Review

A Survey on Hallucination in Large Language and Foundation Models

Pegah Ahadian and Qiang Guan *

Posted Date: 15 April 2025

doi: 10.20944/preprints202504.1236.v1

Keywords: LLM; Hallucination; Foundation Models; Generative AI

 doi:10.20944/preprints202504.1236.v1

Review

A Survey on Hallucination in Large Language and Foundation Models

Pegah Ahadian and Qiang Guan *

Kent State University, OH, USA
* Correspondence: qguan@kent.edu

Abstract: Generative text models, particularly large language models (LLMs) and foundation models, have influenced numerous fields, including high-quality text generation, reasoning, and multimodal synthesis. These models have been widely applied in healthcare, legal analysis, and scientific research. However, where accuracy and reliability are critical, generative text models pose a significant risk due to hallucination, where generated outputs include incorrect factuality, fabricated, or misleading information. In this survey, we present a review of hallucination in generative AI, covering its taxonomy, detection methods, mitigation strategies, and evaluation benchmarks. We first establish a structured taxonomy, distinguishing between intrinsic vs. extrinsic hallucination and factual vs. semantic hallucination, also discussing task-specific variations in areas such as summarization, machine translation, and dialogue generation. Next, we examine state-of-the-art hallucination detection techniques, including uncertainty estimation, retrieval-augmented generation (RAG), self-consistency validation, and internal state monitoring. We further explore mitigation strategies, such as fine-tuning, reinforcement learning from human feedback (RLHF), knowledge injection, adversarial training, and contrastive learning. Additionally, we review key evaluation metrics and benchmarks, including FEVER, TruthfulQA, HALL-E, and Entity-Relationship-Based Hallucination Benchmarks (ERBench), which serve as standardized measures for assessing hallucination severity. Despite notable efforts, hallucination remains an open challenge, necessitating further improvements in real-time detection, multimodal hallucination evaluation, and trustworthiness frameworks. We show critical research gaps including the need for standardized hallucination taxonomies, scalable mitigation techniques, and human-AI hybrid verification m e thods. O ur s urvey a ims t o s erve a s a f oundational resource for researchers and practitioners, providing insights into current methodologies and guiding future advancements in trustworthy and explainable generative AI.

Keywords: LLM; Hallucination; Foundation Models; Generative AI

Introduction

Generative models, particularly large language models (LLMs) and foundation models, have demonstrated unprecedented capabilities in natural language processing (NLP), reasoning, and multimodal generation. Transformer-based architectures such as GPT-4, PaLM, and LLaMA have enabled remarkable progress in machine translation, summarization, dialogue systems, and data-to-text generation [1,2]. However, despite these advancements, a persistent and critical challenge in these models is hallucination, where generated outputs contain factually incorrect, fabricated, or misleading information [3,4]. This phenomenon poses a serious risk in real-world applications, particularly in high-stakes domains such as healthcare, legal analysis, and financial decision-making where misinformation can lead to harmful consequences [5,6].

0.1. Historical Context and Significance

The issue of hallucination in artificial intelligence (AI) traces back to early Natural Language Generation (NLG) models, where researchers observed that statistical and neural-based language

 doi:10.20944/preprints202504.1236.v1

models often generated syntax-fluent text, that was semantically incorrect [7]. Early works in sequence-to-sequence models highlighted that likelihood-based training objectives could lead to degeneration, producing either repetitive or nonsensical outputs [3,8]. With the rise of Transformer-based architectures, hallucination became a more prominent issue, as models trained on large-scale, noisy datasets learned to generate fluent but unreliable information [9,10].

0.2. Types of Hallucination

Hallucination in LLMs can take multiple forms, broadly categorized as intrinsic vs. extrinsic hallucination and factual vs. semantic hallucination [3,11]:

Intrinsic hallucination refers to internal inconsistencies within the generated text, where the model contradicts itself [3].

Extrinsic hallucination occurs when a model introduces content not present in the input context or real-world knowledge [5].

Factual hallucination involves generating statements that contradict verified facts [4].

Semantic hallucination arises when generated text is grammatically correct but lacks logical coherence or relevance to the prompt [12].

The causes of hallucination can be attributed to both data-related and model-related factors:

Training Data Issues:

Source-reference divergence in datasets (e.g., summarization datasets where summaries contain additional, unsupported claims) [13].

Noisy, incomplete, or biased training data, leading the model to learn incorrect associations [9,10].

Model-Specific Factors:

Overgeneralization and misalignment with factual constraints due to autoregressive generation [14].

Reinforcement Learning from Human Feedback (RLHF) over-optimization, which can cause models to prioritize fluency and human preference over factual accuracy [15,16].

Decoding Strategies:

Techniques like beam search and nucleus sampling can amplify hallucination by favoring more probable but less accurate tokens [8].

Exposure bias, where models generate text token-by-token without reevaluating the broader context [17,18].

0.3. Challenges and Open Research Questions

Despite advancements in detection and mitigation strategies, hallucination remains a major unsolved challenge. Current hallucination detection techniques such as uncertainty estimation, retrieval-augmented generation (RAG), and self-consistency validation have limitations in terms of scalability, reliability, and real-time applicability [19,20]. Mitigation strategies including fine-tuning, RLHF, adversarial training, and contrastive learning have shown promise but often introduce new challenges such as overcorrection and reduced model fluency [10,15]. In this survey, we provide a comprehensive review of hallucination in LLMs and foundation models including:

- A structured taxonomy of hallucination types across different NLP tasks
- An analysis of state-of-the-art detection methods, including probabilistic, contrastive, and retrieval-based approaches
- A detailed discussion on mitigation techniques, including reinforcement learning, prompt optimization, and adversarial fine-tuning
- A review of evaluation metrics and benchmarks, such as FEVER, TruthfulQA, and the Hallucination Evaluation Benchmark (HALL-E)
- Research gaps and future directions, including the need for standardized hallucination definitions, real-time detection techniques, and multimodal hallucination evaluation

 doi:10.20944/preprints202504.1236.v1

By synthesizing existing research, our survey aims to provide a foundational resource for researchers and practitioners, offering insights into current methodologies and guiding the development of more trustworthy and explainable generative AI systems.

1. Taxonomy of Hallucination

Hallucination in generative models, particularly Large Language Models (LLMs) and foundation models, has been widely studied due to its implications for trustworthiness and reliability, see Figure 1. Researchers have proposed various classification frameworks to understand the nature, causes, and impact of hallucinations. In this section, we systematically categorize hallucination types, referencing key studies that have shaped this taxonomy.

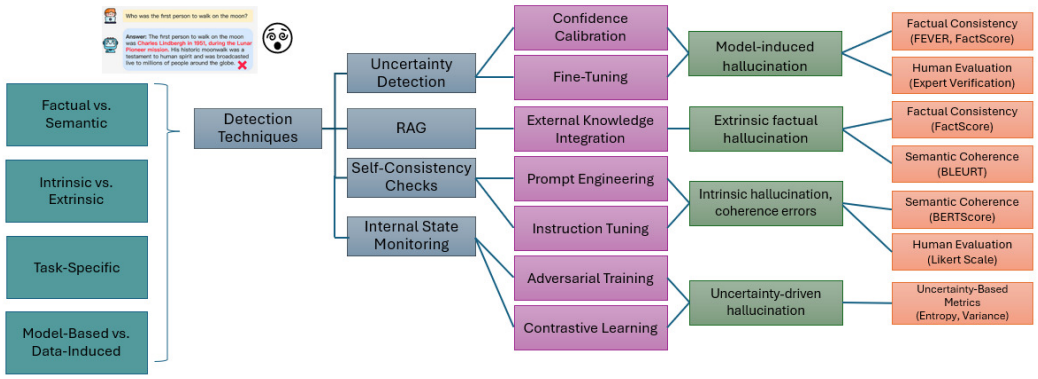

Figure 1. Overview of hallucination taxonomy, detection techniques, mitigation strategies, and evaluation metrics in generative AI models. The figure categorizes hallucinations into factual vs. semantic, intrinsic vs. extrinsic, task-specific, and model-based vs. data-induced. Detection methods include uncertainty detection, retrieval-augmented generation (RAG), self-consistency checks, and internal state monitoring. Mitigation strategies such as confidence calibration, fine-tuning, external knowledge integration, prompt engineering, instruction tuning, adversarial training, and contrastive learning target different types of hallucination. Evaluation metrics, including factual consistency (FEVER, FactScore), semantic coherence (BLEURT, BERTScore), human evaluation, and uncertainty-based metrics, assess hallucination severity and model reliability.

1.1. Intrinsic vs. Extrinsic Hallucination

A widely accepted distinction in hallucination taxonomy is between intrinsic and extrinsic hallucinations. Ji et al. (2023) [3] define intrinsic hallucination as a phenomenon where the generated text contradicts itself internally, without reference to any external knowledge. This type of hallucination arises from logical inconsistencies within the generated output. For example, a model summarizing an article might state in one sentence that a person was born in 1980 and later claim they were born in 1975, indicating an internally inconsistent response. Intrinsic hallucinations often occur in long-form text generation and reasoning tasks, where models struggle to maintain coherence over multiple sentences [3,21].

