent a complex and multifaceted challenge, distinct from those encountered in unimodal Large Language Models (LLMs). This distinction primarily stems from the inherent "modality gap" and the intricate interplay between diverse input types, such as vision, audio, and text ????. MLLMs, including Large Vision-Language Models (LVLMs), Large Audio-Language Models (LALMs), and Audio-Visual Language Models (AV-LLMs), frequently generate outputs that are factually inconsistent with their non-textual inputs, necessitating a specialized understanding of their unique failure modes.

A foundational and extensively studied form of multimodal hallucination is *object hallucination*, where models describe objects that are entirely absent from the visual input. Pioneering work by ? systematically investigated this phenomenon in LVLMs, identifying object frequency and co-occurrence in training data as significant drivers. This initial characterization has since been refined, moving beyond simple object presence to more nuanced errors. For instance, ? extended the analysis to *multi-object hallucination*, revealing the increased complexity and ambiguity when models struggle to simultaneously recognize and reason about multiple entities. Further, ? delved into *fine-grained object hallucination*, where models misrepresent subtle object attributes or behaviors, highlighting a deeper deficiency in perceptual grounding beyond mere existence.

Beyond individual objects, MLLMs exhibit significant struggles in accurately interpret-

ing relationships. ? introduced R-Bench to systematically evaluate *relationship hallucinations* in LVLMs, uncovering a fundamental flaw in models' understanding of inter-object dynamics and their propensity to over-rely on common sense rather than visual evidence. This was further elaborated by ?, who categorized these into *perceptive* (concrete, directly observable) and *cognitive* (abstract, inferential) relation hallucinations with their Reefknot benchmark. Their findings critically demonstrated that MLLMs are often more susceptible to errors in perceiving concrete relationships, suggesting a weakness in direct visual interpretation rather than just high-level reasoning.

The integration of dynamic modalities like video and audio introduces novel categories of hallucinations related to temporal consistency. For video-language models, ? characterized *temporal hallucinations*, distinguishing between intrinsic (contradicting explicit video content) and extrinsic (unverifiable from the video) types, underscoring the challenge of maintaining factual accuracy across sequential frames. Similarly, ? developed VIDHALLUC to specifically assess *action, temporal sequence, and scene transition hallucinations* in video understanding, revealing models' difficulties in discerning visually similar yet semantically different video segments. In the audio domain, ? explored *object hallucination* in LALMs, noting that while models can generate descriptive captions, their performance in discriminative tasks often falters, indicating a weakness in grounding query understanding to audio processing. The most intricate scenarios involve *cross-modal driven hallucinations*, as characterized by ? with AVHBench. Here, models misinterpret information due to subtle relationships or an over-reliance on one modality, such as perceiving imaginary sounds from visual cues or fabricating visual events based on audio inputs, demonstrating a critical failure in multimodal fusion and consistency.

More intricate forms of hallucination arise from the entanglement of language and perception within complex reasoning tasks. ? identified *event hallucination*, where LVLMs invent fictional entities and construct entire narratives around them, a phenomenon observed to escalate with output length and complexity. ? proposed Hallusionbench to diagnose entangled *language hallucination* (over-reliance on language priors) and *visual illusion* (misinterpreting visual inputs) in LVLMs. Their findings revealed a pronounced

language bias in even state-of-the-art models, where linguistic patterns often override direct visual evidence. This problem is further compounded in interactive settings, where ? identified *multimodal hallucination snowballing*, demonstrating that an LVLM's own generated hallucinations can mislead subsequent responses, creating a cascade of errors. Moreover, ? introduced VisDiaHalBench to specifically diagnose hallucination in visual dialogue settings, highlighting how long-term misleading textual history can induce errors. As MLLMs become globally deployed, ? introduced CCHall to evaluate *joint cross-lingual and cross-modal hallucinations*, demonstrating that models struggle significantly more when aligning visual content with text across diverse languages, indicating a complex interplay of linguistic and multimodal biases.

Understanding the root causes of these diverse hallucinations is paramount. A recurring theme in the literature is the *over-reliance on language priors* and the *modality imbalance* between powerful language model components and comparatively weaker visual or audio encoders ??????. This imbalance often leads models to prioritize statistical patterns from text over direct perceptual evidence. For instance, the "language hallucination" observed by ? directly stems from this bias, where models generate plausible but visually unfounded statements. Furthermore, issues like "visual amnesia" ?, where models lose track of visual information during autoregressive decoding, contribute to object and attribute misrepresentations. The internal attention mechanisms also play a critical role, with studies showing how modality priors can negatively impact output quality via attention, leading to biases that exacerbate hallucinations ?. Adversarial attacks, such as those exploiting "attention sink" phenomena, further reveal how internal mechanisms can aggregate misleading information, inducing or amplifying hallucinations ?. The inherent "toxicity in datasets" and "LLM hallucinations" inherited from the underlying language model components also contribute significantly to these multimodal errors ?.

The criticality of hallucinations becomes particularly acute in *domain-specific applications*. For instance, ? highlighted the severe risks of *medical hallucinations* in LVLMs used for medical Visual Question Answering (VQA) or imaging report generation. They introduced Med-HallMark, a benchmark with a novel hierarchical categorization of hallu-

cinations (e.g., Catastrophic, Critical, Attribute, Prompt-induced, Minor) based on their clinical severity, emphasizing the need for context-aware analysis and specialized detection methods beyond general-domain approaches. Similarly, **?** introduced MedVH, another benchmark for medical visual hallucination, revealing that domain-specific LVLMs, despite their promising performance on standard medical tasks, are often more susceptible to hallucinations than general models, raising significant concerns about their reliability in clinical settings. These findings underscore that the unique characteristics and high stakes of specialized domains necessitate tailored frameworks for characterizing and assessing hallucination.

In conclusion, the landscape of hallucination in multimodal and domain-specific models is vastly more intricate than in unimodal LLMs. The "modality gap" and the interplay of diverse inputs give rise to distinct phenomena such as object, relation, temporal, and cross-modal hallucinations, often compounded by issues like fine-grained attribute errors, event fabrication, and cross-lingual inconsistencies. These errors are frequently rooted in an over-reliance on language priors, modality imbalance, and challenges in maintaining visual or audio grounding during generation. While significant progress has been made in characterizing these diverse failure modes, the challenge of achieving truly factual and trustworthy multimodal AI systems, particularly in safety-critical domains, remains a paramount and active area of research. This detailed characterization provides a crucial foundation for the subsequent development of robust evaluation benchmarks.

6.2 Evaluation Benchmarks for Multimodal and Domain-Specific Hallucinations

The increasing sophistication and deployment of Multimodal Large Language Models (MLLMs) necessitate the development of specialized benchmarks and metrics to rigorously assess and quantify the diverse forms of hallucinations they exhibit. Moving beyond generic evaluations, the field has progressed towards fine-grained, context-aware, and domain-specific frameworks crucial for systematically identifying, categorizing, and quantifying these errors.

Initial efforts to evaluate hallucinations predominantly focused on object-level inconsistencies in static images. For instance, while not a new benchmark in itself, the POPE (Polling-based Object Probing Evaluation) framework became a common reference for assessing object existence. However, such methods often suffered from limitations like response bias or an inability to capture the nuances of free-form generation. To address this, **?** introduced **THRONE**, an object-based benchmark specifically designed for "Type I" hallucinations in the free-form generations of Large Vision-Language Models (LVLMs). THRONE leverages the semantic understanding of open-source LMs for Abstractive Question Answering (AQA) to accurately judge object existence within complex, unconstrained text, significantly reducing judgment errors compared to prior methods. Extending the scope to multiple objects, **?** proposed **ROPE** (Recognition-based Object Probing Evaluation), an automated protocol for assessing multi-object hallucination. ROPE uniquely employs visual referring prompts (e.g., bounding boxes) to eliminate referential ambiguity and systematically investigates hallucination across various object class distributions within an image, revealing how LVLMs exploit shortcuts and spurious correlations.

As MLLMs expanded to other modalities, so did the need for tailored benchmarks. For audio-language models, **?** investigated object hallucination, introducing discriminative and generative evaluation tasks along with novel metrics like **ECHO** (Evaluation of Caption Hallucination in audiO) and Cover to quantify hallucination and coverage in audio captioning. This work highlighted a critical weakness in LALMs regarding their understanding of discriminative queries.

Beyond simple object presence, researchers recognized the importance of evaluating more complex relationships between objects. **?** introduced **R-Bench**, a novel benchmark specifically designed to evaluate and analyze relationship hallucinations in LVLMs. A key innovation of R-Bench is its meticulous construction using the `nocaps` validation set, preventing data leakage prevalent in earlier benchmarks that relied on pre-training datasets. Building on this, **?** presented **Reefknot**, a comprehensive benchmark that systematically defines and categorizes relation hallucinations into *perceptive* (concrete) and *cognitive* (abstract) types. Reefknot utilizes real-world semantic triplets and offers

diverse evaluation tasks (Yes/No, Multiple Choice, VQA), providing deeper insights into MLLMs' relational understanding, including the counter-intuitive finding that perceptive hallucinations are often more prevalent.

The dynamic nature of video content introduced new challenges, leading to benchmarks for temporal and event-based hallucinations. **?** developed **VideoHallucer**, the first comprehensive benchmark for hallucination detection in Large Video-Language Models (LVLMs). VideoHallucer categorizes hallucinations into *intrinsic* (contradicting video content) and *extrinsic* (not verifiable from video) and employs an adversarial binary VideoQA method for robust evaluation. Further specializing in temporal errors, **?** introduced **VidHalluc**, the largest benchmark for evaluating temporal hallucinations, including Action Hallucination (ACH), Temporal Sequence Hallucination (TSH), and Scene Transition Hallucination (STH). VidHalluc uses a semi-automated pipeline to identify adversarial video pairs that are visually different but semantically similar, exposing MLLMs' over-reliance on contextual scenes.

The complexity of multimodal interactions also spurred the creation of benchmarks for cross-modal driven hallucinations. **?** presented **AVHBench**, the first cross-modal hallucination benchmark for Audio-Visual Large Language Models (AV-LLMs). AVHBench uniquely targets hallucinations arising from the *interaction* between audio and visual modalities, such as perceiving imaginary sounds from visual cues or fake visual events from audio cues, using a semi-automatic annotation pipeline and synthetic videos. **?** introduced **Hallusionbench**, an advanced diagnostic suite for entangled language hallucination and visual illusion in LVLMs. This benchmark features a novel VQA structure with control groups and human-edited images to quantitatively analyze models' response tendencies and specific failure modes, revealing a pronounced language bias in even state-of-the-art models. Pushing the boundaries further, **?** proposed **CCHall**, a novel benchmark for detecting joint cross-lingual and cross-modal hallucinations, a scenario highly relevant for global applications but largely unaddressed, demonstrating significantly worse performance in MLLMs when both complexities are combined.

To provide a more unified and fine-grained evaluation, **?** developed **Hal-Eval**, a

universal and fine-grained hallucination evaluation framework. Hal-Eval introduces a novel category of "Event Hallucination," which involves inventing fictional entities and weaving entire narratives around them, and integrates both discriminative and generative evaluation methods through an automatic annotation pipeline. Similarly, **?** presented **UNIHD** (Unified Multimodal Hallucination Detection) and its accompanying benchmark **MHaluBench**. UNIHD is a task-agnostic, tool-enhanced framework that unifies hallucination detection across image-to-text and text-to-image generation, covering modality-conflicting (object, attribute, scene-text) and fact-conflicting hallucinations with fine-grained, claim-level annotations.

Beyond general multimodal contexts, the critical implications of hallucinations in specific domains have led to specialized benchmarks. In the medical domain, **?** introduced **Med-HallMark**, the first benchmark dedicated to detecting and evaluating medical hallucinations in LVLMs. Med-HallMark features multi-task support (Med-VQA, Imaging Report Generation), a novel hierarchical hallucination categorization based on clinical severity (e.g., Catastrophic, Critical), and a new evaluation metric, MediHall Score, to provide a nuanced assessment of clinical impact.

Finally, a crucial meta-perspective on benchmark quality was provided by **?**, who introduced the **Hallucination benchmark Quality Measurement (HQM)** framework. HQM systematically assesses benchmark quality based on reliability (test-retest, parallel-forms) and validity (criterion validity, coverage of hallucination types), revealing significant issues like response bias and misalignment with human judgment in many existing benchmarks. They also proposed **HQH** (High-Quality Hallucination Benchmark), which demonstrates superior reliability and validity through a simplified binary hallucination detection metric and comprehensive coverage of eight hallucination types. This work underscores the importance of robust benchmark design for trustworthy progress. Furthermore, **?** investigated "Multimodal Hallucination Snowballing," proposing the **MMHal-Snowball** framework to evaluate how previously generated hallucinations can mislead an LVLM's subsequent responses in conversational settings, highlighting the dynamic and cumulative nature of these errors.

In conclusion, the evolution of hallucination evaluation benchmarks for MLLMs reflects a growing understanding of the complexity and diversity of these errors. From basic object presence to intricate temporal, relational, cross-modal, and domain-specific inconsistencies, researchers are continuously developing more sophisticated tools. However, challenges persist in creating benchmarks that are truly comprehensive, scalable, free from inherent biases (e.g., from LLM-assisted generation or evaluation), and perfectly aligned with human judgment across all contexts. Future directions will likely involve more dynamic, interactive, and adaptive evaluation frameworks that can capture the evolving nature of MLLM capabilities and their failure modes, potentially integrating advanced causal inference techniques to pinpoint the root causes of hallucination more precisely.

