IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

HALLUCINATION IN LARGE LANGUAGE MODELS: A COMPREHENSIVE SURVEY OF TYPES, DETECTION METHODS, MITIGATION STRATEGIES, AND FUTURE DIRECTIONS

1. Jiya Patel
Student, Computer studies and Emerging Technology,
TransStadia University, Ahmedabad

2. Manthan Khopkar
Assistant Peofessor, Computer studies and Emerging Technology,
TransStadia University, Ahmedabad

Abstract: Although they may produce fluent text, large language models (LLMs) like GPT-4, LLaMA, and Gemini are prone to hallucinations, which result in coherent-seeming claims that are untrue or unsupported by evidence [1][2]. Hallucinations are dangerous and damage trust (e.g., medical or legal disinformation [3]). This survey thoroughly examines hallucinations in LLMs, providing comprehensive taxonomy of hallucination kinds, underlying causes, methods for identification, and strategies for mitigation. We highlight important results from recent benchmarks (e.g., TruthfulQA, HaluEval, HalluLens) and compare current efforts and classify solutions. An LLM-specific taxonomy (factual vs. fidelity, intrinsic vs. extrinsic, domain and granularity levels) and a critical examination of detection/mitigation techniques (retrieval-augmentation, fine-tuning, prompting) are among our contributions. While open models continue to lag behind GPT-4 on benchmarks, we find that retrieval-augmented and self-consistency approaches help reduce hallucinations [4][5]. We wrap off by discussing future prospects including neurosymbolic grounding and automated fact-checking pipelines, as well as problems like benchmark standardization and multilingual issues.

Index Terms - Large Language Models, Hallucination, Natural Language Generation, Factual Accuracy, Retrieval-Augmented Generation, Transformer Models, Misinformation, Fact Verification, RLHF, Chain-of-Thought Prompting, TruthfulQA, HaluEval, Self-Consistency, Knowledge Grounding, Trustworthy AI, Text Generation, Deep Learning, GPT, LLaMA, Generative AI

I. INTRODUCTION

With their remarkable generation capabilities, large language models (LLMs), neural models trained on massive text corpora, such as GPT (OpenAI), LLaMA (Meta), Mistral, Gemini (Google), and Claude (Anthropic), have transformed natural language processing (NLP). A crucial issue is hallucination, which occurs when the model produces claims that are confidently presented but are either untrue or not supported by the input, even though it produces fluent, contextually coherent prose [1][2]. In reality, hallucinations can have detrimental effects. For instance, making up medical advice could put patients in peril, giving false legal information could mislead lawyers, and code generating hallucinations could introduce defects [3][1]. Therefore, it is essential for reliable AI to guarantee the faithfulness and factual consistency of LLM outputs.

The first thorough survey on LLM hallucinations at the undergraduate level is presented in this publication. With specific examples, we offer an LLM-focused taxonomy of hallucinations that distinguishes between factual and fidelity types, intrinsic and extrinsic, domain-specific and open-domain, and granularity levels. We examine the underlying causes, including inference factors (decoding sampling, model overconfidence) and training data problems (noise, out-of-date information). Next, we examine detection techniques, classifying them into retrieval-augmented verification, knowledge-based fact-checking, internal consistency checks, uncertainty estimation, and human-evaluation benchmarks (e.g. TruthfulQA, HaluEval, FactScore). Retrieval-Augmented Generation (RAG), reinforcement learning from human feedback (RLHF), chain-of-thought prompting, grounding with external tools or citations, Constitutional AI, and prompt engineering are among the mitigation techniques that are reviewed. We also cover outstanding issues and future

research paths, and we use benchmark data to evaluate hallucination behavior across major LLMs.

The remainder of the paper is organized as follows: The history of LLM and earlier research on hallucination surveys are reviewed in Section 2. The taxonomy and forms of hallucinations are defined in Section 3. Root causes are examined in Section 4. Detection techniques are discussed in Section 5. Strategies for mitigation are covered in Section 6. Hallucinations are compared amongst LLMs in Section 7. Future directions and problems are covered in Sections 8 and 9. In Section 10, we wrap up. We base our presentation on current literature throughout by using benchmark findings and recent studies [6][7][4].

II. Background & Related Work

Evolution of LLMs. Large-scale pretrained language models made possible by the Transformer architecture (Vaswani et al., 2017) sparked a revolution in sequence models, which started with recurrent neural networks and LSTMs in the 1990s. Important turning points include the emergent abilities demonstrated by GPT-2 (Radford et al., 2019) and GPT-3 (Brown et al., 2020), which were followed by instruction-tuned models such as InstructGPT (Ouyang et al., 2022) and ChatGPT (OpenAI, 2022). While Meta's LLaMA-2 (Touvron et al., 2023) and LLaMA-3 (2024) provided open-weight high-performance models, GPT-4 (OpenAI, 2023) introduced multimodal input in 2023. Google Gemini (2024) and Anthropic's Claude series (2024–2025) are examples of more recent arrivals [8]. These developments have made strong chatbots and assistants possible, but because of their scope and generality, they have also made hallucinations more problematic.