In contrast, extrinsic hallucination occurs when a model generates information that is factually incorrect with respect to an external knowledge source. Huang et al. (2025) [22] highlight that this type of hallucination is particularly problematic in factual generation tasks such as question answering and text summarization, where faithfulness to the source is critical [22]. In a machine translation scenario, extrinsic hallucination manifests when ML translated text introduces concepts that are absent in the source text leading to misleading or fabricated content [11]. Pagnoni et al. (2021) demonstrate that hallucination in summarization often stems from extrinsic fabrication where generated text includes facts not present in the original article [23]. These findings underscore the necessity of differentiating hallucination types based on their reliance on external factuality.

 doi:10.20944/preprints202504.1236.v1

1.2. Factual vs. Semantic Hallucination

Another classification of hallucinations is based on the distinction between factual and semantic inconsistencies. Factual hallucination refers to instances where a model generates statements that contradict real-world knowledge. Rawte et al. (2023) [4] categorize this type as one of the most critical issues in applications where factual correctness is paramount such as medical AI and legal reasoning. For example, an LLM generating the statement, "The speed of light in a vacuum is 100,000 km/s," represents a factual hallucination, as the correct value is approximately 299,792 km/s.

In contrast, semantic hallucination [24] refers to text that is grammatically correct and seemingly plausible but lacks logical coherence or meaningful association with the given input. Zhang et al. (2023) [12] describe how dialogue systems often suffer from this form of hallucination, where responses are fluent yet irrelevant or nonsensical. For instance, if a user asks, "What is the capital of Japan?" and the model responds with, "Tokyo is known for its rich culinary traditions," the response, while related to Tokyo, does not directly answer the question. Azaria and Mitchell (2023) [14] explore this issue further by investigating how LLMs internally assess truthfulness and coherence, revealing that hallucinations may not always stem from a lack of knowledge but rather from a model's inability to maintain contextual consistency [21,25].

1.3. Task-Specific Hallucination Categories

Different generative tasks exhibit hallucinations in unique ways, making task-specific classifications essential. In machine translation, Guerreiro et al. (2022) [11] report that hallucination often arises when models produce translations that contain phrases or concepts absent in the original source text. These hallucinations are particularly evident in low-resource language pairs where training data scarcity leads to excessive reliance on learned priors.

In text summarization, hallucinations take the form of overgeneralization or the introduction of unsupported claims. Pagnoni et al. (2021) [23] demonstrate that models frequently hallucinate when forced to generate summaries of documents containing ambiguous or incomplete information. This issue is escalated by reference-summary misalignment in training datasets, where models are incentivised to generate additional information beyond what is present in the source document [13]. In dialogue systems, Azaria and Mitchell (2023) [14] highlight that hallucination manifests as responses that confidently assert false information often due to overgeneralization from training data. This can be particularly misleading in customer service bots and medical AI assistants where incorrect yet authoritative-sounding responses can mislead users (Shuster et al., 2022).

Multimodal hallucination has become increasingly relevant with the rise of vision-language models. Bai et al. (2023) [15] show that image captioning systems can generate descriptions containing objects that are not present in the image leading to perceptual errors. For instance, a model might describe an image as containing a cat when there is none illustrating how hallucination extends beyond textual modalities into visual and multimodal AI systems [26].

1.4. Model-Based vs. Data-Induced Hallucination

The origin of hallucination can be traced to either model-centric or data-centric factors. Shuster et al. (2022) [9] and Mckenna et al. (2023) [27] categorize hallucinations based on their root causes, distinguishing between model-based and data-induced hallucinations.

Model-based hallucination arises from biases in model architecture, fine-tuning strategies, or reinforcement learning objectives. Burns et al. (2023) [28] highlight how reinforcement learning from human feedback (RLHF) can sometimes exacerbate hallucination by reinforcing plausible but incorrect patterns in generated responses. Similarly, autoregressive generation methods can cause models to make errors progressively while constructing long-form text, amplifying small inaccuracies into larger hallucinations [10]. Data-induced hallucination, by contrast, occurs when models are trained on incomplete, noisy, or biased datasets. Zhao et al. (2021) [10] discuss how low-quality or adversarially corrupted datasets contribute to hallucination, particularly in domains where factual accuracy is

 doi:10.20944/preprints202504.1236.v1

crucial. In an analysis of contrast-consistent search (CCS), Burns et al. (2023) [28] show that models trained on datasets with inconsistencies tend to generate hallucinations at a higher frequency, even when prompted with factual queries.

1.5. Summary of Hallucination Taxonomy

The classification of hallucinations into intrinsic vs. extrinsic, factual vs. semantic, and task-specific categories provides a structured framework for understanding their diverse manifestations. The source of hallucination, whether from model biases or data limitations, further influences how they can be effectively detected and mitigated. As shown in Table 1, each category of hallucination presents distinct challenges and necessitates tailored detection techniques.

Table 1. Taxonomy of Hallucination in Generative Models.

Category	Definition	Example	Key References
Intrinsic Hallucination	Internal inconsistency within the generated text	A summary contradicting itself	Ji et al. (2023) [3]
Extrinsic Hallucination	Misinformation that diverges from the input or real-world knowledge	A fabricated fact in a generated response	Huang et al. (2025) [5, 22]
Factual Hallucination	Statements that contradict real-world facts	Incorrect scientific claims	Rawte et al. (2023) [4]
Semantic Hallucination	Fluent but logically incoherent responses	An irrelevant chatbot reply	Zhang et al. (2023) [29]
Task-Specific Hallucination	Hallucinations in different NLP tasks	Incorrect translations, misleading summaries, hallucinated image descriptions	Dale et al. (2022), Bai et al. (2023) [15]
Model-Based Hallucination	Hallucination due to training biases or fine-tuning strategies	Errors introduced by reinforcement learning objectives	Burns et al. (2023) [28], Ouyang et al. (2022) [16]
Data-Induced Hallucination	Hallucination due to incomplete or biased training data	Incorrect outputs stemming from flawed datasets	Zhao et al. (2021)[10]

Understanding the taxonomy of hallucinations is crucial for designing effective detection and mitigation strategies. While intrinsic and extrinsic hallucinations define whether errors stem from internal inconsistency or external factuality, semantic and factual hallucinations differentiate between verifiable inaccuracies and incoherent yet fluent responses. Task-specific and cause-based classifications further refine how hallucination manifests in generative AI models. In the next section, we explore state-of-the-art techniques for hallucination detection, analyzing methods that aim to identify and quantify these errors in large-scale language models.

2. Detection Techniques

Detecting hallucination in generative text models is a critical challenge in ensuring the reliability of large language models (LLMs) and foundation models. Various approaches have been developed to identify hallucinations, leveraging methods such as uncertainty estimation, retrieval-augmented generation (RAG), self-consistency checks, and internal state monitoring. This section explores state-of-the-art detection techniques, categorizing them based on their underlying principles and effectiveness in identifying different types of hallucination.

 doi:10.20944/preprints202504.1236.v1

2.1. Uncertainty Estimation

One of the most widely used techniques for hallucination detection is uncertainty estimation [30], which quantifies how confident a model is in its generated outputs. Models with high uncertainty in their predictions are more likely to hallucinate. Ji et al. (2023) [3] highlight that hallucination often correlates with high entropy in the output probability distribution where multiple plausible words or phrases compete for selection. Bayesian approaches, such as Monte Carlo Dropout [31] and Bayesian Neural Networks [11,32], have been employed to measure uncertainty in generative outputs. Xie et al. (2024) [33] propose using confidence calibration methods to detect hallucination in summarization tasks, showing that hallucinated statements often have lower calibration scores compared to factual outputs [34]. Similarly [35] analyze how LLMs encode epistemic uncertainty, suggesting that lower-probability completions tend to exhibit higher hallucination rates [36].

Beyond probabilistic methods, ensemble-based uncertainty estimation has been explored, where multiple model outputs are compared to identify inconsistencies. Huang et al. (2025) [22] introduce a multi-head decoding approach that generates multiple outputs for the same prompt and measures variance amongst responses, with larger divergence indicating potential hallucination [37]. This technique has proven particularly effective in medical NLP applications [38], where detecting uncertainty is critical for decision support.

2.2. Retrieval-Augmented Generation (RAG) and External Fact Verification

A promising strategy for hallucination detection is retrieval-augmented generation (RAG), where models retrieve relevant knowledge from an external database before generating responses. Lewis et al. (2020) [39] first introduced RAG as a mechanism to improve factual consistency in LLMs by conditioning the generation process on retrieved facts [40]. By comparing generated content with retrieved references, hallucinated outputs can be identified when they diverge significantly from verifiable sources. Shuster et al. (2022) [9] propose a fact-checking mechanism where the generated text is cross-checked against an external knowledge base such as Wikipedia or Wikidata [41]. If key factual claims do not match the retrieved references, the output is flagged as a potential hallucination. Yu et al. (2023) [42] extend this approach by incorporating Neural Fact-Checking Models, which automatically classify statements as supported, refuted, or unverifiable based on retrieval evidence.