6.3 Multimodal Mitigation Strategies

Hallucinations in Large Vision-Language Models (LVLMs) and other multimodal architectures present a formidable challenge, manifesting as generated content factually inconsistent with visual inputs, misinterpretations of audio cues, or flawed cross-modal reasoning **??**. These errors, stemming from issues like the "modality gap," dataset biases, and inherited language model tendencies, severely undermine the dependability and credibility of multimodal AI systems **?**. Given the theoretical inevitability of hallucination in computable models **?**, mitigation strategies are crucial for managing, rather than eliminating, these inconsistencies. This subsection explores a range of techniques, from robust training paradigms to sophisticated inference-time interventions, all designed to enhance factual consistency by accounting for the complex interplay between modalities.

Training-based strategies aim to instill hallucination resistance directly into the model's learning process. Pioneering efforts in robust instruction tuning for LVLMs include **?**'s LRV-Instruction, a dataset incorporating diverse *negative instructions* (e.g., describing nonexistent objects) to explicitly train models to avoid visual hallucinations. This data-centric approach, coupled with the GAVIE evaluation framework, demonstrated the effectiveness of explicitly teaching LMMs what *not* to generate. Building on the success of preference optimization, **?** introduced Hallucination-targeted Direct Preference Op-

timization (HDPO) for MLLMs. HDPO constructs specific preference data to address three distinct causes of multimodal hallucinations: visual distracted hallucination (VDH), long-context hallucination (LCH), and multimodal conflict hallucination (MCH). Unlike general DPO methods, HDPO's targeted data generation, which includes amplifying irrelevant visual information or introducing conflicting text, allows for more precise and consistent improvements across diverse hallucination types. While effective, these instruction tuning and DPO methods demand substantial effort in curating high-quality, hallucination-specific preference data, which can be computationally expensive and may risk stifling generative creativity if not carefully balanced.

Beyond data augmentation, other training-based methods focus on internal model dynamics. **?** introduced NoiseBoost, a generalizable technique that injects noise into visual tokens during supervised fine-tuning (SFT), reinforcement learning (RL), or semi-supervised learning (SSL). This perturbation encourages MLLMs to distribute attention more evenly between visual and linguistic tokens, reducing over-reliance on language priors without incurring significant inference costs. This approach is particularly appealing for its efficiency, as it avoids the need for extensive negative data curation. Similarly, **?** presented an efficient "judge-free" self-improvement framework for MLLMs. This method leverages controllable negative samples and lightweight CLIP-based verification for Direct Preference Optimization (DPO), bypassing the computational expense and potential biases associated with using large models as judges. The challenge with such internal perturbation methods lies in finding the optimal noise level or intervention point to improve robustness without degrading overall performance or introducing new biases.

Inference-time interventions offer "plug-and-play" solutions that do not require expensive retraining, making them highly adaptable. One category focuses on external grounding and post-remedy correction. **?**'s Woodpecker is a training-free pipeline that leverages external expert models (e.g., object detectors, VQA models) to validate visual facts and correct MLLM responses. This approach, akin to active retrieval augmentation for multimodal contexts, enhances interpretability by explicitly adding visual grounding (e.g., bounding boxes) to corrected text. However, its effectiveness is inherently limited

by the accuracy and coverage of the external expert models, potentially leading to cascading errors if the "experts" themselves are flawed, and can introduce additional inference latency.

Another significant area of inference-time intervention involves directly manipulating the model's internal states or attention mechanisms to enhance visual grounding. **?**'s MemVR (Memory-space Visual Retracing) addresses "visual amnesia" by re-injecting visual tokens as supplementary evidence into intermediate layers of the MLLM, dynamically activated by high uncertainty. This "look-twice" mechanism efficiently refreshes visual memory, overcoming the forgetting of visual information that can lead to hallucinations. Similarly, **?**'s ClearSight proposes Visual Amplification Fusion (VAF), a plug-and-play technique that enhances attention to visual signals within MLLM middle layers during the forward pass, mitigating object hallucination without compromising content quality or inference speed. **?** proposed ICT (Image-Object Cross-Level Trusted Intervention), a training-free method that calculates an "intervention direction" to shift the model's focus towards different levels of visual information (overall image and fine-grained objects) by targeting specific attention heads. These methods offer fine-grained control but require deep architectural understanding and careful tuning to avoid indiscriminately suppressing beneficial language priors.

Contrastive decoding methods, which leverage subtle differences in input or model states to steer generation, have also been adapted for multimodal contexts. **?** introduced Visual Contrastive Decoding (VCD), a training-free approach that mitigates object hallucinations by contrasting output distributions from original and *distorted* visual inputs. By applying a Gaussian noise mask to create visual uncertainty, VCD penalizes tokens that are highly probable under the distorted input (which amplifies reliance on language priors), thereby calibrating the model's output. Building upon this, **?** proposed Hallucination-Induced Optimization (HIO), a more precise contrastive decoding strategy for LVLMs. HIO addresses the "uncontrollable nature of global visual uncertainty" in prior methods by training an "Evil LVLM" using a *reversed* Bradley-Terry model to specifically *amplify* targeted hallucinations. This "Evil LVLM" then provides a stronger, more focused

contrastive signal during decoding, leading to superior hallucination reduction. Another novel approach, ?'s ConVis, employs a Text-to-Image (T2I) model to visualize potential hallucinations from an MLLM's initial caption. These visualized discrepancies then serve as a contrastive signal during decoding to penalize hallucinated tokens, offering a unique way to generate visual contrastive signals. While powerful, contrastive decoding methods can introduce additional computational overhead during inference and require careful balancing to prevent over-penalization that might stifle creativity or lead to overly cautious, less fluent outputs.

Beyond direct visual grounding, some methods leverage advanced reasoning or generative capabilities. ?'s Counterfactual Inception prompts LMMs to self-generate "counterfactual keywords" (e.g., non-existent objects) and then explicitly instructs the model to avoid them during response generation. This training-free method, guided by a Plausibility Verification Process (PVP) using CLIP scores, implants counterfactual thinking to reduce hallucinations by making the model explicitly aware of what *not* to say. Furthermore, ? offered CAUSAL MM, a causal inference framework that uses back-door adjustment and counterfactual reasoning to mitigate modality prior-induced hallucinations by deciphering attention causality. This provides a more principled and theoretically grounded way to balance visual and language influences, moving beyond heuristic adjustments.

Unified frameworks and multi-view reasoning approaches aim for more adaptive and comprehensive mitigation. ? introduced "Dentist," a unified hallucination mitigation framework that first classifies the query type (e.g., perception vs. reasoning) and then applies a tailored mitigation strategy within an iterative validation loop. For perception queries, it uses visual verification with sub-questions, while for reasoning queries, it employs Chain-of-Thought (CoT) prompting. This adaptive approach, which refines answers until semantic convergence, directly addresses the need for mitigation strategies that adapt based on query type. Complementing this, ?'s MVP (Multi-View Multi-Path Reasoning) is a training-free framework that enhances LVLM inference by thoroughly perceiving image information from "multi-view" dimensions (top-down, regular, bottom-

up) and then employing "multi-path certainty-driven reasoning" during decoding. This approach explores multiple reasoning paths and aggregates certainty scores to select the most reliable answer, maximizing the innate capabilities of existing LVLMs. While promising for their adaptability, the complexity of these unified frameworks and the accuracy of query classification or certainty aggregation remain critical factors influencing their real-world performance.

In conclusion, the landscape of multimodal hallucination mitigation is characterized by a diverse array of strategies, each with distinct trade-offs. Training-based methods like LRV-Instruction ? and HDPO ? aim for intrinsic robustness but demand significant data curation and computational resources. Inference-time interventions, including external grounding (Woodpecker ?), internal state manipulation (MemVR ?, ClearSight ?), and contrastive decoding (VCD ?, HIO ?, ConVis ?), offer flexible, often training-free, alternatives, but their effectiveness can depend on external tool quality, the intrusiveness of internal manipulations, or the computational overhead of generating contrastive signals. Unified frameworks like Dentist ? and multi-view reasoning approaches like MVP ? represent a move towards more adaptive and comprehensive solutions. Despite these advancements, the inherent complexity of multimodal understanding, coupled with the theoretical limits of hallucination, suggests that future research must continue to balance factual consistency with generative fluency, explore more sophisticated causal interventions, and develop methods that are robust to diverse and evolving hallucination types across various multimodal contexts, including audio and video, to ensure truly dependable and credible AI systems.

7 Advanced Topics and Future Directions

7.1 Adversarial Hallucination and Robustness

The increasing deployment of Large Language Models (LLMs) and multimodal models (MLLMs) in sensitive applications necessitates a critical focus on their resilience against deliberate manipulation. This subsection delves into the emerging and crucial field of

adversarial attacks specifically designed to induce hallucinations, examining how exploiting inherent vulnerabilities reveals fundamental weaknesses and drives the development of proactive robustness engineering. Understanding these techniques is paramount, not merely for creating new attacks, but for building robust AI systems capable of withstanding sophisticated, targeted manipulations.

7.1.1 Adversarial Attack Methods and Exploitable Vulnerabilities

Adversarial hallucination attacks aim to intentionally mislead LLMs and MLLMs into generating factually incorrect or unfaithful content. These attacks often exploit specific architectural weaknesses or biases learned during training. A prominent example in MLLMs is the "attention sink" phenomenon, identified by **?**. Their work proposes a novel attack that manipulates attention scores and hidden embeddings to trigger these sinks, leading to the generation of visually uninterpretable or misleading content. This attack significantly increases hallucinated sentences and words, revealing how instruction-tuning can inadvertently create "two-segment response" patterns with declining image-text relevance, exposing a fundamental weakness in model faithfulness.

Beyond architectural vulnerabilities, linguistic and data-driven biases can also be exploited. **?** reveal a "semantic shift bias" in LVLMs, where the presence of paragraph breaks () in training data leads models to associate subsequent content with a semantic shift, increasing the likelihood of hallucination. Crucially, they demonstrate that strategically inserting can *induce* multimodal hallucinations, providing a novel and simple attack mechanism. This highlights how subtle structural elements can be leveraged to corrupt model output. Similarly, **?** demonstrate the high susceptibility of LLMs to adversarial hallucination attacks in clinical decision support. By embedding fabricated details (e.g., false lab results) into clinical prompts, they show that LLMs frequently elaborate on this false information, posing significant risks in high-stakes domains. This type of attack directly targets the LLM's knowledge retrieval and generation capabilities, forcing it to "hallucinate" details consistent with the fabricated premise.

Another form of adversarial manipulation, closely related to inducing misinformation,

is explored by **?**. They identify that LLMs are "involuntary truth-tellers" and struggle to generate genuinely fallacious or deceptive reasoning. Exploiting this "fallacy failure," they propose a jailbreak attack where an LLM is prompted to generate a fallacious yet seemingly real procedure for harmful behavior. The model, unable to fabricate a truly fallacious solution, instead proposes a truthful (and thus harmful) one, bypassing safety mechanisms. While framed as a jailbreak, this method effectively induces the model to generate factually harmful content under a deceptive premise, akin to an adversarial hallucination of a "safe" procedure.

It is important to distinguish these direct adversarial attacks from diagnostic benchmarks that use adversarial-style inputs to probe model weaknesses. For instance, **?** introduces AVHBench, which uses synthetic videos with swapped audio to create natural mismatches, intentionally inducing cross-modal hallucinations to evaluate an Audio-Visual LLM's discernment. Similarly, **?** utilizes human-edited images within their HALLUSION BENCH to challenge LVLMs' robustness against "language hallucination" and "visual illusion." While these tools are invaluable for revealing how models misinterpret subtle relationships or over-rely on one modality, they are primarily diagnostic rather than malicious attack vectors. However, understanding the vulnerabilities they expose, such as the "Multimodal Hallucination Snowballing" phenomenon identified by **?** where an LVLM's own previously generated hallucinations influence subsequent responses, is critical for anticipating potential adversarial exploitation.

7.1.2 Proactive Robustness Engineering: Defenses Against Adversarial Hallucination

Understanding these vulnerabilities and attack vectors is crucial for developing more robust AI systems, shifting the paradigm from purely reactive mitigation to proactive robustness engineering. This involves building strong defenses against sophisticated, targeted manipulations, ensuring models can withstand deliberate attempts to make them hallucinate. Proactive robustness techniques can be broadly categorized into training-based and inference-time interventions, often designed to strengthen grounding, balance

modality priors, or enhance critical self-correction.

In the realm of **training-based robustness engineering**, methods aim to instill hallucination resistance directly into the model's learned parameters, making them inherently more resilient to adversarial prompts. **?** proposes NoiseBoost, a generalizable technique that injects Gaussian noise into projected visual tokens during supervised fine-tuning (SFT) or reinforcement learning (RL). This perturbation compels the MLLM to distribute attention more evenly between visual and linguistic tokens, thereby reducing over-reliance on language priors that adversaries might exploit to induce visual hallucinations. Addressing fine-grained hallucinations, **?** introduces ReCaption, a framework that fine-tunes Large Vision-Language Models (LVLMs) using diverse rewritten captions generated by ChatGPT. This approach enhances fine-grained visual-text alignment, making models more robust to subtle attribute and behavior inaccuracies that an adversary might attempt to corrupt. Further advancing training-based defenses, **?** presents an efficient "judge-free" self-improvement framework for MLLMs. This method generates controllable negative samples by blending conditional and unconditional decoding paths with a "hallucination ratio" and uses a lightweight CLIP-based verifier for Direct Preference Optimization (DPO), proactively training models to resist generating hallucinatory content even when subtly prompted towards it. More broadly, **?** highlights the role of AI alignment training in reducing LLMs' propensity for misinformation and improving their ability to refuse malicious instructions, forming a foundational layer of defense against adversarial attempts to induce harmful hallucinations.