Prior surveys on hallucination. Hallucinations are included in a number of recent surveys. An overview of hallucinations in natural language generation tasks is given by Ji et al. (2023) [9], who define intrinsic and extrinsic kinds from an NLG perspective (intrinsic: contradictory to source; extrinsic: beyond input). Specifically focusing on LLMs, Huang et al. (2024) reclassify hallucinations into two categories: fidelity (divergence from prompt or context) and factual (contradiction with real-world facts) [10]. Wu et al.'s Frontiers survey from 2025 uses TruthfulQA and other benchmarks to examine hallucination contributions (prompt vs. model) [4][11]. About 20% of generic questions result in hallucinated content, according to Li et al. (2023), who presented HaluEval, a sizable benchmark of human-annotated hallucinations in ChatGPT answers [6]. In a similar vein, Bang et al. (2025) introduced HalluLens, a standard for LLM QA that differentiates between intrinsic and extrinsic hallucinations [12]. Although these efforts provide useful taxonomies and datasets, there is yet no comprehensive survey with an LLM-centric focus that covers all aspects (taxonomy, detection, mitigation, and inter-model comparison). By combining taxonomy, detection/mitigation techniques, and model comparisons, our survey closes this gap.

Gaps filled by this work. In contrast to previous research, our survey provides a comprehensive yet approachable approach to final-year undergraduates. We combine current definitions and suggest a more comprehensive taxonomy that incorporates granularity and domain dimensions. For clarity, we provide comparative tables that summarize and contrast detection and mitigation techniques from various literature sources. Lastly, we compare hallucination tendencies in frontier LLMs using the most recent benchmark evaluations (up to 2026). This all-encompassing viewpoint on kinds, causes, remedies, and upcoming difficulties seeks to direct scholarly research as well as the real-world creation of reliable LLMs.

III. Taxonomy of Hallucinations

- There are several ways to classify hallucinations in LLM outputs. Here, we provide examples to define the main categories. In conclusion, hallucinations happen whenever the generated content—often subtly—deviates from reality or context. Hallucinations were first described as "content that is nonsensical or unfaithful to the source" by Jiang et al. (2022) [13]. Factual hallucinations and fidelity hallucinations are the two main kinds Huang et al. (2024) identify in LLMs [10][2]. We use these along with additional axes (level, domain, and intrinsic/extrinsic).

- **Factual Hallucination.** The model makes a claim that is at odds with what is known in the real world. For instance, an LLM may assert with assurance that "Sydney is the capital of Australia," which is untrue. Such hallucinations ignore verifiable truth and only rely on the parametric knowledge of the model [2]. In essence, this kind of discrepancy is factual. These mistakes are specifically targeted by knowledge retrieval and fact-checking detection techniques. We see that these are occasionally referred to as "unfaithful to reality." In contrast to faithfulness, factual hallucinations deviate from genuine world facts.

- **Faithfulness Hallucination.** In this case, the output contradicts or is not justified by the instructions or context. For example, it is disloyal (context-contradiction) if the prompt is a news item and the model's summary creates events that aren't discussed. Divergence from the source or instruction is the definition of fidelity illusion given by Huang et al. [10]. An additional illustration The model adds information to a translation task that the source text does not imply. Faithfulness to input is violated. Logical inconsistencies and unjustified conclusions are examples of faithfulness mistakes. Even if facts are accurate on their own, they still occur.

- **Intrinsic vs. Extrinsic Hallucination.** According to Ji et al. (2022) and Zhang et al. (2023), extrinsic hallucinations are outputs that transcend the source and have no basis in either the input or the training data, while intrinsic hallucinations are outputs that are inconsistent with the input context (contradicting source document or query) [14][15]. While extrinsic errors involve creating new, unsupported content (particularly in open-ended generation), intrinsic errors compromise context integrity (for example, missing a constraint from the prompt).

- **Closed-Domain vs. Open-Domain Hallucination.** This pertains to grounding sources and has been proposed in recent work [16]. Statements that are not based on the context and information supplied to the model in the prompt or fine-tuning context are referred to as closed-domain hallucinations. The term "open-domain hallucination" describes fabrications that cannot be confirmed by any known information (training data or external facts). In other words, open-domain errors are free-floating erroneous assertions, while closed-domain errors contradict the particular task context. For instance, asserting a fictional historical fact is an open-domain hallucination, whereas adding an unspecified statistic is a violation of closed-domain faithfulness [16].

- **Entity-Level, Relation-Level, Sentence-Level Hallucinations.** The granularity of hallucinations allows us to classify them. Entity-level hallucinations include the incorrect insertion or modification of named entities (e.g., naming the wrong person or place). Relation-level hallucinations include misleading relationships or characteristics, such as saying "X is the CEO of Y." Sentence-level hallucinations are more general fabrications, such as whole sentences or false assertions. Hallucinations are annotated at the span/entity level by recent benchmarks, such as HalluEntity [17]. An LLM biography generator might, for example, create a workplace for an individual (entity-level), ascribe a bogus award (relation-level), or include an irrelevant story (sentence-level). These levels aid in identifying the specific section of the text that is being hallucinated.