In dialogue systems, Zhang et al. (2023) [1] integrate retrieval-based validation, where chatbot responses are compared to a reference corpus. This approach has been shown to significantly reduce extrinsic hallucinations by grounding generated responses into factual data. However, RAG-based methods require high-quality retrieval sources, and their effectiveness diminishes when dealing with novel or emerging topics where no prior reference exists [28].

2.3. Self-Consistency Checks

Self-consistency methods involve generating multiple responses to the same prompt and analyzing their consistency. Wang et al. (2023) [43] propose the majority-vote self-consistency technique, where an LLM generates multiple completions for a given query. If the outputs show significant variance, it indicates possible hallucination. This technique has been applied in multi-step reasoning tasks [44], where logical consistency is critical. In addition to voting-based self-consistency, Mitchell et al. (2023) introduce contrastive consistency evaluation, where a model is prompted with slightly reworded versions of the same query. If responses vary substantially between prompts, the generated information is likely unreliable. This approach has been shown to be effective in legal and scientific text generation, where small linguistic variations should not alter factual correctness [14].

Azaria and Mitchell (2023) [14] further demonstrate that self-consistency techniques can be combined with uncertainty estimation, where high variance among generated responses correlates strongly with high uncertainty scores. This hybrid approach has been successfully tested in detecting hallucination in summarization models, particularly in cases where models struggle to paraphrase input text faithfully.

 doi:10.20944/preprints202504.1236.v1

2.4. Internal State Monitoring

Recent works suggest that hallucination can be detected by analyzing internal activations of LLMs rather than relying solely on output-level evaluations. Azaria and Mitchell (2023) [14] show that LLMs encode implicit signals of truthfulness in their hidden states. By training a classifier on activation patterns in the transformer layers, they achieve 71-83% accuracy in detecting whether a generated statement is true or false. This finding suggests that LLMs may possess an internal notion of correctness even when they produce hallucinated outputs. Building on this idea, Burns et al. (2023) [28] propose activation clustering where neural activations corresponding to factually correct statements cluster differently from those associated with hallucinated outputs. This technique has been used to enhance debiasing mechanisms in generative AI, particularly in medical text generation [38].

Another novel approach, truthfulness probing, is where activations are compared against ground-truth token embeddings to assess factual consistency. Zhao et al. (2021) [10] implement this method in multimodal hallucination detection, demonstrating that vision-language models encode distinct neural patterns for correct and hallucinated captions.

2.5. Benchmarking and Evaluation of Detection Methods

Several benchmarks have been introduced to evaluate hallucination detection techniques. The Hallucination Evaluation for Large Language Models (HALL-E) benchmark [9] provides a standardized dataset for assessing factuality in generated text. Similarly, the Entity-Relationship-Based Hallucination Benchmark (ERBench) [45] measures factual consistency by comparing generated text to structured knowledge graphs.

Other datasets including FEVER [46] and TruthfulQA [6,19,47] are widely used for hallucination detection in question answering tasks, where models are tested on their ability to generate factually accurate responses. Azaria and Mitchell (2023) suggest that dataset quality significantly impacts hallucination detection performance, highlighting the need for real-world, domain-specific evaluation datasets. See Table 2.

Table 2. Comparison of Hallucination Detection Techniques.

Method	Principle	Strengths	Limitations	Key References
Uncertainty Estimation	Measures confidence in model outputs	Detects low-confidence hallucinations	Less effective for confidently incorrect statements	Ji et al. (2023)[3], Liu et al. (2023)[45]
Retrieval-Augmented Generation (RAG)	Compares output against retrieved facts	High accuracy in factual consistency	Requires high-quality external sources	Lewis et al. (2020)[39], Yu et al. (2023)[48]
Self-Consistency Checks	Compares multiple generated outputs	Detects variance-based hallucinations	Computationally expensive	Wang et al. (2023) [43], Kojima et al. (2022)[44]
Internal State Monitoring	Analyzes hidden activations of LLMs	Directly probes model knowledge	Requires model access	Azaria & Mitchell (2023)[14], Burns et al. (2023)[28]

Hallucination detection remains an evolving field, with techniques ranging from uncertainty-based probability estimation to direct neural activation analysis. While retrieval-augmented methods provide strong factual grounding, self-consistency and internal state monitoring offer promising advancements in detecting hallucinations independent of external sources. Future work should focus on developing real-time hallucination detection mechanisms, improving benchmarks for domain-specific applications, and integrating multiple detection techniques for enhanced reliability. In the next

 doi:10.20944/preprints202504.1236.v1

section, we explore mitigation strategies, focusing on how generative models can be refined to reduce hallucination occurrence.

3. Mitigation Strategies

Mitigating hallucinations in large language models (LLMs) and foundation models is a crucial step towards improving their reliability, especially in high-stakes applications such as healthcare, legal reasoning, and scientific research. Various techniques have been proposed to reduce the occurrence of hallucinations, ranging from model fine-tuning and reinforcement learning to retrieval-based augmentation and adversarial training. This section explores the most effective strategies highlighting their strengths and limitations.

3.1. Fine-Tuning and Reinforcement Learning from Human Feedback (RLHF)

Fine-tuning remains one of the most widely used approaches for reducing hallucinations in generative models. By exposing the model to higher-quality, well-annotated datasets, fine-tuning improves factual grounding and response consistency. Ouyang et al. (2022) demonstrate that fine-tuning models using domain-specific datasets significantly reduces factual hallucinations in tasks such as medical question answering. Similarly, Burns et al. (2023) highlight that supervised fine-tuning on curated knowledge sources enhances the ability of the model to generate factually correct responses.

Reinforcement learning from human feedback (RLHF) has gained prominence as an advanced fine-tuning strategy, helping models align better with human expectations. Christiano et al. (2017) first introduced RLHF to train models that optimize for human preference scores, which was later adopted in OpenAI's GPT-4 (Ouyang et al., 2022). By using a reward model trained on human-labeled responses, RLHF discourages the generation of hallucinated or misleading content. Despite its success, RLHF has several limitations. Bai et al. (2022) [49] point out that reward modeling can introduce biases, reinforcing overfitting to subjective human preferences rather than objective factual correctness. Furthermore, Zhao et al. (2023) argue that RLHF can lead to "overcorrection," where the model avoids generating uncertain but factually correct responses due to conservative reward functions. While RLHF improves response alignment, it can lead to overly conservative outputs, limiting model creativity and diversity. An alternative approach is self-supervised learning (SSL) for hallucination reduction, where models can autonomously refine their outputs by learning from contrastive loss objectives. Recent studies in contrastive learning for NLP, such as SimCSE (Gao et al., 2021) and Contrastive Knowledge Distillation (Meng et al., 2022), demonstrate how models can enhance factual accuracy without requiring human labels. By aligning generated outputs with known factual embeddings, contrastive methods improve robustness against hallucination.

3.2. Retrieval-Augmented Generation (RAG) and External Knowledge Integration

One of the most effective strategies to mitigate extrinsic hallucination is retrieval-augmented generation (RAG), where models retrieve relevant facts from an external knowledge base before generating responses (Lewis et al., 2020). This approach grounds LLM outputs in real-world references significantly reducing misinformation. Shuster et al. (2022) evaluate RAG-based models in open-domain question answering, demonstrating that retrieving supporting documents before response generation reduces factual hallucinations by over 40%. Similarly, Yu et al. (2023) integrate retrieval-based verification in chatbot systems, ensuring that model outputs remain factually consistent with external sources.

While RAG is effective in grounding generative models, it has certain drawbacks. Burns et al. (2023) highlight that RAG is only as reliable as the external knowledge source; if the retrieved data is outdated, biased, or incorrect, the model may still generate hallucinations. Additionally, retrieval mechanisms introduce latency, making them less suitable for real-time applications such as conversational AI (Zhang et al., 2023). One of the limitations of RAG is the dependence on static knowledge bases, which can lead to outdated or incomplete retrieval. Recent work explores adaptive retrieval strategies, where self-supervised contrastive learning refines the retrieval process dynamically.

 doi:10.20944/preprints202504.1236.v1

By training models to distinguish between hallucinated and factual retrievals, contrastive fine-tuning enhances RAG's ability to filter reliable sources.

3.3. Prompt Engineering and Instruction Tuning

Prompt engineering has emerged as a lightweight and efficient method to guide models toward accurate responses. Brown et al. (2020) demonstrate that few-shot and chain-of-thought prompting significantly reduces hallucination by structuring the reasoning process more effectively. Kojima et al. (2022) [44] extend this idea with zero-shot chain-of-thought prompting to improve logical consistency in LLM-generated reasoning. Instruction tuning is a related approach that modifies the training process so that models better adhere to task-specific constraints. Mishra et al. (2022) propose instruction-based fine-tuning, showing that explicitly framing instructions for factual accuracy improves model robustness. While prompt engineering is computationally inexpensive, it does not fundamentally alter the knowledge base of the model. Burns et al. (2023) caution that prompt-based methods only work as temporary fixes, requiring frequent updates as hallucination patterns evolve.