Inference-time and post-hoc interventions provide flexible, training-free solutions to enhance robustness in deployed models, often directly countering the vulnerabilities exploited by adversarial attacks. A direct countermeasure to the -induced hallucination attack by **?** is their proposed MiHO (Mitigating Hallucinations during Output) method. By adjusting the decoding strategy to reduce the logits of the token, MiHO effectively prevents its generation, thereby suppressing the semantic shift bias and reducing hallucination without retraining. This exemplifies a targeted defense against a specific adversarial mechanism.

Other inference-time strategies focus on strengthening visual grounding and balancing modality priors, directly countering attacks like the "attention sink." **?** introduces MemVR (Memory-space Visual Retracing), a decoding paradigm that re-injects visual tokens into the MLLM's middle layers, dynamically activating when the model exhibits high uncertainty. This reinforces visual memory, mitigating the "amnesia" of visual information that can be exploited by adversaries to induce hallucinations. Similarly, **?** proposes CAUSAL MM, a causal inference framework that applies back-door adjustment and counterfactual reasoning to attention mechanisms. By treating modality priors as confounding factors, CAUSAL MM balances visual and language attention, making model outputs more aligned with multimodal inputs and less susceptible to prior-induced hallucinations that an adversary might leverage. **?** presents ConVis, a novel contrastive decoding method that leverages a Text-to-Image (T2I) model to visualize potential hallucinations from the MLLM's initial caption. By comparing logit distributions from the original and T2I-reconstructed images, ConVis penalizes the generation of visualized hallucinations, offering a unique visual feedback loop for robustness against visually-grounded attacks. The Residual Visual Decoding (RVD) method by **?** also falls into this category, mitigating hallucination snowballing by integrating residual visual input to revise output distributions during generation, thereby defending against an adversary attempting to exploit this internal vulnerability.

Furthermore, critical reasoning and external validation at inference time are crucial. **?** introduces Counterfactual Inception, a training-free method that prompts LMMs to generate and then explicitly avoid "counterfactual keywords" that deviate from visual content. This encourages a more critical self-correction process, making it harder for adversarial prompts to inject and propagate false information. **?** proposes MVP (Multi-View Multi-Path Reasoning), a framework that maximizes an LVLM's innate capabilities by seeking multi-view information and employing certainty-driven decoding. By exploring multiple reasoning paths and selecting answers with high certainty, MVP significantly alleviates hallucinations without external tools or retraining, offering a robust approach against adversarial attempts to force a single, hallucinated conclusion.

In conclusion, the study of adversarial hallucination marks a crucial shift towards understanding and proactively defending against sophisticated manipulations of AI models. By exposing vulnerabilities like the attention sink, semantic shift biases, and the propensity to elaborate on fabricated details, researchers are developing a diverse arsenal of training-based and inference-time techniques to build more resilient systems. The ongoing challenge lies in creating defenses that are both effective against increasingly sophisticated attacks and efficient enough for real-world deployment, while ensuring that robustness measures do not inadvertently compromise other desirable model characteristics like fluency or creativity. Future work will likely focus on deeper causal understanding of model failures and adaptive, multi-layered defense strategies to ensure the credibility and trustworthiness of AI in complex, real-world scenarios.

7.2 Safety-Critical Applications and Guardrails

In domains where the consequences of AI errors are profound and potentially irreversible, such as medical diagnosis, legal counsel, financial advisories, or critical infrastructure management, the generation of erroneous or hallucinatory content by Large Language Models (LLMs) transitions from a mere inconvenience to a catastrophic risk. These high-stakes environments necessitate a paradigm shift from general factual accuracy to absolute error prevention, aiming to eliminate "never events"—errors that are entirely preventable and unacceptable **?**. For example, in clinical settings, LLMs analyzing unstructured medical notes can exhibit significant hallucination, leading to inaccuracies in extracting critical patient information, which could directly impact diagnosis and treatment **?**. This underscores the profound ethical and practical imperative for robust safeguards. This subsection explores the development and implementation of specialized "guardrails" and application-specific mechanisms designed to ensure the highest levels of dependability, accountability, and credibility in AI systems deployed in these sensitive real-world settings.

Guardrails, in this context, represent explicit enforcement mechanisms designed to constrain LLM behavior and output within predefined safety and operational boundaries. Beyond merely detecting or mitigating hallucinations, they act as a critical layer of de-

fense, preventing the generation of harmful, factually incorrect, or otherwise undesirable information before it reaches end-users **?**. While the term "guardrails" broadly encompasses mechanisms for preventing prompt injection, ensuring topical relevance, and upholding ethical guidelines, our focus here is primarily on **semantic guardrails** designed to prevent factual hallucinations in safety-critical applications. These guardrails specifically focus on the meaning and content of the LLM's output, ensuring strict alignment with domain-specific knowledge, safety protocols, and ethical considerations **?**.

A pioneering example of such a system is presented by **?**, who explicitly address the critical need for guardrails in medical safety-critical settings, specifically pharmacovigilance. Their work highlights how application-specific mechanisms can be engineered to prevent "never event" errors in processing Individual Case Safety Reports (ICSRs). They categorize guardrails into "hard" and "soft" types:

- **Hard Semantic Guardrails**: These enforce strict, non-negotiable rules, actively preventing outputs that violate fundamental safety criteria. For instance, their "MISMATCH Guardrail" identifies and flags discrepancies in drug names or adverse event terms between source and generated text, leveraging custom drug dictionaries and medical ontologies (e.g., MedDRA). This mechanism ensures that critical information is neither hallucinated nor omitted, directly preventing errors that could lead to patient harm or regulatory non-compliance.

- **Soft Semantic Guardrails**: These mechanisms communicate uncertainty or flag outputs for human review when confidence is low or input data is anomalous. **?** introduce Document-wise Uncertainty Quantification (DL-UQ), which uses document embeddings to identify unusual input documents, signaling potential risks in the LLM's processing.

This distinction is crucial: hard guardrails aim for absolute prevention of specific, known error types, while soft guardrails manage residual risk by enabling informed human intervention.

Beyond explicit rule enforcement, guardrails are intrinsically linked to the LLM's ability to manage and communicate its inherent uncertainty. Building on the uncertainty

quantification methods discussed in Section 4.3, the *semantic entropy* proposed by **?** serves as a powerful signal for a guardrail system. This label-free abstention mechanism allows LLMs to self-identify high-risk outputs by quantifying the semantic diversity of potential continuations. A guardrail system can then leverage this signal to trigger an abstention response, preventing potentially harmful hallucinations by signaling when human intervention or further verification is required. This operationalizes uncertainty estimation as an active component of an error prevention strategy, acting as a soft guardrail.

Furthermore, ensuring the verifiability and accountability of AI-generated content is indispensable for safety-critical applications. While not a direct guardrail, the ability to trace information back to its source provides a critical layer of human-in-the-loop verification, often mandated in legal, medical, and financial contexts where accountability is non-negotiable. As discussed in Section 4.1, **?** introduced a reproducible benchmark and automatic evaluation framework for enabling LLMs to generate text with verifiable citations. When integrated into a safety-critical workflow, this capability allows domain experts to audit the factual basis of AI claims, serving as an essential support system for outputs that have passed through guardrail checks. It complements guardrails by providing the necessary transparency for post-hoc analysis and ensuring trust in the system's overall credibility.

Implementing robust guardrails presents several significant challenges that extend beyond generic statements about completeness or overhead. A fundamental challenge stems from the inherent "unavoidable nature" of hallucination in LLMs, as discussed in Section 2.3 **?**. This theoretical limit implies that guardrails cannot completely eradicate all forms of hallucination but must instead manage and contain them. The brittleness of rule-based systems against the fluidity and ambiguity of natural language poses a continuous threat; subtle linguistic variations or adversarial prompt injections can bypass even well-designed guardrails, potentially inducing harmful outputs or hallucinations **??**. Developing comprehensive rule sets for hard guardrails requires deep, often scarce, domain expertise and continuous, labor-intensive iteration, creating a scalability bottleneck. Moreover, formally verifying semantic constraints to guarantee their coverage and prevent

adversarial bypasses remains an open research problem. In multimodal safety-critical applications, such as autonomous driving or medical image analysis, guardrails must also contend with cross-modal hallucinations arising from perturbed inputs or misinterpretations of visual information, adding another layer of complexity **??**. The trade-off between strict error prevention and maintaining the LLM's utility and generative capabilities must be carefully managed, as overly restrictive guardrails can inadvertently limit the system's effectiveness.

Future research must focus on developing more adaptive, dynamic, and self-evolving guardrail systems that can infer rules from documentation or safety cases, reducing reliance on manual expert-driven creation. Advances in formal verification techniques are crucial to provide provable guarantees for guardrail effectiveness. Furthermore, integrating advanced uncertainty quantification and self-reflection mechanisms more deeply into guardrail architectures will enable more nuanced risk management and informed human intervention.

In conclusion, the deployment of AI in safety-critical applications necessitates a dedicated focus on guardrails as explicit enforcement mechanisms for absolute error prevention. The integration of application-specific semantic guardrails, as exemplified by **?**'s work in pharmacovigilance, alongside uncertainty-triggered abstention mechanisms leveraging signals like semantic entropy **?**, represents a significant advancement. These are complemented by supporting infrastructures like verifiable citation generation **?** that enhance accountability. Addressing the inherent limitations of LLMs and the complexities of natural language, the development of robust, verifiable, and ethically sound guardrail architectures is paramount to building truly dependable AI systems capable of responsible deployment in the most sensitive real-world scenarios.

7.3 Meta-Evaluation and Unified Theoretical Frameworks

As the field of AI hallucination research matures, moving beyond initial problem identification and ad-hoc solutions, a critical dual focus has emerged: the rigorous meta-evaluation of existing benchmarks and the ongoing pursuit of a unified theoretical framework. This

forward-looking perspective, building upon the diverse detection methodologies discussed in Section 3 and the mechanistic insights from Section 2.2, aims to ensure the scientific rigor of research, guide the development of future dependable AI systems, and establish a robust foundation for advancements in AI reliability.

The proliferation of diverse and sophisticated evaluation methodologies, while essential for progress, simultaneously necessitates a critical assessment of these tools themselves. True meta-evaluation involves scrutinizing the quality, reliability, and validity of hallucination benchmarks. For instance, while benchmarks like POPE ? were foundational for object hallucination detection in Large Vision-Language Models (LVLMs), subsequent meta-analysis by ? revealed significant prompt bias, demonstrating that these benchmarks could inadvertently exploit model judgment biases rather than accurately measuring real-world hallucination. This highlights a critical flaw in relying solely on primary evaluation metrics without a deeper understanding of their underlying mechanisms and potential vulnerabilities. Similarly, ? provides a review of faithfulness metrics across various tasks, implicitly performing a meta-evaluation by correlating their effectiveness with human judgment, underscoring that LLM-based evaluators often achieve the highest correlation. This kind of comparative analysis is crucial for identifying robust and trustworthy evaluation paradigms. Furthermore, comprehensive surveys like those by ? and ? contribute to meta-evaluation by systematically categorizing existing benchmarks and discussing their limitations, scope, and applicability. They reveal the fragmentation of evaluation efforts and the need for standardized, cross-comparable metrics that are less susceptible to dataset biases or prompt engineering. The challenge lies in developing meta-benchmarks or frameworks that can consistently compare the efficacy, scope, and validity of diverse hallucination detection methods, ensuring that the tools used to measure hallucination are themselves robust and reliable. Without such a meta-evaluative layer, the field risks optimizing for flawed metrics, leading to an illusion of progress rather than genuine advancement in mitigating hallucination.

Complementing the advancements and critiques in evaluation, the field is increasingly gravitating towards a unified theoretical understanding of hallucination, integrating em-

pirical observations with fundamental mathematical insights. This conceptual shift is a direct response to the limitations of purely empirical taxonomies and mechanistic analyses (as seen in Section 2.2), seeking to uncover the underlying principles governing hallucination. A pivotal contribution in this direction is presented by **?**, who propose a unified theoretical framework for LLM hallucination, offering a formal mathematical definition and exploring its origins, such as undecidability principles, alongside empirical causes related to data and architecture. While this framework offers a powerful conceptual lens, its practical utility in guiding the development of specific, universally applicable mitigation techniques remains an underexplored but critical next step. Further deepening this theoretical grounding, **?** rigorously investigate the (im)possibility of automated hallucination detection, proving that while detection is fundamentally impossible for most language collections when relying solely on correct examples, it becomes feasible with expert-labeled feedback (both positive and negative examples). This theoretical insight provides crucial guidance for the design of effective hallucination detectors, validating the importance of methods like Reinforcement Learning with Human Feedback (RLHF) and highlighting the inherent limitations of unsupervised detection.

Mechanistic insights also continue to integrate with theoretical frameworks. For instance, **?** identified "knowledge overshadowing" as a novel root cause of amalgamated hallucinations, where dominant conditions in training data lead to over-generalization. Their work provides a specific mechanistic insight into how factual inaccuracies can arise even with correct data, supported by a derived generalization bound, thereby integrating empirical observation with a mathematical explanation of its limits. Similarly, **?** delves into the inference dynamics of LLMs, revealing how output token probabilities evolve across layers, distinguishing between correct and hallucinated cases. This fine-grained analysis of internal model states offers a mechanistic basis for understanding *why* hallucinations occur even when models possess the correct knowledge, providing a pathway for more principled detection and mitigation. Furthermore, **?** proposes a novel mathematical formulation for Information Quality (IQ) in LLMs, defining it as a weighted function of consistency, relevance, and accuracy. This framework, while conceptual, represents

an important step towards formalizing the very qualities that hallucination undermines, offering a quantifiable target for theoretical and empirical improvements.