Table 1 summarizes these categories. Each dimension provides a different perspective. Importantly, the same output can exhibit multiple hallucination types (e.g. a false statement (factual) that also contradicts context (intrinsic) at the sentence level). Our taxonomy clarifies terms: **factual** vs. **faithful**, **intrinsic** vs. **extrinsic**, **closed/open domain**, and **entity/relation/sentence**. Identifying the type aids in choosing detection or mitigation strategy.

Table 1. *Summary of hallucination types.*

Type	Definition	Example
Factual	Claims contradict verifiable facts or reality[2].	“Sydney is the capital of Australia” (correct is Canberra).
Faithfulness	Output inconsistent with the prompt context or instructions[10].	Summarizing an article by adding events not mentioned in it.
Intrinsic	Inconsistent with or unsupported by the input context[14].	Ignoring a condition given in the prompt, e.g., ignoring time zone.
Extrinsic	Not supported by input context; novel content not based on known knowledge.	Inventing a technology that has no basis in the training data.
Closed-Domain	Ungrounded in the provided context/knowledge in prompt (task-specific)[16].	Answering a query with info not in the supplied documents.
Open-Domain	Ungrounded in any known knowledge base or training data[16].	Claiming a new scientific discovery that lacks evidence.
Entity-Level	Incorrect or fabricated named entities.	Naming the wrong director of a movie.
Relation-Level	Incorrect attributes/relations between entities.	Saying X is CEO of Y when X never held that position.
Sentence-Level	Entire sentence is factually unsupported or irrelevant.	Adding a made-up anecdote unrelated to the task.

（LLMs may exhibit multiple overlapping hallucination types in a single response[18].）

IV. Ease Root Causes of Hallucination

Why do LLMs hallucinate? Understanding root causes is crucial for mitigation[19]. Key factors include:

- **Training Data Issues.** The data used to train LLMs is frequently linked to hallucinations. The model may learn incorrect information due to noise, bias, or factual inaccuracies in web-crawled corpora. Due to the "knowledge cutoff" issue caused by outdated data, models are unable to provide accurate answers to current queries. Stereotypes and falsehoods can be strengthened by skewed or unbalanced datasets. The model may make inaccurate generalizations if some factual claims are uncommon or inconsistently reported. According to Zhang et al. (2023), LLMs provide information that seems credible when training data shortages result in unanswerable or confusing inquiries [20].

- **Model Architecture and Optimization.** Instead of modeling truthfulness, transformers model token probabilities. There isn't a built-in system that guarantees accuracy. The optimization goal (probability of human text) promotes output that is fluid but not always factual. Due of their lack of explicit world information, attention mechanisms may overlook long-range consistency. The model may "hallucinate" correlations as a result of overparameterization. The propensity to mix training instances might lead to false assertions.

- **Decoding Strategies.** Hallucinations are influenced by the decoding method (greedy, beam search, or sampling). While top-k sampling or high temperatures boost diversity, they run the danger of producing low-probability nonsense. Beam search has the potential to spread early errors. According to some research, conservative decoding produces safer but possibly boring language, whereas aggressive sampling frequently increases factual errors. If left unchecked, diversity-promoting decoding can result in greater hallucinations but also creative outputs.

- **Knowledge Cutoff and External World Changes.** Up until their training cutoff, off-the-shelf LLMs have static knowledge (e.g., GPT-4's knowledge might end in 2023). After that cutoff, queries concerning facts or events compel the model to make guesses. It might base responses on out-of-date or irrelevant patterns, so creating false impressions of future or altered realities. For instance, if data is not updated, inquiring about "the president of country X" following an election may result in an inaccurate estimate.

- **Overconfidence and "Sycophancy."** LLMs frequently exhibit overconfidence, producing responses that are highly likely to be incorrect. Because the training promotes confident output, people might state "facts" as if they are certain when they aren't. Additionally, some models display "yes-man" behavior, supporting user claims even when they are questionable in terms of their veracity (a phenomenon known as sycophancy). If the model attempts to satisfy the request instead of truth-checking it, this could make hallucinations worse.

- **Parameter Averaging (Interpolation).** LLMs combine data from various sources. They may combine contradictory data and interpolate in embedding space when prompted for details. This may result in statements that sound realistic yet are inaccurate.

In conclusion, a combination of incomplete data, model constraints, and generating heuristics results in hallucinations [19][20]. Addressing these underlying issues, such as enhancing data coverage, enhancing architectures or training objectives, and employing cautious decoding, is necessary to overcome them. But since there isn't a single solution, hybrid approaches—like those in the sections that follow—are required.

V. Detection Methods

Detecting hallucinations in LLM outputs is vital for safe deployment. Existing approaches include:

- **Knowledge-Based Fact-Checking.** Check LLM claims against outside sources of information (such as databases, search engines, or Wikipedia). For instance, queries can be sent to a knowledge base or search API, and the results are compared against the model's output. Inconsistencies raise the possibility of hallucinations. Web searches are used by systems such as WebGPT (Nakano et al., 2021) to verify information. These approaches depend on the caliber of the knowledge source and may not work well in the absence of precise references [21, 22]. When data is accessible, their strength is high precision; nevertheless, they struggle with subtle facts and may be slow or constrained by API access. HaluEval discovered that supplying outside information greatly enhances the ability to recognize hallucinations [23].