3.4. Adversarial Training and Contrastive Learning

Adversarial training is an effective method for enhancing robustness against hallucination. This approach exposes models to adversarially crafted examples to teach differentiation between plausible but incorrect information and grounded, factual content [50,51]. Hendrycks et al. (2021) employ contrastive adversarial training, where models are trained with factually incorrect distractor samples to improve factual precision. Contrastive learning techniques such as contrast-consistent search (CCS) (Burns et al., 2023) train models by comparing hallucinated outputs to correct responses, refining their internal confidence mechanisms. Zhang et al. (2023) apply contrastive learning to dialogue systems to reduce the frequency of semantically inconsistent responses.

Despite its advantages, adversarial training is computationally expensive and requires large-scale annotated adversarial datasets. Zhao et al. (2023) note that generating adversarial samples for fine-tuning is labor-intensive, limiting the scalability of this approach.

3.5. Hybrid Approaches and Multimodal Mitigation Strategies

Recent research has explored hybrid methods, combining multiple mitigation techniques for greater effectiveness. Liu et al. (2023) propose a two-stage hybrid model integrating uncertainty-aware filtering with retrieval-based knowledge injection, achieving a 30% reduction in hallucinated responses compared to standard fine-tuning.

In multimodal models, Bai et al. (2023) introduce a cross-modal grounding technique, where vision-language models cross-reference textual descriptions with visual evidence to reduce hallucination in image captions. This method has proven effective in mitigating hallucination in image-generation models such as DALL·E and Stable Diffusion. Hybrid approaches show promising results, but also increase model complexity. Burns et al. (2023) note that combining multiple mitigation techniques requires fine-grained calibration, as overcorrection in one area may degrade model fluency. See Table 3.

 doi:10.20944/preprints202504.1236.v1

Table 3. Comprehensive Overview of Hallucination Evaluation Metrics and Mitigation Strategies Across Different Tasks.

Category	Task	Principle	Strengths	Limitations & Key References
Evaluation Metrics	**Dialogue**	Measures factual consistency in conversational AI	Detects inconsistencies	Sensitive to open-ended dialogue, may require human validation [9,52–54]
	Summarization	Evaluates faithfulness of generated summaries	Captures factual errors	Struggles with abstractive models that paraphrase well [23,55–57]
	Translation	Checks alignment between source and translated output	Identifies extrinsic hallucination	Limited for low-resource languages [11,42,58,59]
	Data-to-Text	Assesses alignment of structured data with text output	Task-specific, improves reliability	May not generalize well across datasets [60–62]
	Multimodal	Validates consistency between image and generated text	Reduces visual-text mismatches	Requires strong vision-language benchmarks [14,15,26]
	RAG	Compares generated text with retrieved knowledge	Enhances factual accuracy	Relies on knowledge quality and retrieval effectiveness [9, 39,40]
Mitigation Strategies	**Fine-Tuning & RLHF**	Trains on curated datasets, optimizes via reward models	Reduces factual hallucinations, aligns with human intent	May reinforce biases, risk of overfitting [16,28]
	Retrieval-Augmented Generation (RAG)	Uses external databases for fact verification	Enhances factuality	Dependent on retrieval source quality [39,48]
	Prompt Engineering & Instruction Tuning	Guides models using structured prompts	Lightweight, computationally cheap	Temporary fix, requires frequent updates [44,63]
	Adversarial Training	Exposes model to adversarial examples	Improves model robustness	Computationally expensive, requires large adversarial datasets [50,51]
	Hybrid & Multimodal Approaches	Combines multiple mitigation techniques	Increases adaptability	Complex to implement, needs careful balancing [15,64]

3.6. Contrastive Learning for Hallucination Mitigation

Contrastive learning has emerged as an effective technique for reducing hallucination by enabling models to differentiate between factually correct and hallucinated outputs. Unlike traditional supervised learning, contrastive learning leverages both positive (factually correct) and negative (hallucinated) examples to train models to recognize semantic inconsistencies and factual deviations in generated text.

Principles of Contrastive Learning for Hallucination Mitigation

Contrastive learning operates by maximizing the similarity between correct outputs and reference texts while minimizing the similarity between incorrect (hallucinated) outputs and factual knowledge. This is achieved using contrastive loss functions such as InfoNCE (Information Noise Contrastive

 doi:10.20944/preprints202504.1236.v1

Estimation) loss and Triplet loss which enforce factual alignment. In this paradigm, a model processes three types of inputs:

Anchor: The original input (e.g., a document or question).

Positive Sample: A factually correct reference output.

Negative Sample: A hallucinated or incorrect output.

By minimizing the distance between the anchor and the positive sample while increasing the distance between the anchor and the negative sample, the model learns to distinguish hallucinated outputs from factually consistent responses. Contrastive learning has been successfully applied across multiple NLP tasks to mitigate hallucination in generative models including:

Hallucination Detection in Summarization: Recent studies have integrated contrastive learning into abstractive summarization models to ensure that generated summaries remain faithful to the information from source documents. For example, contrastive fact verification (CFV) [59] fine-tunes large language models (LLMs) by training them to detect inconsistent or extrinsic information in generated summaries

Factually Grounded Dialogue Systems: In conversational AI, contrastive learning is used to enhance retrieval-augmented generation (RAG) by reinforcing responses that are factually grounded while penalizing hallucinated dialogue continuations [65]. Models like Fact-CheckGPT employ contrastive loss to ensure that responses remain coherent and supported by retrieved knowledge

Multimodal Hallucination Mitigation: Vision-language models (VLMs) such as BLIP-2 and PaLI use contrastive learning to align textual and visual features, reducing object-based hallucination where descriptions contain misclassified or non-existent objects [15].

Contrastive learning provides several advantages over conventional fine-tuning and reinforcement learning methods: Self-Supervised Learning Capability: Unlike RLHF, which depends on extensive human feedback, contrastive learning can be applied in a self-supervised manner, reducing the cost of annotation.

Better Generalization: Because contrastive learning enables models to compare a diverse range of factual and hallucinated outputs, it improves robustness across different tasks, including summarization, translation, and question-answering.

Scalability: Unlike retrieval-augmented generation (RAG), which relies on external knowledge sources that can become outdated, contrastive learning operates within the model's latent space, making it adaptable to new domains without requiring frequent updates to knowledge bases.

While contrastive learning has shown promise in hallucination mitigation, several open challenges remain. Future work should focus on integrating contrastive learning with adaptive fine-tuning techniques, allowing models to continuously refine their factual reasoning capabilities. Additionally, multi-view contrastive learning [10], where models learn to contrast hallucinated text from multiple perspectives, including logical consistency, factual grounding, and contextual alignment, could be explored. As generative models become increasingly complex, incorporating contrastive loss functions into hallucination detection pipelines will be crucial in ensuring more trustworthy and reliable AI-generated content.

4. Evaluation Metrics and Benchmarks

Evaluating hallucination in generative models is essential for assessing their reliability, factuality, and coherence across different applications. Given the increasing deployment of large language models (LLMs) in critical domains such as healthcare, law, finance, and scientific research, it is imperative to establish rigorous evaluation methodologies. Various quantitative and qualitative metrics have been proposed to measure the extent, severity, and nature of hallucination, providing insights into how well generative models align with human expectations and factual accuracy. Beyond evaluation metrics, benchmarks and datasets serve as standardized testing environments to ensure consistent comparison across different models and hallucination mitigation techniques. This section provides a

 doi:10.20944/preprints202504.1236.v1

detailed overview of key hallucination evaluation metrics and benchmarks, highlighting their strengths, limitations, and suitability for different NLP tasks.

4.1. Evaluation Metrics for Hallucination

Hallucination detection and quantification rely on a combination of factual consistency measures, semantic coherence metrics, uncertainty-based evaluations, and human annotation techniques. Each class of metrics serves a different purpose, depending on whether hallucination is assessed in structured tasks (summarization, translation, retrieval-augmented generation, etc.) or in open-ended text generation (conversational AI, storytelling, multimodal generation, etc.)

4.1.1. Factual Consistency Metrics

Factual consistency metrics assess whether the generated text aligns with known facts or remains faithful to a given input source. These metrics are particularly relevant in tasks such as abstractive summarization, question answering, fact verification, and data-to-text generation, where the ability of the model to preserve factual integrity is crucial.

FactScore [66] is designed to measure the faithfulness of summaries to their original source documents by computing the semantic and factual overlap between the generation and input text. This metric is widely used in news summarization and scientific document summarization tasks.

Strengths: Captures both semantic and factual alignment, making it effective for evaluating hallucination in structured text generation.

Limitations: Struggles with implicit factual inconsistencies and complex paraphrasing, where the hallucination is not a direct contradiction but rather a distortion of meaning.

Entity-Level Fact Checking [67] involves identifying and verifying named entities, numerical values, and key factual statements using external databases. This metric is commonly employed in medical AI, legal AI, and knowledge-intensive NLP tasks.

Strengths: Effective in detecting misinformation and verifying named entities in structured text.