In conclusion, the dual focus on meta-evaluation and unified theoretical frameworks signifies a critical maturation in hallucination research. The field must now rigorously evaluate its evaluation tools, moving beyond simply applying benchmarks to critically assessing their biases, scope, and correlation with human judgment, as exemplified by critiques of existing LVLM benchmarks **?** and reviews of faithfulness metrics **?**. Concurrently, the pioneering work on unified theoretical frameworks **?**, the theoretical limits of detection **?**, and the deepening mechanistic insights with mathematical grounding **??** are paving the way for a deeper scientific understanding of AI reliability. This integrated approach, combining robust evaluation of evaluation methodologies with fundamental theoretical insights, is essential for guiding the development of truly dependable, trustworthy, and scientifically grounded AI systems in the future, moving beyond reactive fixes to principled, proactive solutions.

8 Conclusion

8.1 Summary of Key Developments

The intellectual journey in addressing hallucination within Large Language Models (LLMs) and their multimodal successors reflects a profound maturation in the pursuit of trustworthy artificial intelligence. This field has rapidly evolved from initial problem characterization to sophisticated management strategies, fundamentally shifting the paradigm from viewing hallucination as a solvable bug to acknowledging its inherent nature. This section provides a concise recap of these major milestones, highlighting the progression from foundational definitions and detection methods to diverse mitigation strategies, the crucial expansion into multimodal AI, and the emerging focus on advanced topics like adversarial robustness and safety-critical applications, underscoring the field's evolving approach to AI dependability across all modalities.

Early research meticulously focused on defining and characterizing hallucinations, pri-

marily within text-based LLMs. Foundational work established critical distinctions between intrinsic errors (contradicting source input) and extrinsic errors (adding unsupportable, often false, information), particularly in tasks like abstractive summarization ?. This initial understanding paved the way for comprehensive taxonomies that categorized hallucinations by their origin and nature, such as input-conflicting, context-conflicting, and fact-conflicting errors ?. The challenge of fact-conflicting hallucinations, lacking immediate ground truth, underscored the need for more robust evaluation. This phase also saw the development of initial profiling frameworks and efforts to attribute hallucination causes through association analysis, laying the groundwork for understanding the mechanistic origins of these failures ?.

As the understanding of hallucination deepened, detection methodologies became increasingly sophisticated, driven by the need for more granular, efficient, and black-box compatible evaluations. The field transitioned from relying on reference-dependent metrics to pioneering reference-free approaches that leveraged the internal consistency of LLM outputs ?. Further refinements included self-contained detection mechanisms employing metamorphic relations and prompt mutation to generate diverse test cases, enhancing detection accuracy and generalizability ?. The increasing complexity of LLM applications also necessitated specialized benchmarks, moving towards fine-grained, token-level or sentence-level analysis, and the verification of LLM reasoning rationales using structured data ?. The imperative for reliability in high-stakes domains led to the development of domain-specific benchmarks, such as Med-HALT for medical contexts, revealing significant challenges even for state-of-the-art models in complex factual recall ?. A notable advancement in verifiability was the development of benchmarks and metrics for enabling LLMs to generate text with explicit and verifiable citations, allowing for automatic evaluation of factual correctness and citation quality ?.

Mitigation efforts evolved along two complementary axes: external grounding and internal model interventions. External grounding techniques, notably Retrieval-Augmented Generation (RAG), became a cornerstone, enabling LLMs to access and integrate external, verifiable knowledge to reduce factual errors. The seminal ReAct paradigm further

advanced this by demonstrating the power of interleaving an LLM's internal reasoning with external actions (e.g., API calls or tool use) to dynamically gather information and self-correct responses in interactive environments ?. This approach was further refined by frameworks that dynamically interwove Chain-of-Thought reasoning with iterative knowledge retrieval ?. While highly effective, RAG's limitations, particularly in noise robustness and handling counterfactuals, spurred diagnostic benchmarks and the development of adaptive RAG frameworks that trigger retrieval only when hallucination is detected in real-time ??. The integration of Knowledge Graphs (KGs) also proved instrumental, allowing LLMs to reason over structured factual repositories and dynamically adapt heterogeneous knowledge sources for enhanced factual consistency ??.

Complementing external grounding, internal strategies focused on steering model behavior during decoding or embedding hallucination resistance during training. Decoding-time techniques included contrastive decoding, which penalizes tokens inconsistent with a reference or subtly altered input, and innovative approaches that leverage "induced hallucinations" to guide models toward factuality ?. Self-correction mechanisms, such as Chain-of-Verification (CoVe), empowered LLMs to internally plan and execute verification steps, iteratively refining responses based on consistency checks ?. Training-based approaches included robust instruction tuning with negative examples and preference optimization techniques like Direct Preference Optimization (DPO), which align model behavior with desired factual accuracy using AI-generated or human-annotated preference data ??. More advanced interventions delved into manipulating internal model states, such as constraining specific attention heads responsible for "false premise hallucinations," demonstrating significant performance gains by targeting mechanistic origins ?.

A pivotal intellectual shift in the field was the theoretical acknowledgment of hallucination's inherent nature. Groundbreaking theoretical proofs, notably employing diagonalization arguments, demonstrated that hallucination is an inevitable and unavoidable characteristic for *any computable LLM*, regardless of architectural advancements or training improvements ?. This fundamental insight, reinforced by theoretical frameworks es-

tablishing the inherent impossibility of automated hallucination detection without expert-labeled negative examples ?, fundamentally reshaped the research paradigm. The focus moved decisively from the ambition of complete elimination to the pragmatic necessity of robust detection, effective mitigation, and responsible management, marking a significant maturation of the field's approach to AI dependability.

This theoretical grounding spurred the crucial expansion of research into multimodal AI, where hallucination presents unique and complex challenges due to the inherent "modality gap" and cross-modal inconsistencies ??. Early work probed object hallucination in Vision-Language Pre-training (VLP) models ?, leading to systematic empirical studies and specialized evaluation methods for various multimodal hallucination types, including object, relationship, temporal, and cross-modal driven errors ?????. The increasing number and complexity of these benchmarks necessitated meta-evaluation to ensure their reliability and validity ?. Multimodal mitigation strategies also diversified, employing both training-based and inference-time approaches, such as robust instruction tuning with negative examples ?, visual grounding techniques ?, and advanced causal inference frameworks that mitigate modality prior-induced hallucinations by deciphering attention causality ?. Recent work has also advanced adversarial generation of diverse visual hallucination instances to build more robust multimodal models ?.

The field's maturation is further evidenced by the emergence of advanced topics that push beyond current paradigms. Research into adversarial hallucination actively probes model vulnerabilities to build more robust systems, revealing that LLMs are highly susceptible to fabricated details in prompts, especially in high-stakes domains like clinical decision support ?. This necessitates a shift towards proactive robustness engineering. Concurrently, the development of specialized guardrails for safety-critical applications has become paramount, focusing on preventing "never event" errors and explicitly communicating uncertainty in domains like pharmacovigilance ?. These developments, alongside the ongoing pursuit of unified theoretical frameworks and meta-evaluation of benchmarks, signify a comprehensive intellectual journey towards fundamentally dependable and trustworthy AI systems.

8.2 Remaining Challenges and Open Questions

Despite significant advancements in understanding, detecting, and mitigating hallucinations in Large Language Models (LLMs) and Multimodal Large Language Models (MLLMs), several persistent challenges and open questions continue to shape the research agenda. The theoretical proof that hallucination is an inherent and inevitable limitation for all computable LLMs, regardless of their architecture or training **?**, fundamentally shifts the paradigm from eradication to robust management, underscoring the enduring nature of this problem.

One primary challenge lies in the **scalability and fine-grained nature of hallucination evaluation**. Early efforts relied on costly human evaluation **?**, which is not scalable for the rapid development of LLMs. While methods like *SelfCheckGPT* **?** introduced zero-resource black-box detection, they still face computational costs and limitations in sampling quality. The need for fine-grained, token-level, and reference-free detection led to benchmarks like HADES **?**, but these are often derived from specific text sources and may not capture the full spectrum of LLM errors. In the multimodal domain, initial object hallucination evaluations like POPE **?** focused on coarse-grained presence/absence, leading to subsequent work addressing more complex relationship hallucinations with R-Bench **?** and cross-modal interactions with AVHBench **?**. However, the quality of these benchmarks themselves has been questioned, with frameworks like HQM/HQH **?** revealing issues like response bias and misalignment with human judgment. The sheer diversity of hallucination types, including multi-object **?**, temporal **?**, and event hallucinations **?**, necessitates increasingly sophisticated and universal evaluation frameworks. Furthermore, the challenge of evaluating hallucinations in unconstrained, free-form generation **?** and dialogue-level contexts **?** remains formidable, even for unified detection frameworks like UNIHD **?**. Automated dataset generation tools like AutoHall **?** offer scalability but often focus on specific hallucination types and rely on other LLMs for classification, potentially inheriting their biases.

Another critical area is the **generalizability of mitigation techniques across diverse tasks and domains**, coupled with the need for a **deeper understanding**

of complex causal mechanisms. Initial retrieval-augmented generation (RAG) approaches like ReAct ? and IRCoT ? improved grounding by interleaving reasoning with external tool use or retrieval. However, these methods often faced limitations in handling complex, multi-step scenarios, were sensitive to prompt engineering, or struggled with the quality and quantity of retrieved information, as highlighted by the ALCE benchmark for citation generation ? and the RGB benchmark's analysis of RAG's shortcomings in noise robustness and information integration ?. Self-correction strategies, as surveyed by ??, such as Chain-of-Verification (CoVe) ?, empower LLMs to self-critique but can be computationally expensive or rely solely on the model's internal knowledge. Integrating Knowledge Graphs (KGs) offers structured grounding ????, but faces challenges in KG quality, maintenance, and the effective integration of graphical structures into LLM reasoning. Dynamic RAG methods like DRAD ? aim for efficiency by triggering retrieval only when hallucinations are detected, yet their effectiveness relies on the accuracy of real-time uncertainty correlation. In multimodal models, mitigation becomes even more complex. Training-based methods like robust instruction tuning with negative examples ? or fine-tuning with caption rewrites ? require extensive data. Inference-time methods like Woodpecker ? and Dentist ? leverage external expert models or LLM-as-judge paradigms, introducing dependencies. The phenomenon of "multimodal hallucination snowballing" ?, where initial errors propagate, further complicates mitigation. Efforts to understand the causal mechanisms, such as CAUSAL MM ? which intervenes on attention to balance modality priors, are emerging but often rely on assumptions about confounding factors. The challenge of balancing visual and linguistic priors without compromising content quality or inference speed remains, as seen in methods like ClearSight/VAF ?, MemVR ?, and ICT ?.

Finally, the **ethical implications of deploying systems with inherent hallucinatory tendencies** are paramount. The inevitability of hallucination ? means that LLMs, even with the best mitigation, will occasionally produce untruthful content. This necessitates a focus on **truly transparent and accountable LLMs**. The ability to generate text with verifiable citations, as benchmarked by ALCE ?, is a step towards ac-

countability, as is the generation of "mind maps" for reasoning transparency ?. However, the economic and societal impact of unreliable LLM outputs ? underscores the urgency of preventing and detecting misinformation ?. Furthermore, the vulnerability to adversarial attacks that exploit internal mechanisms like attention sinks to induce hallucinations ? highlights the need for robust security. Future directions must emphasize the ongoing pursuit of **trustworthy AI and responsible innovation**. This includes developing more adaptive and context-aware AI that can dynamically assess its own uncertainty and seek external validation, as well as creating systems that are not only accurate but also explainable, allowing users to understand and verify their outputs. The goal is to build LLMs that are not just powerful, but also reliable, transparent, and ethically sound for deployment in high-stakes domains like medicine ??.

References

References

Tu Vu, Mohit Iyyer, Xuezhi Wang, et al. (2023). *FreshLLMs: Refreshing Large Language Models with Search Engine Augmentation*. Annual Meeting of the Association for Computational Linguistics.

Yue Chang, Liqiang Jing, Xiaopeng Zhang, et al. (2024). *A Unified Hallucination Mitigation Framework for Large Vision-Language Models*. Trans. Mach. Learn. Res..

Jiaqi Wang, Yifei Gao, and Jitao Sang (2024). *VaLiD: Mitigating the Hallucination of Large Vision Language Models by Visual Layer Fusion Contrastive Decoding*. arXiv.org.

Mengjia Niu, Hao Li, Jie Shi, et al. (2024). *Mitigating Hallucinations in Large Language Models via Self-Refinement-Enhanced Knowledge Retrieval*. arXiv.org.

Aiwei Liu, Qiang Sheng, and Xuming Hu (2024). *Preventing and Detecting Misinformation Generated by Large Language Models*. Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.

Xingxuan Li, Ruochen Zhao, Yew Ken Chia, et al. (2023). *Chain-of-Knowledge: Grounding Large Language Models via Dynamic Knowledge Adapting over Heterogeneous Sources*. International Conference on Learning Representations.

Mengfei Liang, Archish Arun, Zekun Wu, et al. (2024). *THaMES: An End-to-End Tool for Hallucination Mitigation and Evaluation in Large Language Models*. arXiv.org.

Yue Zhang, Yafu Li, Leyang Cui, et al. (2023). *Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models*. Computational Linguistics.

Guanyu Zhou, Yibo Yan, Xin Zou, et al. (2024). *Mitigating Modality Prior-Induced Hallucinations in Multimodal Large Language Models via Deciphering Attention Causality*. International Conference on Learning Representations.