- **Consistency-Based Detection (Self-Checks).** These techniques make use of the LLM's own logic. An LLM (such as GPT-3.5) is prompted by SelfCheckGPT (Manakul et al., 2023) to evaluate its original response by generating alternative replies or chain-of-thought reasoning before comparing consistency. A possible hallucination is indicated if several runs are inconsistent. Self-consistency decoding (Wang et al., 2023) creates several lines of reasoning and chooses the most common response. It makes sense that a right response is more frequently discovered in separate samples. These methods involve overhead because they need to execute the model several times. They are limited if the model repeatedly exhibits the same delusion, but they can identify internal conflicts [24, 25]. Nevertheless, they are attractive as they don't require outside data.

- **Uncertainty Estimation and Confidence Scoring.** Token probabilities, which can be understood as confidence, are generated by LLMs. Hallucinations may be suggested by low confidence (low probability, high entropy). Uncertainty is assessed at the token, phrase, or entity levels in recent work [26][27]. For instance, entropy scores are calculated by Guerrairo et al. (2023) and Malinin & Gales (2021). Uncalibrated probabilities, however, frequently overstate confidence and are unable to accurately identify errors [27]. Yeh et al. (2024) demonstrate that while context-aware scoring, which accounts for semantic relevance, increases reliability, basic token-probability techniques overpredict hallucinations [27]. Although uncertainty-based approaches are generally lightweight and reference-free, they currently lack accuracy and granularity (they flag entire phrases) [26, 27].

- **Retrieval-Augmented Verification.** Similar to RAG but used for verification, the LLM is used to confirm its claim against pertinent documents or facts that are first retrieved for a specific query. This could entail creating a search query that reflects the statement. The output is probably hallucinated if the evidence that was retrieved contradicts it. This integrates model thinking with outside knowledge. Although it increases verification robustness, it depends on retrieval quality and adds latency.

- **Human-Evaluation Benchmarks.** Empirical detection is provided via extensive annotated benchmarks. TruthfulQA (Lin et al., 2022) examines whether models perpetuate prevalent misconceptions. 5k+ ChatGPT replies classified for hallucinations are provided by HaluEval (Li et al., 2023) [6]. FActScore (Min et al., 2023) calculates the fraction of generation that is accurate by breaking it down into atomic facts [7]. These benchmarks are assessment metrics rather than online detection technologies. They draw attention to model variations, such as the fact that FactScore found ChatGPT to be roughly 58% true on biographies and that GPT-4 was superior to GPT-3-based models [7]. They are used to compare and calibrate detection techniques. Methods are summarized in Table 2.

Table 2. *Comparison of hallucination detection approaches.*

Method	Approach	Strengths	Limitations
Knowledge-Based	Use external KB or search to fact-check model output[21].	High precision if relevant info exists; explainable.	Coverage limited; relies on query matching.
Self-Consistency	Generate multiple answers/reasoning, check agreement[24].	Model-internal, no external data needed.	High compute cost; fails if model repeats error.
Uncertainty Scoring	Compute token-level entropy/confidence (e.g. likelihood, entropy)[27].	Simple, fast, reference-free.	Poor calibration; coarse (sentence-level); over/under-predicts.
Retrieval Verification	Retrieve evidence then check output (LLM or rule-based) against it.	Uses up-to-date info; adaptable to domains.	Latency; requires good retriever; extra complexity.
Benchmarks (TruthfulQA, HaluEval, FactScore)	Large annotated QA sets or generative tasks to measure hallucination rate[6][7].	Standardized evaluation across models.	Not a real-time method; not fine-grained detection.

Each detection method has tradeoffs. For instance, Li et al. (2023) showed that adding external knowledge or reasoning steps can help LLMs identify their own hallucinations[23], suggesting hybrid strategies. Uncertainty methods are improving (entity-level detection in Yeh et al., 2024[28]), but are not yet turnkey solutions. Ultimately, best practices often combine methods (e.g., self-checking plus retrieval) to balance precision and recall.

VI. Mitigation Strategies

Reducing hallucinations involves aligning the model more closely with truth and context. Key strategies include:

- Retrieval-Augmented Generation (RAG). External papers are incorporated into the creation process via RAG [29]. The model grounds its output by conditioning on retrieved factual texts. RAG was developed by Lewis et al. (2020) by fusing parametric LMs with non-parametric memory. In actuality, a retrieval step obtains pertinent data (such as Wikipedia) before to producing an answer. Since the model can replicate or paraphrase recovered material instead of making guesses, this can significantly increase factual accuracy. RAG's efficacy, however, is dependent on the retriever and knowledge base; if pertinent information is either ignored by the model or not retrieved, it may still experience hallucinations. RAG is not suitable for isolated domains without external corpora and adds complexity.
- Fine-Tuning with Curated Data (RLHF). To minimize unwanted outputs, instruction tailoring and reinforcement learning from human feedback (RLHF) have been employed. InstructGPT (via RLHF) produces more accurate responses than vanilla GPT-3, as demonstrated by Ouyang et al. (2022) [30]. By explicitly rewarding helpful and truthful responses, RLHF trains the model to favor correct material. In a similar vein, knowledge-grounded conversations or factual QA pairings can benefit from supervised fine-tuning. Although RLHF greatly decreased toxic/hallucinated outputs in InstructGPT, it is still flawed (e.g., InstructGPT is not entirely genuine) and labor-intensive (requires human labels) [30]. Additionally, RLHF may produce new biases or overfit to training challenges.
- Chain-of-Thought Prompting. Answer accuracy is frequently increased by asking the model to generate intermediate reasoning stages (chain-of-thought, or CoT) [31]. It makes sense that if the model follows a methodical approach, it is less likely to reach an incorrect result. The model can be guided to confirm facts before declaring them, for instance, by providing exemplars with reasoning in the prompt. Sampling several CoT pathways and choosing the majority response is related to self-consistency (Wang et al., 2023) [32]. Empirically, self-consistency and CoT significantly raised reasoning task scores. Since a fallacy arises in thinking, CoT can detect logical mistakes in hallucination reduction and reduce answers that are overconfident. But not all jobs are beneficial, and it can also introduce more text that may involve hallucinations.
- Grounding with Tools and Citations. Current methods enable LLMs to give sources or invoke tools (calculation, knowledge bases, web search). For example, the model can access search or APIs during generation thanks to OpenAI's WebGPT and more recent plugins. Hallucinations are easier to identify or prevent if the model references sources or facts. To lessen hallucinations by refusal, another strategy is to teach models to communicate hesitation or say "I don't know" when unsure. This "abstention" is a potentially safer method of grounding (grounding in ambiguity). The drawback is that over-refusal diminishes utility, and research on tool integration is currently ongoing.
- Constitutional AI and Self-Correction. According to a set of guidelines, Anthropic's Constitutional AI (Bai et al., 2022) teaches models to evaluate and modify their own

outputs [33]. After producing a response, the model evaluates its veracity or harmfulness in light of "constitutional" guidelines before refining the result. If the model "concludes" that its response is inconsistent with its constitutional values, such as "be factual," this process may detect some hallucinations. The final model is typically safer because it was taught using reinforcement learning on self-critiques. Although it necessitates properly thought-out ideas, constitutional AI lessens the need for human designations. Additionally, it makes the erroneous assumption that the model is capable of accurate self-criticism.

- Prompt Engineering. The model can be guided away from hallucinations by properly crafting instructions. Pure guesswork can be discouraged, for instance, by specifically requesting proof or a line of reasoning. The model may be prompted to credit sources or quote facts. Fabrications can be decreased by using templates that remind the model to verify its facts, such as "Base your answer on known facts" or "If uncertain, say you don't know." Empirically, the model may be biased to rely on training knowledge by making even little adjustments, such as adding "according to research." Prompt engineering is not infallible, though; astute users can still induce hallucinations with hostile prompts, and prompts that cause a model to become unduly cautious may hinder innovation.

Table 3 compares these strategies.

Table 3. *Comparison of hallucination mitigation strategies.*

Strategy	Description	Effectiveness	Limitations
Retrieval-Augmented Gen (RAG)[29]	Incorporate retrieved documents into generation.	Often significantly reduces factual errors (if KB covers query).	Requires external datastore and retriever; may ignore irrelevant info.
RLHF / Fine-tuning[30]	Fine-tune on human feedback or curated factual data to prefer truthful output.	Improves overall truthfulness (OpenAI found higher accuracy).	Expensive (human labels); may not generalize outside training tasks.
Chain-of-Thought (CoT)[31]	Prompt model to produce reasoning steps.	Can markedly increase reasoning accuracy and reduce logical errors.	Increases response length (more hallucination opportunities); inconsistent benefits.
Grounding/Tool Use	Enable model to query tools (web, DBs) or cite sources in answers.	Tethers output to real data; can improve accuracy if tools are reliable.	Dependent on tool quality; API calls add complexity and latency.
Constitutional AI[33] & Self-correction	Models generate critiques and revisions of their own answers under a fixed constitution.	Reduces unacceptable/hallucinated answers without human labels.	Depends on model's ability to self-critique; complex RL pipeline.
Prompt Engineering	Design prompts that emphasize accuracy (e.g. "Answer with sources").	Sometimes reduces hallucination by framing expectations.	Unpredictable; may require trial and error; not guaranteed.

There are trade-offs for every mitigation. Ensemble techniques are often the most effective. For example, RLHF-trained models may make better use of tools, and combining RAG with CoT results in grounded yet rational responses. The literature points to advancements but offers no magic bullet. Notably, Meta's work on Llama-3 (2024) reflects the field's move toward grounded LLM usage by emphasizing tool use and retrieval to further eliminate fabrication (e.g., Llama-3 can be made to call Bing search). In the end, minimizing hallucinations requires more integration of alignment techniques and outside knowledge.