Limitations: Performs poorly in open-ended text generation, where hallucinations may involve abstract reasoning rather than entity-level distortions.

FEVER Score The Fact Extraction and Verification (FEVER) Score [46] evaluates the factual accuracy of model-generated claims by retrieving evidence from Wikipedia and classifying statements as "Supported", "Refuted", or "Not Enough Information".

Strengths: Provides a structured framework for fact verification, making it a benchmark standard in fact-checking tasks.

Limitations: Limited to Wikipedia-based claims, making it less effective for domain-specific applications such as biomedical text generation.

4.1.2. Semantic Coherence and Fluency Metrics

While factual consistency focuses on objective truthfulness, semantic coherence and fluency metrics assess whether the generated text maintains logical consistency and readability. These metrics are especially useful in dialogue systems, open-ended storytelling, and multimodal generation, where hallucination may manifest as logically inconsistent or nonsensical outputs.

BERTScore [68] measures the semantic similarity between generated text and reference text using contextual embeddings from pretrained transformer models. Unlike traditional BLEU or ROUGE scores, BERTScore captures fine-grained semantic relationships rather than surface-level word overlap.

Strengths: Effective in capturing meaning preservation and semantic relevance.

Limitations: Does not explicitly measure factual consistency, meaning a hallucinated yet semantically plausible output may still receive a high BERTScore.

BLEURT [69] is a learning-based evaluation metric that predicts human ratings for text generation quality. It improves upon traditional fluency metrics by incorporating contextual embeddings and pretrained transformer knowledge.

Strengths: Robust for open-ended generation tasks where human-like fluency is critical.

 doi:10.20944/preprints202504.1236.v1

Limitations: Less suitable for strict fact-based tasks, as it focuses more on readability than factuality.

4.1.3. Uncertainty-Based Metrics

Uncertainty-based metrics attempt to quantify the confidence of the model in the generated outputs. These techniques are especially useful in risk-sensitive domains such as medical AI, legal NLP, and financial AI, where low-confidence responses often correlate with hallucination.

Entropy-Based Confidence Score [70] calculates the probability distribution of generated tokens, flagging sentences where the model is highly uncertain about its predictions.

Strengths: Helps detect low-confidence hallucinations.

Limitations: Struggles with confident yet incorrect outputs, a common issue in LLMs trained with reinforcement learning.

Monte Carlo Dropout [71] generates multiple outputs for the same input and measures the variance in predictions to estimate uncertainty.

Strengths: Effective in assessing model confidence, particularly in question answering and medical AI.

Limitations: Computationally expensive, requiring multiple forward passes through the model.

4.1.4. Human Evaluation Metrics

Despite advancements in automated evaluation, human annotation remains the gold standard for hallucination assessment. These evaluations involve domain experts or crowd workers manually assessing model outputs for factual correctness, coherence, and logical consistency.

Expert Fact Verification In high-stakes applications like biomedical AI and legal text generation, expert fact verification [38] is used to validate AI-generated content against authoritative sources.

Strengths: Provides the highest level of reliability in domain-specific tasks.

Limitations: Expensive, time-consuming, and not scalable for large-scale model evaluations.

Likert-Scale Faithfulness Ratings Likert-scale annotations [12] is a method where human annotators rate the factual faithfulness of generated text on a numerical scale (1 to 5), providing quantitative insight into the human perception of hallucination.

Strengths: Useful for qualitative feedback and comparative model evaluation.

Limitations: Subjective and prone to annotator bias.

4.2. Benchmarks and Datasets for Hallucination Evaluation

Several large-scale benchmarks have been developed to standardize hallucination assessment across different tasks. These datasets provide a consistent evaluation framework, enabling researchers to compare hallucination detection and mitigation techniques.

FEVER Benchmark [46] The Fact Extraction and Verification (FEVER) dataset is a widely used benchmark for factual consistency evaluation. It consists of 185,000 claims labeled as "Supported", "Refuted", or "Not Enough Information" with corresponding Wikipedia evidence. FEVER has been applied in hallucination detection for summarization [56] and dialogue generation [68].

Hallucination Evaluation for Large Language Models (HALL-E) [9] The HALL-E benchmark was introduced to evaluate hallucination in open-ended text generation [72]. It contains human-annotated hallucination labels for responses generated by LLMs, covering diverse domains such as medical, legal, and financial text generation.

TruthfulQA Benchmark [6] TruthfulQA tests models on their ability to generate factually correct answers by posing truth-based adversarial questions. Unlike standard question-answering datasets, TruthfulQA evaluates how well generative models resist common misconceptions and misinformation.

Entity-Relationship-Based Hallucination Benchmark (ERBench) [48] ERBench focuses on hallucination detection in structured text by evaluating how well LLMs preserve entity relationships in generated outputs. It is particularly useful in biomedical and legal AI applications, where hallucination often involves misrepresenting factual associations.

 doi:10.20944/preprints202504.1236.v1

HaluEval Benchmark [45] HaluEval is a recent large-scale benchmark designed specifically to evaluate hallucination in large language models. Unlike previous datasets that focus on narrow domains, HaluEval provides a diverse set of 35,000 examples, covering multiple NLP tasks such as question answering, knowledge-grounded dialogue, and summarization. The dataset evaluates LLM-generated responses for factual consistency, coherence, and faithfulness, making it an essential resource for researchers aiming to assess and mitigate hallucination in generative AI systems. HaluEval also incorporates human annotations, where responses are manually assessed for hallucination severity, providing a gold standard for evaluating automatic hallucination detection techniques. The dataset has been tested across state-of-the-art models, including GPT-4, LLaMA, and PaLM, allowing comparative performance analysis across different architectures. Future works utilizing HaluEval could refine current hallucination detection methodologies by integrating model confidence scores, retrieval-based verification, and adaptive fine-tuning techniques.

OpenAI's Real-World Hallucination Dataset [14] Mitchell et al. (2023) introduced a dataset containing real-world examples of hallucination from deployed AI systems. This benchmark includes hallucinated chatbot responses, mistranslations, and incorrect factual claims, serving as a practical evaluation set for real-world applications, see Table 4.

Table 4. Evaluation Metrics and Benchmarks for Hallucination Assessment.

Category	Metric / Benchmark	Application	Key References
Factual Consistency	FEVER Score, FactScore, Entity-Level Fact Checking	Summarization, Question Answering	Thorne et al. (2018)[46], Kryscinski et al. (2020)[56]
Semantic Coherence	BERTScore, BLEURT, Self-BLEU	Open-ended text generation	Zhang et al. (2020)[68], Sellam et al. (2020)[69]
Uncertainty Estimation	Entropy-Based Confidence, Prediction Variance Analysis	Long-form text, Multi-step reasoning	Jiang et al. (2023)[70], Liu et al. (2023)
Human Evaluation	Likert-Scale Ratings, Expert Verification	High-risk AI applications	Nori et al. (2023)[38], Zhang et al. (2023)[68]
Benchmarks	FEVER, HALL-E, TruthfulQA, ERBench, HaluEval	Standardized hallucination assessment	Thorne et al. (2018)[46], Shuster et al. (2022)[9], Lin et al. (2022), Yu et al. (2023)[48]

The evaluation of hallucination in generative models is a complex task requiring multiple assessment techniques. Automated metrics such as fact verification scores, semantic similarity measures, and uncertainty estimation provide scalable evaluation methods. However, human annotation remains essential for high-stakes applications where factual correctness is critical. The development of standardized benchmarks like FEVER, HALL-E, and TruthfulQA has significantly advanced comparative evaluation, but future work should focus on real-world hallucination detection in deployed AI systems. In the next section, we discuss open challenges and future directions in hallucination research.

5. Research Gaps and Future Directions

Despite significant progress in detecting and mitigating hallucinations in generative models, several open challenges remain. As AI systems continue to scale and integrate into real-world applications, addressing these challenges will be critical for ensuring the trustworthiness, interpretability, and robustness of large language models (LLMs) and foundation models. This section highlights key research gaps and proposes future directions to advance hallucination detection and mitigation.

 doi:10.20944/preprints202504.1236.v1

5.1. Lack of Standardized Hallucination Definitions and Taxonomies

One of the persistent challenges in hallucination research is the lack of a universally accepted definition and classification framework. While researchers have categorized hallucinations into intrinsic vs. extrinsic [3], factual vs. semantic [4], and model-based vs. data-induced [28], these classifications often overlap, leading to inconsistent evaluation methodologies across studies. Future research should focus on developing a standardized hallucination taxonomy that integrates multiple perspectives, ensuring a unified evaluation framework across different NLP tasks. Establishing benchmark datasets with clearly labeled hallucination types will further aid in improving cross-comparability of research outcomes.

5.2. Limitations of Existing Hallucination Detection Techniques

Most hallucination detection techniques rely on retrieval-augmented generation (RAG) [39], uncertainty estimation [70], and self-consistency checks [43]. While these approaches have demonstrated effectiveness in identifying hallucinated outputs, they exhibit several significant limitations that hinder their scalability, reliability, and generalizability across diverse tasks and domains.