Zechen Bai, Pichao Wang, Tianjun Xiao, et al. (2024). *Hallucination of Multimodal Large Language Models: A Survey.* arXiv.org.

Shukang Yin, Chaoyou Fu, Sirui Zhao, et al. (2023). *Woodpecker: Hallucination Correction for Multimodal Large Language Models.* Science China Information Sciences.

Zouying Cao, Yifei Yang, and Hai Zhao (2023). *AutoHall: Automated Hallucination Dataset Generation for Large Language Models.* arXiv.org.

Ming-Kuan Wu, Jiayi Ji, Oucheng Huang, et al. (2024). *Evaluating and Analyzing Relationship Hallucinations in Large Vision-Language Models.* International Conference on Machine Learning.

Bishwamittra Ghosh, Sarah Hasan, Naheed Anjum Arafat, et al. (2024). *Logical Consistency of Large Language Models in Fact-checking.* International Conference on Learning Representations.

Tianyu Gao, Howard Yen, Jiatong Yu, et al. (2023). *Enabling Large Language Models to Generate Text with Citations.* Conference on Empirical Methods in Natural Language Processing.

Borui Yang, Md Afif Al Mamun, Jie M. Zhang, et al. (2025). *Hallucination Detection in Large Language Models with Metamorphic Relations.* Proc. ACM Softw. Eng..

Yongheng Zhang, Xu Liu, Ruoxi Zhou, et al. (2025). *CCHall: A Novel Benchmark for Joint Cross-Lingual and Cross-Modal Hallucinations Detection in Large Language Models.* Annual Meeting of the Association for Computational Linguistics.

Shangyu Xing, Fei Zhao, Zhen Wu, et al. (2024). *EFUF: Efficient Fine-Grained Unlearning Framework for Mitigating Hallucinations in Multimodal Large Language Models.* Conference on Empirical Methods in Natural Language Processing.

Sicong Leng, Hang Zhang, Guanzheng Chen, et al. (2023). *Mitigating Object Hallucinations in Large Vision-Language Models through Visual Contrastive Decoding.* Computer Vision and Pattern Recognition.

Junho Kim, Yeonju Kim, and Yonghyun Ro (2024). *What if...?: Thinking Counterfactual Keywords Helps to Mitigate Hallucination in Large Multi-modal Models*. Conference on Empirical Methods in Natural Language Processing.

Chaoyu Li, Eun Woo Im, and Pooyan Fazli (2024). *VidHalluc: Evaluating Temporal Hallucinations in Multimodal Large Language Models for Video Understanding*. Computer Vision and Pattern Recognition.

Ziwei Ji, Yuzhe Gu, Wenwei Zhang, et al. (2024). *ANAH: Analytical Annotation of Hallucinations in Large Language Models*. Annual Meeting of the Association for Computational Linguistics.

Yunfan Gao, Yun Xiong, Xinyu Gao, et al. (2023). *Retrieval-Augmented Generation for Large Language Models: A Survey*. arXiv.org.

Ziwei Ji, Zihan Liu, Nayeon Lee, et al. (2022). *RHO (): Reducing Hallucination in Open-domain Dialogues with Knowledge Grounding*. Annual Meeting of the Association for Computational Linguistics.

Griffin Adams, Han-Chin Shing, Q. Sun, et al. (2022). *Learning to Revise References for Faithful Summarization*. Conference on Empirical Methods in Natural Language Processing.

Weihang Su, Yichen Tang, Qingyao Ai, et al. (2024). *Mitigating Entity-Level Hallucination in Large Language Models*. SIGIR-AP.

LI DU, Yequan Wang, Xingrun Xing, et al. (2023). *Quantifying and Attributing the Hallucination of Large Language Models via Association Analysis*. arXiv.org.

Ziwei Ji, Tiezheng Yu, Yan Xu, et al. (2023). *Towards Mitigating Hallucination in Large Language Models via Self-Reflection*. arXiv.org.

Liangming Pan, Michael Stephen Saxon, Wenda Xu, et al. (2023). *Automatically Correcting Large Language Models: Surveying the landscape of diverse self-correction strategies*. arXiv.org.

Haoqiang Kang, and Xiao-Yang Liu (2023). *Deficiency of Large Language Models in Finance: An Empirical Examination of Hallucination*. arXiv.org.

Haoqiang Kang, Juntong Ni, and Huaxiu Yao (2023). *Ever: Mitigating Hallucination in Large Language Models through Real-Time Verification and Rectification*. arXiv.org.

Yue Dong, J. Wieting, and Pat Verga (2022). *Faithful to the Document or to the World? Mitigating Hallucinations via Entity-linked Knowledge in Abstractive Summarization*. Conference on Empirical Methods in Natural Language Processing.

Xiaoye Qu, Mingyang Song, Wei Wei, et al. (2024). *Mitigating Multilingual Hallucination in Large Vision-Language Models*. arXiv.org.

Nick McKenna, Tianyi Li, Liang Cheng, et al. (2023). *Sources of Hallucination by Large Language Models on Inference Tasks*. Conference on Empirical Methods in Natural Language Processing.

Rick Rejeleene, Xiaowei Xu, and John R. Talburt (2024). *Towards Trustable Language Models: Investigating Information Quality of Large Language Models*. arXiv.org.

Xinxin Liu (2024). *A Survey of Hallucination Problems Based on Large Language Models*. Applied and Computational Engineering.

Hanchao Liu, Wenyuan Xue, Yifei Chen, et al. (2024). *A Survey on Hallucination in Large Vision-Language Models*. arXiv.org.

Jiawei Chen, Dingkang Yang, Tong Wu, et al. (2024). *Detecting and Evaluating Medical Hallucinations in Large Vision Language Models*. arXiv.org.

Qing Li, Chenyang Lyu, Jiahui Geng, et al. (2024). *Reference-free Hallucination Detection for Large Vision-Language Models*. Conference on Empirical Methods in Natural Language Processing.

Bei Yan, Jie Zhang, Zheng Yuan, et al. (2024). *Evaluating the Quality of Hallucination Benchmarks for Large Vision-Language Models*. arXiv.org.

Yuxuan Wang, Yueqian Wang, Dongyan Zhao, et al. (2024). *VideoHallucer: Evaluating Intrinsic and Extrinsic Hallucinations in Large Video-Language Models*. arXiv.org.

Yuxi Xie, Guanzhen Li, Xiao Xu, et al. (2024). *V-DPO: Mitigating Hallucination in Large Vision Language Models via Vision-Guided Direct Preference Optimization*. Conference on Empirical Methods in Natural Language Processing.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, et al. (2023). *Hallusionbench: An Advanced Diagnostic Suite for Entangled Language Hallucination and Visual Illusion in Large Vision-Language Models*. Computer Vision and Pattern Recognition.

Xun Liang, Shichao Song, Simin Niu, et al. (2023). *UHGEval: Benchmarking the Hallucination of Chinese Large Language Models via Unconstrained Generation*. Annual Meeting of the Association for Computational Linguistics.

Lei Huang, Weijiang Yu, Weitao Ma, et al. (2023). *A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions*. ACM Trans. Inf. Syst..

Qitan Lv, Jie Wang, Hanzhu Chen, et al. (2024). *Coarse-to-Fine Highlighting: Reducing Knowledge Hallucination in Large Language Models*. International Conference on Machine Learning.

Yuzhe Gu, Ziwei Ji, Wenwei Zhang, et al. (2024). *ANAH-v2: Scaling Analytical Hallucination Annotation of Large Language Models*. Neural Information Processing Systems.

Wen Huang, Hongbin Liu, Minxin Guo, et al. (2024). *Visual Hallucinations of Multimodal Large Language Models*. Annual Meeting of the Association for Computational Linguistics.

Fuxiao Liu, Kevin Lin, Linjie Li, et al. (2023). *Mitigating Hallucination in Large Multi-Modal Models via Robust Instruction Tuning*. International Conference on Learning Representations.

Peng Ding, Jingyu Wu, Jun Kuang, et al. (2024). *Hallu-PI: Evaluating Hallucination in Multi-modal Large Language Models within Perturbed Inputs*. ACM Multimedia.

Vipula Rawte, Swagata Chakraborty, Agnibh Pathak, et al. (2023). *The Troubling Emergence of Hallucination in Large Language Models - An Extensive Definition, Quantification, and Prescriptive Remediations*. Conference on Empirical Methods in Natural Language Processing.

Liangming Pan, Michael Stephen Saxon, Wenda Xu, et al. (2024). *Automatically Correcting Large Language Models: Surveying the Landscape of Diverse Automated Correction Strategies*. Transactions of the Association for Computational Linguistics.

Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, et al. (2023). *HaluEval: A Large-Scale Hallucination Evaluation Benchmark for Large Language Models*. Conference on Empirical Methods in Natural Language Processing.

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, et al. (2023). *Analyzing and Mitigating Object Hallucination in Large Vision-Language Models*. International Conference on Learning Representations.

Zongbo Han, Zechen Bai, Haiyang Mei, et al. (2024). *Skip : A Simple Method to Reduce Hallucination in Large Vision-Language Models*. arXiv.org.

Yining Wang, Mi Zhang, Junjie Sun, et al. (2025). *Mirage in the Eyes: Hallucination Attack on Multi-modal Large Language Models with Only Attention Sink*. arXiv.org.

Xiaoye Qu, Jiashuo Sun, Wei Wei, et al. (2024). *Look, Compare, Decide: Alleviating Hallucination in Large Vision-Language Models via Multi-View Multi-Path Reasoning*. International Conference on Computational Linguistics.

Wenliang Dai, Zihan Liu, Ziwei Ji, et al. (2022). *Plausible May Not Be Faithful: Probing Object Hallucination in Vision-Language Pre-training*. Conference of the European Chapter of the Association for Computational Linguistics.

Nouha Dziri, Andrea Madotto, Osmar Zaiane, et al. (2021). *Neural Path Hunter: Reducing Hallucination in Dialogue Systems via Path Grounding*. Conference on Empirical Methods in Natural Language Processing.

Kim Sung-Bin, Oh Hyun-Bin, JungMok Lee, et al. (2024). *AVHBench: A Cross-Modal Hallucination Benchmark for Audio-Visual Large Language Models*. International Conference on Learning Representations.

Joe B Hakim, Jeffery L. Painter, D. Ramcharran, et al. (2024). *The Need for Guardrails with Large Language Models in Medical Safety-Critical Settings: An Artificial Intelligence Application in the Pharmacovigilance Ecosystem*. arXiv.org.

Chaozhuo Li, Pengbo Wang, Chenxu Wang, et al. (2025). *Loki's Dance of Illusions: A Comprehensive Survey of Hallucination in Large Language Models*. arXiv.org.

Lei Wang, Jiabang He, Shenshen Li, et al. (2023). *Mitigating Fine-Grained Hallucination by Fine-Tuning Large Vision-Language Models with Caption Rewrites*. Conference on Multimedia Modeling.

Kedi Chen, Qin Chen, Jie Zhou, et al. (2024). *DiaHalu: A Dialogue-level Hallucination Evaluation Benchmark for Large Language Models*. Conference on Empirical Methods in Natural Language Processing.

Hanxing Ding, Liang Pang, Zihao Wei, et al. (2024). *Retrieve Only When It Needs: Adaptive Retrieval Augmentation for Hallucination Mitigation in Large Language Models*. arXiv.org.

Shijian Deng, Wentian Zhao, Yu-Jhe Li, et al. (2024). *Efficient Self-Improvement in Multimodal Large Language Models: A Model-Level Judge-Free Approach*. arXiv.org.

Junzhe Chen, Tianshu Zhang, Shiyu Huang, et al. (2024). *ICT: Image-Object Cross-Level Trusted Intervention for Mitigating Object Hallucination in Large Vision-Language Models*. Computer Vision and Pattern Recognition.

Beitao Chen, Xinyu Lyu, Lianli Gao, et al. (2024). *Alleviating Hallucinations in Large Vision-Language Models through Hallucination-Induced Optimization*. Neural Information Processing Systems.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, et al. (2020). *On Faithfulness and Factuality in Abstractive Summarization*. Annual Meeting of the Association for Computational Linguistics.

Xiaoye Qu, Qiyuan Chen, Wei Wei, et al. (2024). *Alleviating Hallucination in Large Vision-Language Models with Active Retrieval Augmentation*. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP).

Jiawei Chen, Hongyu Lin, Xianpei Han, et al. (2023). *Benchmarking Large Language Models in Retrieval-Augmented Generation*. AAAI Conference on Artificial Intelligence.

Yiyang Zhou, Chenhang Cui, Rafael Rafailov, et al. (2024). *Aligning Modalities in Vision Large Language Models via Preference Fine-tuning*. arXiv.org.

Chaoya Jiang, Wei Ye, Mengfan Dong, et al. (2024). *Hal-Eval: A Universal and Fine-grained Hallucination Evaluation Framework for Large Vision Language Models*. ACM Multimedia.

Benedict Aaron Tjandra, Muhammed Razzak, Jannik Kossen, et al. (2024). *Fine-Tuning Large Language Models to Appropriately Abstain with Semantic Entropy*. arXiv.org.

Logesh Kumar Umapathi, Ankit Pal, and Malaikannan Sankarasubbu (2023). *Med-HALT: Medical Domain Hallucination Test for Large Language Models*. Conference on Computational Natural Language Learning.

Xin Zou, Yizhou Wang, Yibo Yan, et al. (2024). *Look Twice Before You Answer: Memory-Space Visual Retracing for Hallucination Mitigation in Multimodal Large Language Models*. arXiv.org.

Ningke Li, Yuekang Li, Yi Liu, et al. (2024). *Drowzee: Metamorphic Testing for Fact-Conflicting Hallucination Detection in Large Language Models*. Proc. ACM Program. Lang..