VII. Prepare Comparative Analysis of LLMs

How do different LLMs compare in their propensity to hallucinate? Benchmarks like TruthfulQA, HalluEval, and HalluLens provide empirical comparisons. While new models (GPT-5, Gemini 3 etc.) have emerged, publicly available evaluations suggest trends:

- **GPT-4 vs GPT-3.5.** GPT-4 consistently had less hallucinations than GPT-3.5 in a number of tests. For instance, GPT-4 accurately responded about 85% of items on TruthfulQA (a factual QA benchmark), resulting in a 15% hallucination rate, while open-source analogs performed far worse [4]. According to the Frontiers poll, GPT-4 had only about 14% errors on TruthfulQA, compared to about 31% for LLaMA 2 (and similarly higher for GPT-3.5) [4]. In practical use, the RLHF and greater size of GPT-4 seem to lessen obvious factual inaccuracies. Additionally, much fewer problems with GPT-4 are found using SelfCheckGPT-style techniques.

- **Open Models (LLaMA 2/3, Mistral, Gemma, etc.).** In general, smaller models have greater hallucinations. In contrast to substantially lower rates for larger siblings, Bang et al. (2025) discovered that Llama-3.1-8B-Instruct had a ~48% hallucination rate (when it tried an answer) on their QA suite [34]. Llama-3.1-405B experienced only about 26.8% hallucinations in one HalluLens task

(although it also rejected more than 50% of queries) [5]. Mistral-7B performed poorly (~81% of questions answered involved hallucinations) [34]. Claude-3 (Anthropic) tended to decline more inquiries, which reduces hallucinations at the expense of coverage [35]. Larger models generally result in fewer hallucinations when they respond, according to trends. Comparisons are made more difficult by the fact that many smaller models just refuse to respond (high rejection).

- **Benchmarked Results.** GPT-4 had the greatest (factually) score when FactScore (Min et al., 2023) assessed GPT-4, ChatGPT, Vicuna, and others on biography generation [7]. About 19.5% of general searches had ChatGPT hallucinations, according to HaluEval [6]. According to the 2026 Hallucination Leaderboard (unofficial), the most recent models (GPT-5, Gemini 2.x) have hallucination rates on summary tasks that are less than 5%. Reports on TruthfulQA and HalluLens-type evaluations are compiled in Table 4.

Table 4. *Hallucination benchmarks: model comparisons.* (Values approximate based on cited sources)

Model	TruthfulQA Correct (%)[4]	HalluLens QA Hallucination (%)[5]
GPT-4	~86	~45 (while answering)
GPT-3.5 (Curie)	~58–60 (prior reports)	–
Llama-2 13B	~69	–
Llama-3.1 405B	–	~27 (when answering)
Mistral-7B	–	~81 (answering rate)
Claude-3 Sonnet	–	>50 (when answering)
ChatGPT (GPT-3.5)	~60 (Lin et al., 58)	~19 (Hallucination rate)[6]

These numbers show that GPT-4/ChatGPT performs better on factuality than smaller open models. Metrics vary, though, with some measuring "incorrect answer rate" and others "hallucinated response rate among answered queries." While smaller open

models can be improved by fine-tuning or augmentation, larger closed models typically achieve higher fidelity but are less transparent. Crucially, as demonstrated by Claude and LLaMA-3, rejection behavior (saying "I don't know") can artificially lower hallucination rates. GPT-4's responses continue to be more accurate when normalized by answer attempts [5].

In conclusion, open research models (LLaMA, Mistral, Qwen, Gemma) vary and typically trail behind frontier models like GPT-4 and future GPT-5 versions on truthfulness benchmarks [4][5]. As new versions (Gemini 3.1, Grok, etc.) appear with purported hallucination mitigations, ongoing assessment is required.

VIII. Challenges & Open Problems

Despite progress, many issues in LLM hallucination remain unsolved:

- **Lack of Standardized Benchmarks.** Current benchmarks (TruthfulQA, HalluLens, HaluEval, etc.) employ distinct definitions and methods, as observed by recent surveys [36][12]. For hallucinations, there is no universally accepted measure. While some tests emphasize consistency, others concentrate on factuality. This makes direct comparisons of models or approaches challenging. It is necessary to have a single evaluation package that addresses various aspects of hallucinations.

- **Subjective Evaluation.** Human assessment of factuality is frequently required to determine hallucinations, which can be subjective for complex claims. It can be difficult to distinguish between a hallucination and an appropriate inference. Creative extrapolations, such as the creation of stories, blur the boundaries. It is still difficult to automate evaluation without gold references.

- **Multilingual Hallucination.** The majority of research focuses on English. There is not enough research on LLM hallucination behavior in other languages. According to preliminary research, translation may cause hallucinations if training data is inconsistent. Research on cross-lingual hallucinations, or hallucinations in low-resource languages, is still ongoing.

- **Domain-Specific Hallucination.** There are high stakes in several professions, such as medical, law, science, and finance. In these sectors, hallucinations are very important, but they are also more difficult to identify (needs professional knowledge). Specialized benchmarks (like medical quality assurance) and domain adaptability are required. For example, an LLM may mistakenly hallucinate a rare illness; specialist testing is need to identify this.