One major limitation of retrieval-augmented models is their over-reliance on external knowledge sources, which can introduce inaccuracies if the retrieved documents contain incomplete, outdated, or biased information [9]. Since these models condition their responses on retrieved content, hallucinations may still persist if the retrieved evidence is irrelevant, inconsistent, or contradictory. This is particularly problematic in low-resource or rapidly evolving fields, where reliable external sources may be scarce or prone to errors. Moreover, retrieval-augmented models struggle with novel or counterfactual queries, where no pre-existing factual evidence is available, leading to either incorrect completions or high-confidence hallucinations. Another significant drawback of hallucination detection techniques is their high computational cost, especially for self-consistency decoding and contrastive reasoning methods. These techniques rely on multiple forward passes to generate different responses for the same query, which are then compared to identify inconsistencies. While this approach is effective for certain reasoning-intensive tasks, it becomes computationally prohibitive in real-time applications such as chatbots, virtual assistants, and autonomous decision-making systems [14]. Large-scale deployment of such methods in high-throughput environments remains impractical due to the increased inference latency and resource consumption, required for processing multiple generations per input.

A further challenge arises from false positives in uncertainty-based hallucination detection techniques. Confidence-based methods often flag responses with low-probability scores as potential hallucinations, even when they are factually correct but uncommon [10]. This issue is particularly pronounced in low-resource languages, domain-specific terminology, and creative text generation, where models may assign lower confidence scores to valid but less frequently occurring outputs. Consequently, hallucination detection systems may introduce an unintended trade-off: over-filtering valid responses while still failing to catch high-confidence hallucinations in cases where the model confidently generates plausible but incorrect information. To overcome these challenges, future research should focus on developing hybrid detection models that integrate uncertainty-aware signals with factual verification techniques, rather than relying solely on confidence estimation or retrieval grounding. Additionally, contrastive learning and reinforcement learning techniques can be leveraged to enhance hallucination detection by encouraging models to differentiate between factually grounded outputs and plausible-sounding hallucinations. By combining multiple detection strategies, it may be possible to reduce computational overhead, minimize false positives, and improve detection accuracy in diverse NLP tasks.

5.3. Challenges in Hallucination Mitigation Strategies

Existing mitigation strategies, such as fine-tuning with high-quality data [16] and reinforcement learning from human feedback (RLHF) [49], have demonstrated effectiveness in reducing hallucina-

 doi:10.20944/preprints202504.1236.v1

tion rates across various generative tasks. Despite these advancements, several challenges remain unresolved, limiting the scalability, adaptability, and generalizability of these approaches.

One of the key challenges is overcorrection bias introduced by RLHF. While reinforcement learning helps align model outputs with human expectations, it often results in models that avoid generating uncertain responses altogether [15]. This leads to "safe but uninformative" outputs, where the model prioritizes cautiousness over diversity, ultimately affecting response richness and utility in applications such as creative writing, open-domain dialogue, and exploratory reasoning. The trade-off between suppressing hallucinations and maintaining response diversity remains unresolved, requiring further optimization to balance safety, informativeness, and factual accuracy. Another major challenge lies in the scalability of retrieval-augmented generation (RAG). While RAG-based models improve factual consistency by retrieving relevant external knowledge before generating responses, they are constrained by retrieval bottlenecks and knowledge source limitations [48]. The effectiveness of RAG depends on the quality, coverage, and timeliness of the retrieved knowledge, which can lead to hallucinations if the retrieved content is incomplete, outdated, or contextually irrelevant. Additionally, real-time retrieval processes introduce latency issues, making these models less efficient for high-speed applications such as real-time dialogue generation, legal AI, and automated decision-making systems.

A further limitation exists in prompt engineering and instruction tuning, which require frequent manual adjustments to maintain their effectiveness [44]. While prompt-based methods have proven useful in guiding model behavior and reducing hallucination, they lack adaptability to dynamically evolving topics and domain-specific nuances. This makes them impractical for long-term deployments, as researchers and practitioners must continually refine and update prompt strategies to keep up with emerging trends, new knowledge, and evolving linguistic patterns.

To address these limitations, adaptive fine-tuning techniques have been proposed as a promising future direction. By incorporating continuous learning mechanisms, generative models can learn from real-time user feedback and self-verification signals, allowing them to gradually correct their own hallucinations over time. Additionally, self-supervised contrastive learning offers another avenue for improvement, enabling models to differentiate between factual and fabricated outputs dynamically. By integrating self-adaptive training methodologies, hybrid RAG-based verification, and uncertainty-aware optimization, future generative models may achieve a more balanced trade-off between factual reliability, response diversity, and computational efficiency.

5.4. Addressing Hallucination in Multimodal and Multilingual Models

While hallucination research has predominantly focused on text-based LLMs, multimodal models such as DALL·E, Stable Diffusion, CLIP, and GPT-4V introduce new challenges in hallucination across image, video, and speech generation [73]. These models must generate coherent and accurate outputs across multiple modalities, yet they often suffer from misalignment between visual, textual, and auditory data. The complexity of integrating different modalities amplifies the risk of hallucination making this a critical area of study. One of the most prevalent issues in vision-language models is object hallucination, where models generate descriptions that contain non-existent or misclassified objects [14]. For example, an image-captioning system may describe a landscape as containing "a river" when no such feature exists in the image. This phenomenon is particularly problematic in medical imaging, where AI-generated reports may describe non-existent abnormalities, leading to severe consequences. To mitigate object hallucination, cross-modal grounding techniques are being developed to ensure that generated text remains faithful to the visual content. Models such as BLIP-2 and PaLI attempt to improve textual alignment with visual information by incorporating retrieval-augmented vision-language learning. However, these methods still face challenges in maintaining semantic consistency across multimodal inputs.

In speech-based models, hallucination manifests as misinterpretations of spoken content, often resulting in incorrect speech-to-text transcriptions or misleading audio descriptions [12]. This issue is particularly relevant in automated subtitle generation and virtual assistant interactions, where AI-generated transcripts may fabricate words or misinterpret speaker intent. Future research should

 doi:10.20944/preprints202504.1236.v1

focus on joint multimodal fact-checking frameworks, where speech recognition models cross-verify their outputs with textual and contextual cues. Methods such as self-supervised contrastive learning for speech-text alignment have shown promise in reducing contextual drift in speech hallucinations.

Multilingual hallucination detection presents another challenge, as most hallucination detection methods are optimized for English-language LLMs, leaving low-resource languages underexplored [50]. Hallucination patterns vary significantly across linguistic contexts due to differences in grammatical structures, semantic ambiguity, and training data availability. Current models often hallucinate non-existent translations or fabricate culturally irrelevant details, particularly in machine translation systems such as Google Translate and MarianMT. Addressing multilingual hallucination requires the development of language-agnostic hallucination detection models capable of adapting across diverse linguistic environments. Multilingual retrieval-augmented generation (RAG) could help mitigate this issue by grounding model outputs in multilingual knowledge bases, ensuring factual consistency across languages.

To tackle these challenges, future research should prioritize benchmarking multimodal hallucination across various domains. Standardized datasets such as HaluEval, OpenAI's Real-World Hallucination Dataset, and multimodal fact-checking corpora should be expanded to include cross-lingual and cross-modal hallucination detection tasks. Furthermore, integrating self-supervised training techniques for cross-lingual hallucination verification could help improve the robustness and accuracy of generative models across different modalities and languages. As multimodal and multilingual AI systems become more prevalent, refining hallucination detection and mitigation strategies will be crucial for building trustworthy AI applications.

5.5. Real-World Deployment and Trustworthy AI Considerations

While hallucination detection has advanced significantly in academic settings, practical deployment in high-risk applications remains a major challenge. Nori et al. (2023) [38] highlight that hallucinations in medical AI can result in severe misinformation, potentially leading to misdiagnoses and incorrect treatment recommendations. Similarly, hallucinations in legal AI systems raise regulatory and ethical concerns, as demonstrated by Yu et al. (2023)[48], where incorrect legal interpretations generated by AI could have serious legal ramifications.

One of the key research gaps in real-world AI deployment is the development of hallucination-resistant AI systems. Future models should integrate self-regulation mechanisms capable of detecting and filtering potentially misleading outputs before presenting them to users. Techniques such as confidence-aware output filtering and automatic retrieval-based fact-checking could help mitigate the risks associated with high-stakes AI applications, ensuring that generative models provide more trustworthy and verifiable responses. Another critical challenge is the need for human-AI collaboration in hallucination verification. Instead of focusing solely on eliminating hallucinations, researchers should explore "Human-AI Hybrid Models," where AI-generated content undergoes expert validation before being deployed in sensitive fields such as healthcare, law, and finance. Mitchell et al. (2023) [14] propose a collaborative framework in which AI models assist human professionals while continuously learning from human feedback, thereby improving factual reliability without sacrificing efficiency.