Yung-Sung Chuang, Linlu Qiu, Cheng-Yu Hsieh, et al. (2024). *Lookback Lens: Detecting and Mitigating Contextual Hallucinations in Large Language Models Using Only Attention Maps*. Conference on Empirical Methods in Natural Language Processing.

Junyi Li, Jie Chen, Ruiyang Ren, et al. (2024). *The Dawn After the Dark: An Empirical Study on Factuality Hallucination in Large Language Models*. Annual Meeting of the Association for Computational Linguistics.

Junyan Wang, Yi Zhou, Guohai Xu, et al. (2023). *Evaluation and Analysis of Hallucination in Large Vision-Language Models*. arXiv.org.

Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli (2024). *Hallucination is Inevitable: An Innate Limitation of Large Language Models*. arXiv.org.

Tianyu Liu, Yizhe Zhang, C. Brockett, et al. (2021). *A Token-level Reference-free Hallucination Detection Benchmark for Free-form Text Generation*. Annual Meeting of the Association for Computational Linguistics.

Potsawee Manakul, Adian Liusie, and M. Gales (2023). *SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models*. Conference on Empirical Methods in Natural Language Processing.

Weihong Zhong, Xiaocheng Feng, Liang Zhao, et al. (2024). *Investigating and Mitigating the Multimodal Hallucination Snowballing in Large Vision-Language Models*. Annual Meeting of the Association for Computational Linguistics.

Yuji Zhang, Sha Li, Jiateng Liu, et al. (2024). *Knowledge Overshadowing Causes Amalgamated Hallucination in Large Language Models*. arXiv.org.

J. Wu, Tsz Ting Chung, Kai Chen, et al. (2024). *Unified Triplet-Level Hallucination Evaluation for Large Vision-Language Models*. Trans. Mach. Learn. Res..

S. Tonmoy, S. M. M. Zaman, Vinija Jain, et al. (2024). *A Comprehensive Survey of Hallucination Mitigation Techniques in Large Language Models*. arXiv.org.

Kai Wu, Boyuan Jiang, Zhengkai Jiang, et al. (2024). *NoiseBoost: Alleviating Hallucination with Noise Perturbation for Multimodal Large Language Models*. arXiv.org.

Yilin Wen, Zifeng Wang, and Jimeng Sun (2023). *MindMap: Knowledge Graph Prompting Sparks Graph of Thoughts in Large Language Models*. Annual Meeting of the Association for Computational Linguistics.

Yifan Li, Yifan Du, Kun Zhou, et al. (2023). *Evaluating Object Hallucination in Large Vision-Language Models*. Conference on Empirical Methods in Natural Language Processing.

Wei Lan, Wenyi Chen, Qingfeng Chen, et al. (2024). *A Survey of Hallucination in Large Visual Language Models*. arXiv.org.

S. Dhuliawala, M. Komeili, Jing Xu, et al. (2023). *Chain-of-Verification Reduces Hallucination in Large Language Models*. Annual Meeting of the Association for Computational Linguistics.

Yuan Sui, and Bryan Hooi (2024). *Can Knowledge Graphs Make Large Language Models More Trustworthy? An Empirical Study over Open-ended Question Answering*. Annual Meeting of the Association for Computational Linguistics.

Wenyi Xiao, Ziwei Huang, Leilei Gan, et al. (2024). *Detecting and Mitigating Hallucination in Large Vision Language Models via Fine-Grained AI Feedback*. arXiv.org.

H. Trivedi, Niranjan Balasubramanian, Tushar Khot, et al. (2022). *Interleaving Retrieval with Chain-of-Thought Reasoning for Knowledge-Intensive Multi-Step Questions*. Annual Meeting of the Association for Computational Linguistics.

Yeji Park, Deokyeong Lee, Junsuk Choe, et al. (2024). *ConVis: Contrastive Decoding with Hallucination Visualization for Mitigating Hallucinations in Multimodal Large Language Models*. AAAI Conference on Artificial Intelligence.

A. Sridhar, and Erik M. Visser (2022). *Improved Beam Search for Hallucination Mitigation in Abstractive Summarization.* arXiv.org.

Weihang Su, Changyue Wang, Qingyao Ai, et al. (2024). *Unsupervised Real-Time Hallucination Detection based on the Internal States of Large Language Models.* Annual Meeting of the Association for Computational Linguistics.

Junyu Luo, Cao Xiao, and Fenglong Ma (2023). *Zero-Resource Hallucination Prevention for Large Language Models.* Conference on Empirical Methods in Natural Language Processing.

Jun Wu, Q. Liu, Ding Wang, et al. (2024). *Logical Closed Loop: Uncovering Object Hallucinations in Large Vision-Language Models.* Annual Meeting of the Association for Computational Linguistics.

Xuweiyi Chen, Ziqiao Ma, Xuejun Zhang, et al. (2024). *Multi-Object Hallucination in Vision-Language Models.* Neural Information Processing Systems.

Kening Zheng, Junkai Chen, Yibo Yan, et al. (2024). *Reefknot: A Comprehensive Benchmark for Relation Hallucination Evaluation, Analysis and Mitigation in Multimodal Large Language Models.* Annual Meeting of the Association for Computational Linguistics.

Xiang Chen, Chenxi Wang, Yida Xue, et al. (2024). *Unified Hallucination Detection for Multimodal Large Language Models.* Annual Meeting of the Association for Computational Linguistics.

Pranab Sahoo, Prabhash Meharia, Akash Ghosh, et al. (2024). *A Comprehensive Survey of Hallucination in Large Language, Image, Video and Audio Foundation Models.* Conference on Empirical Methods in Natural Language Processing.

Ruiyang Zhang, Hu Zhang, and Zhedong Zheng (2024). *VL-Uncertainty: Detecting Hallucination in Large Vision-Language Model via Uncertainty Estimation.* arXiv.org.

Han Qiu, Jiaxing Huang, Peng Gao, et al. (2024). *LongHalQA: Long-Context Hallucination Evaluation for MultiModal Large Language Models*. arXiv.org.

Jio Oh, Soyeon Kim, Junseok Seo, et al. (2024). *ERBench: An Entity-Relationship based Automatically Verifiable Hallucination Benchmark for Large Language Models*. Neural Information Processing Systems.

Yue Zhang, Leyang Cui, Wei Bi, et al. (2023). *Alleviating Hallucinations of Large Language Models through Induced Hallucinations*. North American Chapter of the Association for Computational Linguistics.

Hao Yin, Guangzong Si, and Zilei Wang (2025). *ClearSight: Visual Signal Enhancement for Object Hallucination Mitigation in Multimodal Large Language Models*. Computer Vision and Pattern Recognition.

Tanya Goyal, Jiacheng Xu, J. Li, et al. (2021). *Training Dynamics for Text Summarization Models*. Findings.

Chun-Yi Kuan, Wei-Ping Huang, and Hung-yi Lee (2024). *Understanding Sounds, Missing the Questions: The Challenge of Object Hallucination in Large Audio-Language Models*. Interspeech.

Prannay Kaul, Zhizhong Li, Hao Yang, et al. (2024). *THRONE: An Object-Based Hallucination Benchmark for the Free-Form Generations of Large Vision-Language Models*. Computer Vision and Pattern Recognition.

Moon Ye-Bin, Nam Hyeon-Woo, Wonseok Choi, et al. (2024). *BEAF: Observing BEfore-AFter Changes to Evaluate Hallucination in Vision-language Models*. European Conference on Computer Vision.

Linxi Zhao, Yihe Deng, Weitong Zhang, et al. (2024). *Mitigating Object Hallucination in Large Vision-Language Models via Image-Grounded Guidance*. Unpublished manuscript.

Hongbin Ye, Tong Liu, Aijia Zhang, et al. (2023). *Cognitive Mirage: A Review of Hallucinations in Large Language Models*. LKM@IJCAI.

Hanning Zhang, Shizhe Diao, Yong Lin, et al. (2023). *R-Tuning: Instructing Large Language Models to Say 'I Don't Know'*. North American Chapter of the Association for Computational Linguistics.

Yanyang Li, Jianqiao Zhao, M. Lyu, et al. (2022). *Eliciting Knowledge from Large Pre-Trained Models for Unsupervised Knowledge-Grounded Conversation*. Conference on Empirical Methods in Natural Language Processing.

Qidong Huang, Xiao-wen Dong, Pan Zhang, et al. (2023). *OPERA: Alleviating Hallucination in Multi-Modal Large Language Models via Over-Trust Penalty and Retrospection-Allocation*. Computer Vision and Pattern Recognition.

Jianheng Tang, Qifan Zhang, Yuhan Li, et al. (2024). *GraphArena: Evaluating and Exploring Large Language Models on Graph Computation*. International Conference on Learning Representations.

Yuhan Fu, Ruobing Xie, Xingwu Sun, et al. (2024). *Mitigating Hallucination in Multimodal Large Language Model via Hallucination-targeted Direct Preference Optimization*. Annual Meeting of the Association for Computational Linguistics.

Shunyu Yao, Jeffrey Zhao, Dian Yu, et al. (2022). *ReAct: Synergizing Reasoning and Acting in Language Models*. International Conference on Learning Representations.

Shankar Kanthara, Rixie Tiffany Ko Leong, Xiang Lin, et al. (2022). *Chart-to-Text: A Large-Scale Benchmark for Chart Summarization*. Annual Meeting of the Association for Computational Linguistics.

Boseop Kim, Hyoungseok Kim, Sang-Woo Lee, et al. (2021). *What Changes Can Large-scale Language Models Bring? Intensive Study on HyperCLOVA: Billions-scale Korean Generative Pretrained Transformers*. Conference on Empirical Methods in Natural Language Processing.

Mengzhou Xia, Mikel Artetxe, Chunting Zhou, et al. (2022). *Training Trajectories of Language Models Across Scales*. Annual Meeting of the Association for Computational Linguistics.

Roee Aharoni, Shashi Narayan, Joshua Maynez, et al. (2022). *mFACE: Multilingual Summarization with Factual Consistency Evaluation*. Annual Meeting of the Association for Computational Linguistics.

Haopeng Zhang, Semih Yavuz, Wojciech Kryscinski, et al. (2022). *Improving the Faithfulness of Abstractive Summarization via Entity Coverage Control*. NAACL-HLT.

Wenhui Jiang, Minwei Zhu, Yuming Fang, et al. (2022). *Visual Cluster Grounding for Image Captioning*. IEEE Transactions on Image Processing.

Hongmin Wang (2020). *Revisiting Challenges in Data-to-Text Generation with Fact Grounding*. International Conference on Natural Language Generation.

Sihao Chen, S. Buthpitiya, Alex Fabrikant, et al. (2022). *PropSegmEnt: A Large-Scale Corpus for Proposition-Level Segmentation and Entailment Recognition*. Annual Meeting of the Association for Computational Linguistics.

Tomasz Korbak, Hady ElSahar, Germán Kruszewski, et al. (2021). *Controlling Conditional Language Models without Catastrophic Forgetting*. International Conference on Machine Learning.

K. Raman, Iftekhar Naim, Jiecao Chen, et al. (2022). *Transforming Sequence Tagging Into A Seq2Seq Task*. Conference on Empirical Methods in Natural Language Processing.

Ling Liu, and Mans Hulden (2021). *Can a Transformer Pass the Wug Test? Tuning Copying Bias in Neural Morphological Inflection Models*. Annual Meeting of the Association for Computational Linguistics.

Tobias Norlund, Lovisa Hagström, and Richard Johansson (2021). *Transferring Knowledge from Vision to Language: How to Achieve it and how to Measure it?*. BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP.

Tao Wu, E. Chio, Heng-Tze Cheng, et al. (2020). *Zero-Shot Heterogeneous Transfer Learning from Recommender Systems to Cold-Start Search Retrieval*. International Conference on Information and Knowledge Management.

Yongtai Liu, Joshua Maynez, Gonçalo Simões, et al. (2022). *Data Augmentation for Low-Resource Dialogue Summarization*. NAACL-HLT.

L. Jelinek, M. Hauschildt, C. Wittekind, et al. (2016). *Efficacy of Metacognitive Training for Depression: A Randomized Controlled Trial*. Psychotherapy and Psychosomatics.

Vikas Raunak, and Arul Menezes (2022). *Finding Memo: Extractive Memorization in Constrained Sequence Generation Tasks*. Conference on Empirical Methods in Natural Language Processing.

Swarnadeep Saha, Xinyan Velocity Yu, Mohit Bansal, et al. (2022). *MURMUR: Modular Multi-Step Reasoning for Semi-Structured Data-to-Text Generation*. Annual Meeting of the Association for Computational Linguistics.

Soomin Ham, Kibaek Park, Yeongjun Jang, et al. (2021). *KSL-Guide: A Large-scale Korean Sign Language Dataset Including Interrogative Sentences for Guiding the Deaf and Hard-of-Hearing*. IEEE International Conference on Automatic Face Gesture Recognition.

Eunji Jeong, Sungwoo Cho, Gyeong-In Yu, et al. (2018). *JANUS: Fast and Flexible Deep Learning via Symbolic Graph Execution of Imperative Programs*. Symposium on Networked Systems Design and Implementation.

Akhil Kedia, Mohd Abbas Zaidi, and Haejun Lee (2022). *FiE: Building a Global Probability Space by Leveraging Early Fusion in Encoder for Open-Domain Question Answering*. Conference on Empirical Methods in Natural Language Processing.

A. Gallace, and A. Gallace (2010). *Touch and the body: The role of the somatosensory cortex in tactile awareness*. Unpublished manuscript.