- **Faithfulness vs. Creativity Tradeoff.** There is a basic conflict: more imaginative outputs run the risk of hallucination, while extremely limited factual solutions can appear inflexible. It's challenging to find the ideal balance so that models are both

accurate and captivating. If loyalty is overemphasized, language innovation or fluency may suffer. It is a constant struggle to comprehend and manage this trade-off.

- **Calibration and Trust.** LLMs frequently make misleading claims with excessive confidence. The problem of "calibration," or creating models that can recognize when they could be incorrect, remains unresolved. Probabilistic modeling techniques might be useful, but their large-scale integration is difficult. Users require accurate answers as well as trustworthy uncertainty estimations.

These difficulties demonstrate that hallucinations are systemic problems related to model design, evaluation culture, and deployment setting rather than merely a technological flaw. Interdisciplinary work is needed to address these, from improved benchmarks to regulatory norms.

IX. Future Directions

We conclude by outlining promising research directions:

- **Neuro-Symbolic & Structured Grounding.** LLM outputs could be grounded by using symbolic reasoning or logic constraints. Some hallucinations can be avoided, for example, by utilizing rule-based checks or connecting LLMs to knowledge graphs. Formal reasoning layers may guarantee consistency with fundamental facts, according to early research on neuro-symbolic frameworks. One important approach is to create LLMs that can cross-verify with structured knowledge.
- **Real-Time Fact Verification Pipelines.** constructing end-to-end systems in which a search engine or auxiliary model instantly verifies an LLM response before it is shown. For on-the-spot verification, this "fact guard" might employ quick retrieval or more compact, specialized models. These processors could automatically annotate or fix hallucinations, much as modern grammar checkers.
- **LLM Self-Awareness & Calibration.** There is an increasing amount of research on educating LMs about their uncertainty. Techniques such as measuring token-level probabilities (Verbalized Uncertainty) could be enhanced. It would improve safety to train LLMs to identify possible hallucinations in their own work, possibly through introspective prompting. It is still unclear how to incorporate explicit truthfulness calibration into training.
- **Multimodal Hallucination.** Visual hallucinations occur when LLMs expand to vision (e.g., GPT-4's image inputs), describing imaginary objects. The interaction between textual and visual hallucinations should be studied. As systems integrate vision and language, benchmarks and techniques for multimodal factuality (e.g., cross-checking visual captions versus picture data) are required.
- **Community-Driven Benchmarks and Open Challenges.** Shared benchmarks, such as the Hallucination Leaderboard and AI multi-benchmarks, are beneficial to the profession. Models can be stress-tested through cooperative efforts to crowdsource difficult hallucination scenarios (such as hostile factual inquiry). Standardization would be encouraged by holding open competitions, similar to fact-checking competitions.
- **Regulatory and Ethical Frameworks.** The way society views hallucinations should expand beyond technical solutions. Labeling material produced by AI or establishing truthfulness requirements in specific applications are two examples. Development might be guided by working with the ethics and policy groups to establish trustworthiness standards, similar to those used by the FDA for medical claims.

In conclusion, resolving LLM hallucinations requires a multidisciplinary approach that includes system engineering, model innovation, evaluation science, and ethics. The community's actions in the coming years will be essential to ensuring that AI outputs are consistently accurate.

X. Conclusion

One major obstacle to reliable AI is still hallucinations. We have examined the issue from every perspective in this survey, including defining a comprehensive taxonomy of hallucination types (factual/faithfulness, intrinsic/extrinsic, domain levels) [2][10], identifying underlying causes (data flaws, inference mechanisms) [20], and examining cutting-edge detection and mitigation techniques. Although no model is impervious to inaccuracy, our comparison analysis reveals that leading models (GPT-4, Claude) outperform older or smaller models on truthfulness metrics [4][5]. Both post-hoc detection (fact-checkers, consistency checks) and proactive mitigation (RAG, RLHF, prompting) have advanced, yet often at the cost of complexity or responsiveness. Important discoveries include: factuality is significantly enhanced by incorporating outside knowledge (via tools or retrieval) [29][23]; Truthfulness is greatly increased by fine-tuning with human feedback [30], and chain-of-thought and self-consistency aid in identifying faults during generation [31][24]. However, there is still more to be done to solve subtle failure modes (long-tail, multilingual instances) and standardize evaluation.

In conclusion, reducing hallucinations is essential for the responsible use of LLMs in fields including research, law, and healthcare [3]. We hope that this survey gives researchers and students a thorough foundation, encouraging more developments. Making sure AI systems "tell the truth" will be just as crucial as making them fluent as they grow more ingrained in society. Thus, the pursuit of faithful AI generation is both socially and technically necessary.