Beyond technical robustness, ethical and explainability concerns present additional obstacles to AI deployment in real-world applications. Future research must prioritize explainable hallucination detection, ensuring that users can understand why a model's response is flagged as a hallucination. Zhao et al. (2021) [10] emphasize that black-box AI models with unclear decision-making processes could reduce user trust, particularly in sectors where accountability and transparency are essential. Explainable AI (XAI) techniques, including model interpretability frameworks and attribution-based validation methods, could enhance user confidence and regulatory compliance in AI-assisted decision-making.

 doi:10.20944/preprints202504.1236.v1

5.6. Ethical Considerations and Explainability in Hallucination Detection

The ethical implications of hallucination in generative AI extend beyond simple factual inaccuracies, affecting user trust, decision-making, and AI accountability. As large language models (LLMs) and multimodal AI systems are integrated into critical applications such as healthcare, legal analysis, and education, ensuring transparency in hallucination detection and mitigation becomes imperative. One of the key challenges is developing explainable AI (XAI) techniques that provide clear, interpretable justifications for why an AI-generated response is considered hallucinated. Current hallucination detection methods often operate as black-box classifiers, where models flag outputs as hallucinated but fail to provide underlying justifications. To improve explainability, models should explicitly justify their outputs when hallucination is detected. One approach is retrieval-augmented reasoning, where the model displays retrieved source documents or knowledge graphs that support or contradict the generated response [9,39]. This allows users to compare AI-generated claims against verifiable external knowledge, ensuring greater transparency.

In addition to retrieval sources, generative models should incorporate step-by-step reasoning visualization, inspired by chain-of-thought (CoT) prompting [44]. Instead of only producing a final response, AI systems could break down their reasoning process, highlighting factual dependencies and uncertainty scores. For instance, a medical AI generating a patient diagnosis should present multiple supporting factors and confidence levels rather than an unverified assertion.

Another promising direction is faithfulness verification through self-checking mechanisms, where models internally validate generated outputs by cross-referencing prior factual knowledge. SelfCheckGPT [20] applies this concept by prompting the model to re-evaluate its responses multiple times and check for inconsistencies before finalizing an output. This approach mimics human-style critical reasoning and can significantly improve the interpretability of hallucination detection systems.

Ethical risks associated with generative model hallucination are particularly severe in domains that require factual precision, such as medical diagnostics, legal AI, and financial forecasting. In medical AI, hallucinated disease symptoms or misdiagnosed conditions can lead to harmful treatments [38]. Legal AI systems that hallucinate case law references could introduce false legal precedents, undermining trust in automated legal decision-making. Similarly, AI-driven financial models may generate fabricated stock trends or false economic indicators, leading to misinformed investments. To mitigate these risks, AI systems should incorporate hallucination confidence scores alongside their outputs. These confidence estimates could be presented as factual accuracy scores or probability distributions over multiple possible responses, allowing users to assess the reliability of AI-generated information. Additionally, integrating human-AI collaboration mechanisms, such as expert validation checkpoints, can ensure that high-risk hallucinations do not propagate without human review. Building trustworthy AI requires more than just improving factual accuracy—it demands greater interpretability, accountability, and ethical alignment. Future research should focus on designing explainable hallucination detection models that not only detect hallucination but also justify and correct their outputs in real time. By integrating retrieval-based justifications, step-by-step reasoning, and uncertainty-aware evaluations, AI can transition from an opaque system to a more accountable and transparent assistant.

To address these challenges, future AI systems should incorporate "trustworthiness metrics" capable of quantifying hallucination risk and model confidence levels. Developing scalable, real-time trust assessment frameworks would significantly enhance AI reliability and adoption in mission-critical environments, where hallucination-related risks must be carefully managed to ensure both safety and accountability (see Table 5).

 doi:10.20944/preprints202504.1236.v1

Table 5. Research Gaps and Future Directions in Hallucination Studies.

Research Gap	Future Direction	Key References
Lack of Standardized Taxonomy	Develop unified hallucination classification frameworks	Ji et al. (2023)[3], Rawte et al. (2023)[4] [4]
Limitations of Detection Methods	Hybrid uncertainty + fact verification models	Lewis et al. (2020)[39], Jiang et al. (2023)[70]
Challenges in Mitigation Strategies	Adaptive fine-tuning and self-supervised learning	Ouyang et al. (2022)[16], Bai et al. (2022)[49]
Hallucination in Multimodal Models	Cross-modal grounding for vision-language AI	Mitchell et al. (2023)[14], Zhang et al. (2023)[68]
Trustworthy AI and Deployment	Human-AI hybrid verification systems	Nori et al. (2023)[38], Yu et al. (2023)[48]

6. Conclusion

The rapid advancements in Large Language Models (LLMs) and foundation models have significantly transformed natural language generation, enabling applications in healthcare, law, education, and creative writing. However, hallucination, where models generate misleading, fabricated, or factually incorrect information, remains a major challenge. This survey provides a comprehensive review of hallucination detection, mitigation strategies, evaluation metrics, and research gaps, offering a structured perspective on how to enhance the reliability and trustworthiness of generative AI systems. We first established a taxonomy of hallucination, categorizing it into intrinsic vs. extrinsic, factual vs. semantic, and task-specific types, emphasizing how different AI tasks require tailored detection methods. The literature explores hallucination detection techniques, including uncertainty estimation, retrieval-augmented generation (RAG), self-consistency checks, and internal state monitoring, while highlighting both their strengths and limitations. Mitigation strategies, such as fine-tuning, reinforcement learning from human feedback (RLHF), adversarial training, and prompt engineering, were analyzed to assess their effectiveness in reducing hallucination frequency.

The survey examines evaluation metrics and benchmarks, including FEVER Score, BERTScore, TruthfulQA, and human expert verification, which play a crucial role in assessing hallucination levels across different domains. Despite progress, several research gaps remain. These include the lack of standardized hallucination taxonomies, limitations in current detection methods, hallucination in multimodal AI, and the need for trustworthy AI deployment in real-world applications.

Future work should focus on developing self-regulating AI models that can detect, explain, and correct their own hallucinations, integrating hybrid uncertainty-aware systems with factual grounding techniques. Additionally, cross-modal hallucination detection, low-resource language evaluation, and human-AI hybrid verification frameworks should be explored to enhance AI transparency and reliability. Ensuring trustworthy and explainable AI is not just a technical challenge but also an ethical imperative, as generative models become deeply integrated into decision-making processes. Advancements in hallucination detection and mitigation will be critical in shaping more reliable, transparent, and accountable AI systems in the future.

;)The things I do for you, moody

References

1. Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.
2. Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

 doi:10.20944/preprints202504.1236.v1

3. Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12):1–38, 2023.
4. Vipula Rawte, Amit Sheth, and Amitava Das. A survey of hallucination in large foundation models. *arXiv preprint arXiv:2309.05922*, 2023.
5. Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55, 2025.
6. Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*, 2021.
7. Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On faithfulness and factuality in abstractive summarization. *arXiv preprint arXiv:2005.00661*, 2020.
8. Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*, 2019.
9. Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. Retrieval augmentation reduces hallucination in conversation. *arXiv preprint arXiv:2104.07567*, 2021.
10. Zheng Zhao, Shay B Cohen, and Bonnie Webber. Reducing quantity hallucinations in abstractive summarization. *arXiv preprint arXiv:2009.13312*, 2020.
11. Nuno M Guerreiro, Elena Voita, and André FT Martins. Looking for a needle in a haystack: A comprehensive study of hallucinations in neural machine translation. *arXiv preprint arXiv:2208.05309*, 2022.
12. Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. Siren's song in the ai ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*, 2023.
13. Bhuwan Dhingra, Manaal Faruqui, Ankur Parikh, Ming-Wei Chang, Dipanjan Das, and William W Cohen. Handling divergent reference texts when evaluating table-to-text generation. *arXiv preprint arXiv:1906.01081*, 2019.
14. Amos Azaria and Tom Mitchell. The internal state of an llm knows when it's lying. *arXiv preprint arXiv:2304.13734*, 2023.
15. Haoyue Bai, Xuefeng Du, Katie Rainey, Shibin Parameswaran, and Yixuan Li. Out-of-distribution learning with human feedback. *arXiv preprint arXiv:2408.07772*, 2024.
16. Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.
17. Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. Neural text generation with unlikelihood training. *arXiv preprint arXiv:1908.04319*, 2019.
18. Sean Welleck, Jason Weston, Arthur Szlam, and Kyunghyun Cho. Dialogue natural language inference. *arXiv preprint arXiv:1811.00671*, 2018.
19. Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. Generating with confidence: Uncertainty quantification for black-box large language models. *arXiv preprint arXiv:2305.19187*, 2023.
20. Potsawee Manakul, Adian Liusie, and Mark JF Gales. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *arXiv preprint arXiv:2303.08896*, 2023.
21. Muqing Miao and Michael Kearns. Hallucination, monofacts, and miscalibration: An empirical investigation. *arXiv preprint arXiv:2502.08666*, 2025.
22. Yuheng Huang, Jiayang Song, Zhijie Wang, Shengming Zhao, Huaming Chen, Felix Juefei-Xu, and Lei Ma. Look before you leap: An exploratory study of uncertainty measurement for large language models. *arXiv preprint arXiv:2307.10236*, 2023.
23. Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. Understanding factuality in abstractive summarization with frank: A benchmark for factuality metrics. *arXiv preprint arXiv:2104.13346*, 2021.
24. Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *arXiv preprint arXiv:2302.09664*, 2023.
25. Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