Xintong Li, Aleksandre Maskharashvili, S. Stevens-Guille, et al. (2020). *Leveraging Large Pretrained Models for WebNLG 2020*. WEBNLG.

Tom Chao Zhou, Chin-Yew Lin, Irwin King, et al. (2011). *Learning to Suggest Questions in Online Forums*. AAAI Conference on Artificial Intelligence.

Shin In-Jae, Byungkwen Song, and D. Eom (2016). *Auto-Mapping and Configuration Method of IEC 61850 Information Model Based on OPC UA*. Unpublished manuscript.

Eunji Jeong, Sungwoo Cho, Gyeong-In Yu, et al. (2019). *Speculative Symbolic Graph Execution of Imperative Deep Learning Programs*. ACM SIGOPS Operating Systems Review.

Nazneen Rajani, Mihaela A. Bornea, and Ken Barker (2017). *Stacking With Auxiliary Features for Entity Linking in the Medical Domain*. Workshop on Biomedical Natural Language Processing.

Shuhe Wang, Xiaofei Sun, Xiaoya Li, et al. (2023). *GPT-NER: Named Entity Recognition via Large Language Models*. North American Chapter of the Association for Computational Linguistics.

N. Alshahwan, Jubin Chheda, Anastasia Finogenova, et al. (2024). *Automated Unit Test Improvement using Large Language Models at Meta*. SIGSOFT FSE Companion.

Wei Zou, Runpeng Geng, Binghui Wang, et al. (2024). *PoisonedRAG: Knowledge Poisoning Attacks to Retrieval-Augmented Generation of Large Language Models*. arXiv.org.

Ekaterina Fadeeva, Aleksandr Rubashevskii, Artem Shelmanov, et al. (2024). *Fact-Checking the Output of Large Language Models via Token-Level Uncertainty Quantification*. Annual Meeting of the Association for Computational Linguistics.

Thuat Nguyen, C. Nguyen, Viet Dac Lai, et al. (2023). *CulturaX: A Cleaned, Enormous, and Multilingual Dataset for Large Language Models in 167 Languages*. International Conference on Language Resources and Evaluation.

Wei Zou, Runpeng Geng, Binghui Wang, et al. (2024). *PoisonedRAG: Knowledge Corruption Attacks to Retrieval-Augmented Generation of Large Language Models*. Unpublished manuscript.

Huao Li, Yu Quan Chong, Simon Stepputtis, et al. (2023). *Theory of Mind for Multi-Agent Collaboration via Large Language Models*. Conference on Empirical Methods in Natural Language Processing.

Yiming Wang, Zhuosheng Zhang, and Rui Wang (2023). *Element-aware Summarization with Large Language Models: Expert-aligned Evaluation and Chain-of-Thought Method*. Annual Meeting of the Association for Computational Linguistics.

Yuyan Chen, Qiang Fu, Yichen Yuan, et al. (2023). *Hallucination Detection: Robustly Discerning Reliable Answers in Large Language Models*. International Conference on Information and Knowledge Management.

S. Gilbert, J. Kather, and Aidan Hogan (2024). *Augmented non-hallucinating large language models as medical information curators*. npj Digit. Medicine.

Sunkyu Kim, Choong-kun Lee, and Seung-seob Kim (2024). *Large Language Models: A Guide for Radiologists*. Korean Journal of Radiology.

Jianing Wang, Junda Wu, Yupeng Hou, et al. (2024). *InstructGraph: Boosting Large Language Models via Graph-centric Instruction Tuning and Preference Alignment*. Annual Meeting of the Association for Computational Linguistics.

Yuheng Huang, Jiayang Song, Zhijie Wang, et al. (2023). *Look Before You Leap: An Exploratory Study of Uncertainty Measurement for Large Language Models*. arXiv.org.

Linxi Zhao, Yihe Deng, Weitong Zhang, et al. (2024). *Mitigating Object Hallucination in Large Vision-Language Models via Classifier-Free Guidance*. arXiv.org.

Florian Leiser, S. Eckhardt, Valentin Leuthe, et al. (2024). *HILL: A Hallucination Identifier for Large Language Models*. International Conference on Human Factors in Computing Systems.

Lars Malmqvist (2024). *Sycophancy in Large Language Models: Causes and Mitigations.* arXiv.org.

Sheng-Chieh Lin, Luyu Gao, Barlas Oğuz, et al. (2024). *FLAME: Factuality-Aware Alignment for Large Language Models.* Neural Information Processing Systems.

Xingxuan Li, Ruochen Zhao, Yew Ken Chia, et al. (2023). *Chain of Knowledge: A Framework for Grounding Large Language Models with Structured Knowledge Bases.* arXiv.org.

Fan Ma, Xiaojie Jin, Heng Wang, et al. (2023). *Vista-llama: Reducing Hallucination in Video Language Models via Equal Distance to Visual Tokens.* Computer Vision and Pattern Recognition.

Peiyang Song, Kaiyu Yang, and Anima Anandkumar (2024). *Towards Large Language Models as Copilots for Theorem Proving in Lean.* arXiv.org.

Che Jiang, Biqing Qi, Xiangyu Hong, et al. (2024). *On Large Language Models' Hallucination with Regard to Known Facts.* North American Chapter of the Association for Computational Linguistics.

Peiyang Song, Kaiyu Yang, and Anima Anandkumar (2024). *Lean Copilot: Large Language Models as Copilots for Theorem Proving in Lean.* NeuS.

Haochen Liu, Song Wang, Yaochen Zhu, et al. (2024). *Knowledge Graph-Enhanced Large Language Models via Path Selection.* Annual Meeting of the Association for Computational Linguistics.

Yuhong Sun, Zhangyue Yin, Qipeng Guo, et al. (2024). *Benchmarking Hallucination in Large Language Models Based on Unanswerable Math Word Problem.* International Conference on Language Resources and Evaluation.

Chen Ling, Xujiang Zhao, Wei Cheng, et al. (2024). *Uncertainty Quantification for In-Context Learning of Large Language Models.* North American Chapter of the Association for Computational Linguistics.

Moxin Li, Wenjie Wang, Fuli Feng, et al. (2024). *Think Twice Before Trusting: Self-Detection for Large Language Models through Comprehensive Answer Reflection*. Conference on Empirical Methods in Natural Language Processing.

Andrew L Smith, Felix Greaves, and T. Panch (2023). *Hallucination or Confabulation? Neuroanatomy as metaphor in Large Language Models*. PLOS Digital Health.

Savyasachi V. Shah (2024). *Accuracy, Consistency, and Hallucination of Large Language Models When Analyzing Unstructured Clinical Notes in Electronic Medical Records.*. JAMA Network Open.

Zhenyu Pan, Haozheng Luo, Manling Li, et al. (2024). *Chain-of-Action: Faithful and Multimodal Question Answering through Large Language Models*. International Conference on Learning Representations.

Hengran Zhang, Ruqing Zhang, J. Guo, et al. (2024). *Are Large Language Models Good at Utility Judgments?*. Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.

Jianheng Tang, Qifan Zhang, Yuhan Li, et al. (2024). *GraphArena: Benchmarking Large Language Models on Graph Computational Problems*. arXiv.org.

Xiongtao Zhou, Jie He, Yuhua Ke, et al. (2024). *An Empirical Study on Parameter-Efficient Fine-Tuning for MultiModal Large Language Models*. Annual Meeting of the Association for Computational Linguistics.

Ruilin Zhao, Feng Zhao, Long Wang, et al. (2024). *KG-CoT: Chain-of-Thought Prompting of Large Language Models over Knowledge Graphs for Knowledge-Aware Question Answering*. International Joint Conference on Artificial Intelligence.

Zhenyu Pan, Haozheng Luo, Manling Li, et al. (2024). *Conv-CoA: Improving Open-domain Question Answering in Large Language Models via Conversational Chain-of-Action*. arXiv.org.

Derui Zhu, Dingfan Chen, Qing Li, et al. (2024). *PoLLMgraph: Unraveling Hallucinations in Large Language Models via State Transition Dynamics*. NAACL-HLT.

Wan Zhang, and Jing Zhang (2025). *Hallucination Mitigation for Retrieval-Augmented Large Language Models: A Review*. Mathematics.

Lida Chen, Zujie Liang, Xintao Wang, et al. (2024). *Teaching Large Language Models to Express Knowledge Boundary from Their Own Signals*. Proceedings of the 3rd Workshop on Towards Knowledgeable Foundation Models (KnowFM).

Stefan Dernbach, Khushbu Agarwal, Alejandro Zuniga, et al. (2024). *GLaM: Fine-Tuning Large Language Models for Domain Knowledge Graph Alignment via Neighborhood Partitioning and Generative Subgraph Encoding*. AAAI Spring Symposia.

Jiageng Wu, Xian Wu, and Jie Yang (2024). *Guiding Clinical Reasoning with Large Language Models via Knowledge Seeds*. International Joint Conference on Artificial Intelligence.

Jaewoo Ahn, Taehyun Lee, Junyoung Lim, et al. (2024). *TimeChara: Evaluating Point-in-Time Character Hallucination of Role-Playing Large Language Models*. Annual Meeting of the Association for Computational Linguistics.

Xinyi Mou, Zejun Li, Hanjia Lyu, et al. (2024). *Unifying Local and Global Knowledge: Empowering Large Language Models as Political Experts with Knowledge Graphs*. The Web Conference.

Derong Xu, Ziheng Zhang, Zhihong Zhu, et al. (2024). *Editing Factual Knowledge and Explanatory Ability of Medical Large Language Models*. International Conference on Information and Knowledge Management.

Xiangkun Hu, Dongyu Ru, Lin Qiu, et al. (2024). *RefChecker: Reference-based Fine-grained Hallucination Checker and Benchmark for Large Language Models*. arXiv.org.

Subhojyoti Mukherjee, Anusha Lalitha, Sailik Sengupta, et al. (2024). *Multi-Objective Alignment of Large Language Models Through Hypervolume Maximization*. arXiv.org.

Qirui Jiao, Daoyuan Chen, Yilun Huang, et al. (2024). *Enhancing Multimodal Large Language Models with Vision Detection Models: An Empirical Study*. arXiv.org.

Hao Ding, Ziwei Fan, Ingo Gühring, et al. (2024). *Reasoning and Planning with Large Language Models in Code Development*. Knowledge Discovery and Data Mining.

Chengpeng Wang, Wuqi Zhang, Zian Su, et al. (2024). *Sanitizing Large Language Models in Bug Detection with Data-Flow*. Conference on Empirical Methods in Natural Language Processing.

Zhuo Chen, Jiawei Liu, Haotan Liu, et al. (2024). *Black-Box Opinion Manipulation Attacks to Retrieval-Augmented Generation of Large Language Models*. arXiv.org.

Xiaohua Wang, Yuliang Yan, Longtao Huang, et al. (2023). *Hallucination Detection for Generative Large Language Models by Bayesian Sequential Estimation*. Conference on Empirical Methods in Natural Language Processing.

Wei Jie Yeo, Teddy Ferdinan, Przemysław Kazienko, et al. (2024). *Self-training Large Language Models through Knowledge Detection*. Conference on Empirical Methods in Natural Language Processing.

Sihao Hu, Tiansheng Huang, and Ling Liu (2024). *PokeLLMon: A Human-Parity Agent for Pokemon Battles with Large Language Models*. arXiv.org.

Zhenhong Zhang, Jiajing Chen, Weiyan Shi, et al. (2024). *Contrastive Learning for Knowledge-Based Question Generation in Large Language Models*. 2024 5th International Conference on Intelligent Computing and Human-Computer Interaction (ICHCI).

Dingkang Yang, Dongling Xiao, Jinjie Wei, et al. (2024). *Improving Factuality in Large Language Models via Decoding-Time Hallucinatory and Truthful Comparators*. AAAI Conference on Artificial Intelligence.

Xiaodong Yu, Hao Cheng, Xiaodong Liu, et al. (2023). *ReEval: Automatic Hallucination Evaluation for Retrieval-Augmented Large Language Models via Transferable Adversarial Attacks*. NAACL-HLT.

Zishan Gu, Changchang Yin, Fenglin Liu, et al. (2024). *MedVH: Towards Systematic Evaluation of Hallucination for Large Vision Language Models in the Medical Context.* arXiv.org.

Lihui Liu, Zihao Wang, Ruizhong Qiu, et al. (2024). *Logic Query of Thoughts: Guiding Large Language Models to Answer Complex Logic Queries with Knowledge Graphs.* arXiv.org.

B. Woo, Tom Huynh, Arthur Tang, et al. (2024). *Transforming nursing with large language models: from concept to practice.*. European Journal of Cardiovascular Nursing.

Xiutian Zhao, Ke Wang, and Wei Peng (2024). *Measuring the Inconsistency of Large Language Models in Preferential Ranking.* KNOWLLM.

Xinying Qian, Ying Zhang, Yu Zhao, et al. (2024). $TimeR^4 : Time-aware Retrieval-Augmented Large Language Models for Temporal Knowledge Graph Question Answering. Conference$

J. Butler, James Puleo, Michael Harrington, et al. (2024). *From technical to understandable: Artificial Intelligence Large Language Models improve the readability of knee radiology reports.*. Knee Surgery, Sports Traumatology, Arthroscopy.

N. R. Sahoo, Ashita Saxena, Kishan Maharaj, et al. (2024). *Addressing Bias and Hallucination in Large Language Models.* International Conference on Language Resources and Evaluation.

Kaiwen Zuo, and Yirui Jiang (2024). *MedHallBench: A New Benchmark for Assessing Hallucination in Medical Large Language Models.* arXiv.org.

D. J. Parente (2024). *Generative Artificial Intelligence and Large Language Models in Primary Care Medical Education.*. Family Medicine.