References

[1] S. Lin, J. Hilton, and O. Evans, "TruthfulQA: Measuring how models mimic human falsehoods," in *ACL*, 2022.
[2] J. Li, X. Cheng, W. X. Zhao, J.-Y. Nie, and J.-R. Wen, "HaluEval: A large-scale hallucination evaluation benchmark for large language models," *Proc. EMNLP*, 2023.
[3] S. Min, K. Krishna, X. Lyu, M. Lewis, W.-T. Yih, P. W. Koh, M. Iyyer, and L. Zettlemoyer, "FACTSCORE: Fine-grained atomic evaluation of factual precision in long form text generation," in *EMNLP*, 2023.
[4] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin et al., "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *NeurIPS*, 2020.
[5] L. Ouyang, J. Wu, X. Jiang, D. Almeida et al., "Training language models to follow instructions with human feedback," in *NeurIPS*, 2022.
[6] J. Wei, X. Wang, D. Schuurmans, M. Bosma et al., "Chain-of-thought prompting elicits reasoning in large language models," in *ICLR*, 2023.
[7] X. Wang, J. Wei, D. Schuurmans, Q. Le et al., "Self-consistency improves chain of thought reasoning in language models," in *ICLR*, 2023.
[8] Y. Bai, S. Kadavath, S. Kundu, A. Askell et al., "Constitutional AI: Harmlessness from AI feedback," *ArXiv*, 2022.
[9] Z. Ji, N. Lee, R. Frieske, T. Yu et al., "Survey of hallucination in natural language generation," *ACM TACL*, vol. 8, 2023.
[10] Y. Huang, Z. Wang, Y. Yang, H. Liu et al., "A survey on hallucination in large language models: principles, taxonomy, challenges, and open questions," *ACM Trans. Inf. Syst.*, vol. 42, no. 2, 2024.
[11] M. E. Yeh, A. Calligaro, N. Johnson et al., "HalluEntity: Detecting hallucinated entities in large language model outputs," *ArXiv*, 2024.
[12] P. Manakul, A. Liusie, and M. Gales, "SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models," in *EMNLP*, 2023.
[13] Y. Bang, Z. Ji, A. Schelten, A. Hartshorn, T. Fowler, C. Zhang, N. Cancedda, and P. Fung, "HalluLens: LLM hallucination benchmark," in *ACL*, 2025.
[14] Q. Huang, J. Zhou, K. He et al., "FaithLens: Detecting and explaining faithfulness hallucination in large language models," *ArXiv*, 2026.
[15] J. Z. Yang, J. Gao, Y. Nie et al., "HalluRAG: Detecting and mitigating closed-domain hallucinations in retrieval-augmented generation," *ArXiv*, 2024.
[16] J. Ji, S. He, Y. Huang, and T.-S. Chua, "Reviewing factuality in text generation: A taxonomy and perspective," *ArXiv*, 2023.
[17] R. Chen, A. Kumar, and X. Xu, "Calibrating the confidence of large language models," *ArXiv*, 2024.
[18] R. Chen *et al.*, "Hallucination in medical foundation models and their impact on decision-making," *ArXiv*, 2024.
[19] "GPT-4 Technical Report," OpenAI, 2023.
[20] T. Touvron *et al.*, "LLaMA: Open and Efficient Foundation Language Models," *NeurIPS Workshops*, 2023.

[1] Hallucination (artificial intelligence) - Wikipedia
https://en.wikipedia.org/wiki/Hallucination_(artificial_intelligence)
[2] [24] FaithLens: Detecting and Explaining Faithfulness Hallucination
https://arxiv.org/html/2512.20182v3
[3] [17] [21] [22] [25] [26] [27] [28] Can Your Uncertainty Scores Detect Hallucinated Entity?
https://arxiv.org/html/2502.11948v1
[4] [8] [11] [36] Frontiers | Survey and analysis of hallucinations in large language models: attribution to prompting strategies or model behavior
https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2025.1622292/full
[5] [15] [18] [20] [34] [35] HalluLens: LLM Hallucination Benchmark
https://arxiv.org/html/2504.17550v1
[6] [23] [2305.11747] HaluEval: A Large-Scale Hallucination Evaluation Benchmark for Large Language Models
https://arxiv.org/abs/2305.11747
[7] [2305.14251] FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation
https://arxiv.org/abs/2305.14251
[9] [13] [2202.03629] Survey of Hallucination in Natural Language Generation
https://ar5iv.labs.arxiv.org/html/2202.03629
[10] [14] [19] A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions
https://arxiv.org/pdf/2311.05232
[12] HalluLens: LLM Hallucination Benchmark - ACL Anthology
https://aclanthology.org/2025.acl-long.1176/

[16] [29] The HalluRAG Dataset: Detecting Closed-Domain Hallucinations in RAG Applications Using an LLM's Internal States
https://arxiv.org/html/2412.17056v2
[30] [2203.02155] Training language models to follow instructions with human feedback
https://arxiv.org/abs/2203.02155
[31] [2201.11903] Chain-of-Thought Prompting Elicits Reasoning in Large Language Models
https://arxiv.org/abs/2201.11903
[32] [2203.11171] Self-Consistency Improves Chain of Thought Reasoning in Language Models
https://arxiv.org/abs/2203.11171
[33] [2212.08073] Constitutional AI: Harmlessness from AI Feedback
https://arxiv.org/abs/2212.08073