 doi:10.20944/preprints202504.1236.v1

26. Shiliang Sun, Zhilin Lin, and Xuhan Wu. Hallucinations of large multimodal models: Problem and countermeasures. *Information Fusion*, page 102970, 2025.
27. Nick McKenna, Tianyi Li, Liang Cheng, Mohammad Javad Hosseini, Mark Johnson, and Mark Steedman. Sources of hallucination by large language models on inference tasks. *arXiv preprint arXiv:2305.14552*, 2023.
28. Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. *arXiv preprint arXiv:2212.03827*, 2022.
29. Tianhang Zhang, Lin Qiu, Qipeng Guo, Cheng Deng, Yue Zhang, Zheng Zhang, Chenghu Zhou, Xinbing Wang, and Luoyi Fu. Enhancing uncertainty-based hallucination detection with stronger focus. *arXiv preprint arXiv:2311.13230*, 2023.
30. Andrey Malinin and Mark Gales. Uncertainty estimation in autoregressive structured prediction. *arXiv preprint arXiv:2002.07650*, 2020.
31. Gabriel Yanci Arteaga. Hallucination detection in llms: Using bayesian neural network ensembling, 2024.
32. Yijun Xiao and William Yang Wang. On hallucination and predictive uncertainty in conditional language generation. *arXiv preprint arXiv:2103.15025*, 2021.
33. Liangru Xie, Hui Liu, Jingying Zeng, Xianfeng Tang, Yan Han, Chen Luo, Jing Huang, Zhen Li, Suhang Wang, and Qi He. A survey of calibration process for black-box llms. *arXiv preprint arXiv:2412.12767*, 2024.
34. Kavita Ganesan. Rouge 2.0: Updated and improved measures for evaluation of summarization tasks. *arXiv preprint arXiv:1803.01937*, 2018.
35. Charlotte Nicks, Eric Mitchell, Rafael Rafailov, Archit Sharma, Christopher D Manning, Chelsea Finn, and Stefano Ermon. Language model detectors are easily optimized against. In *The twelfth international conference on learning representations*, 2023.
36. Kaitlyn Zhou, Dan Jurafsky, and Tatsunori Hashimoto. Navigating the grey area: How expressions of uncertainty and overconfidence affect language models. *arXiv preprint arXiv:2302.13439*, 2023.
37. Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Wen-tau Yih. Trusting your evidence: Hallucinate less with context-aware decoding. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 783–791, 2024.
38. Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, et al. Can generalist foundation models outcompete special-purpose tuning? case study in medicine. *arXiv preprint arXiv:2311.16452*, 2023.
39. Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474, 2020.
40. Jie Ren, Jiaming Luo, Yao Zhao, Kundan Krishna, Mohammad Saleh, Balaji Lakshminarayanan, and Peter J Liu. Out-of-distribution detection and selective generation for conditional language models. *arXiv preprint arXiv:2209.15558*, 2022.
41. Roi Cohen, May Hamri, Mor Geva, and Amir Globerson. Lm vs lm: Detecting factual errors via cross examination. *arXiv preprint arXiv:2305.13281*, 2023.
42. Huiling Tu, Shuo Yu, Vidya Saikrishna, Feng Xia, and Karin Verspoor. Deep outdated fact detection in knowledge graphs. In *2023 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 1443–1452. IEEE, 2023.
43. Fen Wang, Bomiao Wang, Xueli Shu, Zhen Liu, Zekai Shao, Chao Liu, and Siming Chen. Chartinsighter: An approach for mitigating hallucination in time-series chart summary generation with a benchmark dataset. *arXiv preprint arXiv:2501.09349*, 2025.
44. Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022.
45. Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. Halueval: A large-scale hallucination evaluation benchmark for large language models. *arXiv preprint arXiv:2305.11747*, 2023.
46. James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. Fever: a large-scale dataset for fact extraction and verification. *arXiv preprint arXiv:1803.05355*, 2018.
47. Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*, 2017.
48. Wenhao Yu, Hongming Zhang, Xiaoman Pan, Kaixin Ma, Hongwei Wang, and Dong Yu. Chain-of-note: Enhancing robustness in retrieval-augmented language models. *arXiv preprint arXiv:2311.09210*, 2023.

 doi:10.20944/preprints202504.1236.v1

49. Haoyue Bai, Gregory Canal, Xuefeng Du, Jeongyeol Kwon, Robert D Nowak, and Yixuan Li. Feed two birds with one scone: Exploiting wild data for both out-of-distribution generalization and detection. In *International Conference on Machine Learning*, pages 1454–1471. PMLR, 2023.
50. Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.
51. Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.
52. Or Honovich, Leshem Choshen, Roee Aharoni, Ella Neeman, Idan Szpektor, and Omri Abend. Evaluating factual consistency in knowledge-grounded dialogues via question generation and question answering. *arXiv preprint arXiv:2104.08202*, 2021.
53. Nouha Dziri, Andrea Madotto, Osmar Zaïane, and Avishek Joey Bose. Neural path hunter: Reducing hallucination in dialogue systems via path grounding. *arXiv preprint arXiv:2104.08455*, 2021.
54. Prakhar Gupta, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. Dialfact: A benchmark for fact-checking in dialogue. *arXiv preprint arXiv:2110.08222*, 2021.
55. Tobias Falke, Leonardo FR Ribeiro, Prasetya Ajie Utama, Ido Dagan, and Iryna Gurevych. Ranking generated summaries by correctness: An interesting but challenging application for natural language inference. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 2214–2220, 2019.
56. Wojciech Kryściński, Bryan McCann, Caiming Xiong, and Richard Socher. Evaluating the factual consistency of abstractive text summarization. *arXiv preprint arXiv:1910.12840*, 2019.
57. Saadia Gabriel, Asli Celikyilmaz, Rahul Jha, Yejin Choi, and Jianfeng Gao. Go figure: A meta evaluation of factuality in summarization. *arXiv preprint arXiv:2010.12834*, 2020.
58. Yang Feng, Wanying Xie, Shuhao Gu, Chenze Shao, Wen Zhang, Zhengxin Yang, and Dong Yu. Modeling fluency and faithfulness for diverse neural machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 59–66, 2020.
59. Katherine Lee, Orhan Firat, Ashish Agarwal, Clara Fannjiang, and David Sussillo. Hallucinations in neural machine translation. 2018.
60. Clément Rebuffel, Marco Roberti, Laure Soulier, Geoffrey Scoutheeten, Rossella Cancelliere, and Patrick Gallinari. Controlling hallucinations at word level in data-to-text generation. *Data Mining and Knowledge Discovery*, pages 1–37, 2022.
61. Sam Wiseman, Stuart M Shieber, and Alexander M Rush. Challenges in data-to-document generation. *arXiv preprint arXiv:1707.08052*, 2017.
62. Ondřej Dušek and Zdeněk Kasner. Evaluating semantic accuracy of data-to-text generation with natural language inference. *arXiv preprint arXiv:2011.10819*, 2020.
63. Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
64. Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14375–14385, 2024.
65. I Chern, Steffi Chern, Shiqi Chen, Weizhe Yuan, Kehua Feng, Chunting Zhou, Junxian He, Graham Neubig, Pengfei Liu, et al. Factool: Factuality detection in generative ai–a tool augmented framework for multi-task and multi-domain scenarios. *arXiv preprint arXiv:2307.13528*, 2023.
66. Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. *arXiv preprint arXiv:2305.14251*, 2023.
67. Ran Tian, Shashi Narayan, Thibault Sellam, and Ankur P Parikh. Sticking to the facts: confident decoding for faithful data-to-text generation (2019). *arXiv preprint arXiv:1910.08684*, 2019.
68. Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.
69. Thibault Sellam, Dipanjan Das, and Ankur P Parikh. Bleurt: Learning robust metrics for text generation. *arXiv preprint arXiv:2004.04696*, 2020.
70. Ping Jiang, Xiaoheng Deng, Shaohua Wan, Huamei Qi, and Shichao Zhang. Confidence-enhanced mutual knowledge for uncertain segmentation. *IEEE Transactions on Intelligent Transportation Systems*, 25(1):725–737, 2023.

 doi:10.20944/preprints202504.1236.v1

71. Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016.
72. Nayeon Lee, Wei Ping, Peng Xu, Mostofa Patwary, Pascale N Fung, Mohammad Shoeybi, and Bryan Catanzaro. Factuality enhanced language models for open-ended text generation. *Advances in Neural Information Processing Systems*, 35:34586–34599, 2022.
73. Yingbo Zhang, Shumin Ren, Jiao Wang, Chaoying Zhan, Mengqiao He, Xingyun Liu, Rongrong Wu, Jing Zhao, Cong Wu, Chuanzhu Fan, et al. Expertise or hallucination? a comprehensive evaluation of chatgpt's aptitude in clinical genetics. *IEEE Transactions on Big Data*, 2025.