Haiyan Zhao, Fan Yang, Himabindu Lakkaraju, et al. (2024). *Opening the Black Box of Large Language Models: Two Views on Holistic Interpretability.* arXiv.org.

Yue Zhou, Henry Peng Zou, Barbara Di Eugenio, et al. (2024). *Large Language Models Are Involuntary Truth-Tellers: Exploiting Fallacy Failure for Jailbreak Attacks*. Conference on Empirical Methods in Natural Language Processing.

Tosin P. Adewumi, Nudrat Habib, Lama Alkhaled, et al. (2024). *On the Limitations of Large Language Models (LLMs): False Attribution*. arXiv.org.

Armin Toroghi, Willis Guo, Mohammad Mahdi Torabi pour, et al. (2024). *Right for Right Reasons: Large Language Models for Verifiable Commonsense Knowledge Graph Question Answering*. Conference on Empirical Methods in Natural Language Processing.

Pittawat Taveekitworachai, Febri Abdullah, and R. Thawonmas (2024). *Null-Shot Prompting: Rethinking Prompting Large Language Models With Hallucination*. Conference on Empirical Methods in Natural Language Processing.

Fanqi Wan, Xinting Huang, Leyang Cui, et al. (2024). *Mitigating Hallucinations of Large Language Models via Knowledge Consistent Alignment*. arXiv.org.

Shayan Meshkat Alsadat, Jean-Raphael Gaglione, D. Neider, et al. (2024). *Using Large Language Models to Automate and Expedite Reinforcement Learning with Reward Machine*. American Control Conference.

Qingxing Cao, Junhao Cheng, Xiaodan Liang, et al. (2024). *VisDiaHalBench: A Visual Dialogue Benchmark For Diagnosing Hallucination in Large Vision-Language Models*. Annual Meeting of the Association for Computational Linguistics.

Kenza Benkirane, Laura Gongas, Shahar Pelles, et al. (2024). *Machine Translation Hallucination Detection for Low and High Resource Languages using Large Language Models*. Conference on Empirical Methods in Natural Language Processing.

Lifeng Jin, Baolin Peng, Linfeng Song, et al. (2024). *Collaborative decoding of critical tokens for boosting factuality of large language models*. arXiv.org.

Yida Mu, Peizhen Bai, Kalina Bontcheva, et al. (2024). *Addressing Topic Granularity and Hallucination in Large Language Models for Topic Modelling*. arXiv.org.

Jun Yu, Yunxiang Zhang, Zerui Zhang, et al. (2024). *RAG-Guided Large Language Models for Visual Spatial Description with Adaptive Hallucination Corrector*. ACM Multimedia.

Maryam Amirizaniani, Jihan Yao, Adrian Lavergne, et al. (2024). *LLMAuditor: A Framework for Auditing Large Language Models Using Human-in-the-Loop*. Unpublished manuscript.

Chen Ling, Xujiang Zhao, Wei Cheng, et al. (2024). *Uncertainty Decomposition and Quantification for In-Context Learning of Large Language Models*. arXiv.org.

Bhaskarjit Sarmah, Dhagash Mehta, Stefano Pasquali, et al. (2023). *Towards reducing hallucination in extracting information from financial reports using Large Language Models*. International Conference on AI-ML-Systems.

Deepak Nathani, David Wang, Liangming Pan, et al. (2023). *MAF: Multi-Aspect Feedback for Improving Reasoning in Large Language Models*. Conference on Empirical Methods in Natural Language Processing.

Cl'ea Chataigner, Afaf Taïk, and G. Farnadi (2024). *Multilingual Hallucination Gaps in Large Language Models*. arXiv.org.

Q. Liu, Xinlong Chen, Yue Ding, et al. (2025). *Attention-guided Self-reflection for Zero-shot Hallucination Detection in Large Language Models*. arXiv.org.

Jongyoon Song, Sangwon Yu, and Sungroh Yoon (2024). *Large Language Models are Skeptics: False Negative Problem of Input-conflicting Hallucination*. arXiv.org.

Liam Barkley, and Brink van der Merwe (2024). *Investigating the Role of Prompting and External Tools in Hallucination Rates of Large Language Models*. arXiv.org.

Chao-Wei Huang, and Yun-Nung Chen (2024). *FactAlign: Long-form Factuality Alignment of Large Language Models*. Conference on Empirical Methods in Natural Language Processing.

B. Malin, Tatiana Kalganova, and Nikoloas Boulgouris (2024). *A review of faithfulness metrics for hallucination assessment in Large Language Models*. IEEE Journal on Selected Topics in Signal Processing.

Hongbang Yuan, Pengfei Cao, Zhuoran Jin, et al. (2024). *Whispers that Shake Foundations: Analyzing and Mitigating False Premise Hallucinations in Large Language Models*. Conference on Empirical Methods in Natural Language Processing.

Shrey Pandit, Jiawei Xu, Junyuan Hong, et al. (2025). *MedHallu: A Comprehensive Benchmark for Detecting Medical Hallucinations in Large Language Models*. arXiv.org.

Samuel C. Bellini-Leite (2023). *Dual Process Theory for Large Language Models: An overview of using Psychology to address hallucination and reliability issues*. Adaptive Behavior.

M. Omar, V. Sorin, J. Collins, et al. (2025). *Large Language Models Are Highly Vulnerable to Adversarial Hallucination Attacks in Clinical Decision Support: A Multi-Model Assurance Analysis*. medRxiv.

Yi-Lun Lee, Yi-Hsuan Tsai, and Wei-Chen Chiu (2024). *Delve into Visual Contrastive Decoding for Hallucination Mitigation of Large Vision-Language Models*. arXiv.org.

Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, et al. (2023). *HELMA: A Large-Scale Hallucination Evaluation Benchmark for Large Language Models*. Unpublished manuscript.

Chen Zhang (2023). *User-Controlled Knowledge Fusion in Large Language Models: Balancing Creativity and Hallucination*. arXiv.org.

Muneeswaran Irulandi, Shreya Saxena, Siva Prasad, et al. (2023). *Minimizing Factual Inconsistency and Hallucination in Large Language Models*. arXiv.org.

Derong Xu Xinhang Li, Ziheng Zhang, Zhenxi Lin, et al. (2024). *Harnessing Large Language Models for Knowledge Graph Question Answering via Adaptive Multi-Aspect Retrieval-Augmentation*. arXiv.org.

Hongyi Guo, Zhihan Liu, Yufeng Zhang, et al. (2024). *Can Large Language Models Play Games? A Case Study of A Self-Play Approach.* arXiv.org.

Zikai Xie (2024). *Order Matters in Hallucination: Reasoning Order as Benchmark and Reflexive Prompting for Large-Language-Models.* arXiv.org.

Maryam Amirizaniani, Jihan Yao, Adrian Lavergne, et al. (2024). *Developing a Framework for Auditing Large Language Models Using Human-in-the-Loop.* arXiv.org.

Tianyun Yang, Ziniu Li, Juan Cao, et al. (2025). *Understanding and Mitigating Hallucination in Large Vision-Language Models via Modular Attribution and Intervention.* International Conference on Learning Representations.

Vibhor Agarwal, Yulong Pei, Salwa Alamir, et al. (2024). *CodeMirage: Hallucinations in Code Generated by Large Language Models.* arXiv.org.

Aswin Rrv, Nemika Tyagi, Md Nayem Uddin, et al. (2024). *Chaos with Keywords: Exposing Large Language Models Sycophantic Hallucination to Misleading Keywords and Evaluating Defense Strategies.* Unpublished manuscript.

Mingchen Li, Zaifu Zhan, Han Yang, et al. (2024). *Benchmarking Retrieval-Augmented Large Language Models in Biomedical NLP: Application, Robustness, and Self-Awareness.* arXiv.org.

Shirui Wang, Bohan Xie, Ling Ding, et al. (2024). *SeCor: Aligning Semantic and Collaborative Representations by Large Language Models for Next-Point-of-Interest Recommendations.* ACM Conference on Recommender Systems.

S. Hegselmann, Zejiang Shen, Florian Gierse, et al. (2024). *A Data-Centric Approach To Generate Faithful and High Quality Patient Summaries with Large Language Models.* ACM Conference on Health, Inference, and Learning.

Jun Gao, Huan Zhao, Wei Wang, et al. (2024). *EventRL: Enhancing Event Extraction with Outcome Supervision for Large Language Models.* arXiv.org.

Yao-Hung Tsai, Walter Talbott, and Jian Zhang (2024). *Efficient Non-Parametric Uncertainty Quantification for Black-Box Large Language Models and Decision Planning.* arXiv.org.

Xinxi Chen, Li Wang, Wei Wu, et al. (2024). *Honest AI: Fine-Tuning "Small" Language Models to Say "I Don't Know", and Reducing Hallucination in RAG.* arXiv.org.

MingShan Liu, Shi Bo, and Jialing Fang (2025). *Enhancing Mathematical Reasoning in Large Language Models with Self-Consistency-Based Hallucination Detection.* arXiv.org.

Taolin Zhang, Qizhou Chen, Dongyang Li, et al. (2024). *DAFNet: Dynamic Auxiliary Fusion for Sequential Model Editing in Large Language Models.* Annual Meeting of the Association for Computational Linguistics.

Yuhang Guo, and Zhiyu Wan (2024). *Performance Evaluation of Multimodal Large Language Models (LLaVA and GPT-4-based ChatGPT) in Medical Image Classification Tasks.* IEEE International Conference on Healthcare Informatics.

Weiqing Luo, Chonggang Song, Lingling Yi, et al. (2024). *KELLMRec: Knowledge-Enhanced Large Language Models for Recommendation.* arXiv.org.

Oussama H. Hamid (2024). *Beyond Probabilities: Unveiling the Delicate Dance of Large Language Models (LLMs) and AI-Hallucination.* Conference on Cognitive and Computational Aspects of Situation Management.

Xinxin Zheng, Feihu Che, Jinyang Wu, et al. (2024). *KS-LLM: Knowledge Selection of Large Language Models with Evidence Document for Question Answering.* arXiv.org.

Vipula Rawte, Aman Chadha, Amit P. Sheth, et al. (2024). *Tutorial Proposal: Hallucination in Large Language Models.* International Conference on Language Resources and Evaluation.

Zhibo Yin (2024). *A review of methods for alleviating hallucination issues in large language models.* Applied and Computational Engineering.

Zilu Tang, Rajen Chatterjee, and Sarthak Garg (2025). *Mitigating Hallucinated Translations in Large Language Models with Hallucination-focused Preference Optimization.* North American Chapter of the Association for Computational Linguistics.

Souvik Das, Lifeng Jin, Linfeng Song, et al. (2024). *Entropy Guided Extrapolative Decoding to Improve Factuality in Large Language Models.* International Conference on Computational Linguistics.

Yuxiang Zhang, Jing Chen, Junjie Wang, et al. (2024). *ToolBeHonest: A Multi-level Hallucination Diagnostic Benchmark for Tool-Augmented Large Language Models.* Conference on Empirical Methods in Natural Language Processing.

Derong Xu, Ziheng Zhang, Zhihong Zhu, et al. (2024). *Mitigating Hallucinations of Large Language Models in Medical Information Extraction via Contrastive Decoding.* Conference on Empirical Methods in Natural Language Processing.

Hongjie Zhang, Hourui Deng, Jie Ou, et al. (2025). *Mitigating spatial hallucination in large language models for path planning via prompt engineering.* Scientific Reports.

Ali Ahmadi (2024). *Unravelling the Mysteries of Hallucination in Large Language Models: Strategies for Precision in Artificial Intelligence Language Generation.* Asian Journal of Computer Science and Technology.

M. Abdelghafour, Mohammed Mabrouk, and Zaki Taha (2024). *Hallucination Mitigation Techniques in Large Language Models.* International Journal of Intelligent Computing and Information Sciences.

Xiaoling Zhou, Mingjie Zhang, Zhemg Lee, et al. (2025). *HaDeMiF: Hallucination Detection and Mitigation in Large Language Models.* International Conference on Learning Representations.

Amin Karbasi, Omar Montasser, John Sous, et al. (2025). *(Im)possibility of Automated Hallucination Detection in Large Language Models.* arXiv.org.

Hamdy Mubarak, Hend Suliman Al-Khalifa, and Khaloud Suliman Alkhalefah (2024). *Halwasa: Quantify and Analyze Hallucinations in Large Language Models: Arabic as a Case Study*. International Conference on Language Resources and Evaluation.

Wenbo Zhang, Zihang Xu, and Hengrui Cai (2024). *Recognizing Limits: Investigating Infeasibility in Large Language Models*. Unpublished manuscript.

Mahmud Omar, Vera Sorin, Jeremy D. Collins, et al. (2025). *Multi-model assurance analysis showing large language models are highly vulnerable to adversarial hallucination attacks during clinical decision support*. Communications Medicine.

Zhengjie Gao, Xuanzi Liu, Yuanshuai Lan, et al. (2024). *A Brief Survey on Safety of Large Language Models*. Journal of computer information technology.

Andreas Pester, Ahmed Tammaa, Christian Gütl, et al. (2024). *Conversational Agents, Virtual Worlds, and Beyond: A Review of Large Language Models Enabling Immersive Learning*. IEEE Global Engineering Education Conference.

Kangxi Wu, Liang Pang, Huawei Shen, et al. (2024). *Enhancing Training Data Attribution for Large Language Models with Fitting Error Consideration*. Conference on Empirical Methods in Natural Language Processing.

Yahan Tu, Rui Hu, and Jitao Sang (2024). *ODE: Open-Set Evaluation of Hallucinations in Multimodal Large Language Models*. Computer Vision and Pattern Recognition.