

# A comprehensive taxonomy of hallucinations in Large Language Models

Manuel Cossio MMed, MEng  
*Universitat de Barcelona*  
*manuel.cossio@ub.edu*

---

## Abstract

Large language models (LLMs) have revolutionized natural language processing, yet their propensity for "hallucination"—generating plausible but factually incorrect or fabricated content—remains a critical challenge. This report provides a comprehensive taxonomy of LLM hallucinations, beginning with a formal definition and a theoretical framework that posits its inherent inevitability in computable LLMs, irrespective of architecture or training. It explores core distinctions, differentiating between intrinsic (contradicting input context) and extrinsic (inconsistent with training data or reality), as well as factuality (absolute correctness) and faithfulness (adherence to input). The report then details specific manifestations, including factual errors, contextual and logical inconsistencies, temporal disorientation, ethical violations, and task-specific hallucinations across domains like code generation and multimodal applications. It analyzes the underlying causes, categorizing them into data-related issues, model-related factors, and prompt-related influences. Furthermore, the report examines cognitive and human factors influencing hallucination perception, surveys evaluation benchmarks and metrics for detection, and outlines architectural and systemic mitigation strategies. Finally, it introduces web-based resources for monitoring LLM releases and performance. This report underscores the complex, multifaceted nature of LLM hallucinations and emphasizes that, given their theoretical inevitability, future efforts must focus on robust detection, mitigation, and continuous human oversight for responsible and reliable deployment in critical applications.

---

## Contents

<b>1</b>	<b>Introduction</b>	<b>4</b>
<b>2</b>	<b>Defining hallucination in LLMs</b>	<b>5</b>
2.1	General conceptual definition . . . . .	5
2.2	Formal definition and inevitability . . . . .	5
2.2.1	Formal definition . . . . .	5
2.2.2	Implications for inevitability . . . . .	6
<b>3</b>	<b>Core taxonomies of LLM hallucinations</b>	<b>9</b>
3.1	Intrinsic vs. extrinsic hallucinations . . . . .	9
3.2	Factuality vs. faithfulness hallucinations . . . . .	10
<b>4</b>	<b>Specific categories and manifestations of hallucinations</b>	<b>11</b>
4.1	Factual errors and fabrications . . . . .	11
4.1.1	Incorrect facts . . . . .	11
4.1.2	Fabricated entities/information . . . . .	11

4.1.3	Adversarial attacks . . . . .	11
4.2	Contextual inconsistencies . . . . .	11
4.3	Instruction inconsistencies/deviation . . . . .	12
4.4	Logical inconsistencies . . . . .	12
4.5	Temporal disorientation . . . . .	12
4.6	Ethical violations . . . . .	12
4.6.1	Defamation/misinformation . . . . .	12
4.6.2	Financial misinformation . . . . .	13
4.6.3	Legal inaccuracies . . . . .	13
4.7	Amalgamated hallucinations . . . . .	13
4.8	Nonsensical responses . . . . .	13
4.9	Task-specific hallucinations . . . . .	13
4.9.1	Dialogue history-based hallucination . . . . .	13
4.9.2	Abstractive summarization hallucination . . . . .	13
4.9.3	Generative question answering hallucination . . . . .	14
4.9.4	Code generation hallucination . . . . .	14
4.9.5	Multimodal large language models hallucination . . . . .	14
4.10	Complexities and critical implications of diverse hallucination types . . . . .	14
<b>5</b>	<b>Underlying causes of hallucinations</b>	<b>15</b>
5.1	Data-related factors . . . . .	16
5.1.1	Quality and volume of training dataset . . . . .	16
5.1.2	Inadequate representation and biases . . . . .	16
5.1.3	Outdated training data and knowledge boundary . . . . .	16
5.1.4	Source-reference divergence . . . . .	16
5.2	Model-related factors . . . . .	16
5.2.1	Auto-regressive nature . . . . .	16
5.2.2	Architecture flaws and internal design . . . . .	17
5.2.3	Training processes . . . . .	17
5.2.4	Decoding strategies . . . . .	17
5.2.5	Overconfidence and calibration . . . . .	17
5.2.6	Generalization to unseen cases . . . . .	17
5.2.7	Lack of reasoning and nuanced language understanding . . . . .	18
5.2.8	Knowledge overshadowing . . . . .	18
5.2.9	Insufficient knowledge representation . . . . .	18
5.2.10	Failure in information extraction . . . . .	18
5.3	Prompt-related factors . . . . .	18
5.3.1	Adversarial attacks . . . . .	18
5.3.2	Overly confirmatory tendency . . . . .	18
5.3.3	Prompting methods . . . . .	19
5.4	An emergent property requiring systemic solutions . . . . .	19
<b>6</b>	<b>Cognitive and human factors in hallucination perception</b>	<b>21</b>
6.1	User trust and interpretability . . . . .	21
6.2	Hallucinations often go unnoticed . . . . .	21
6.3	Cognitive biases amplifying hallucination risks . . . . .	21
6.3.1	Automation bias . . . . .	21
6.3.2	Confirmation bias . . . . .	22

6.3.3	Illusion of explanatory depth . . . . .	22
6.3.4	Persistence of biases despite warnings . . . . .	22
6.4	Design implications and mitigation strategies . . . . .	22
6.4.1	Calibrated uncertainty displays . . . . .	22
6.4.2	Source-grounding indicators . . . . .	23
6.4.3	Justification prompts . . . . .	23
6.4.4	Factuality-aware interface prototypes . . . . .	23
6.5	Human-in-the-loop evaluation and oversight . . . . .	23
<b>7</b>	<b>Evaluation benchmarks and metrics for hallucination detection</b>	<b>23</b>
7.1	Benchmark datasets . . . . .	24
7.1.1	TruthfulQA . . . . .	24
7.1.2	HalluLens . . . . .	24
7.1.3	FActScore . . . . .	24
7.1.4	Q2 and QuestEval . . . . .	25
7.1.5	Domain-specific benchmarks . . . . .	25
7.2	Quantitative metrics . . . . .	26
7.2.1	Faithfulness metrics . . . . .	26
7.2.2	Factuality metrics . . . . .	27
7.2.3	Human evaluation . . . . .	28
7.3	Limitations and open challenges . . . . .	28
7.3.1	Lack of standardization . . . . .	28
7.3.2	Task dependence . . . . .	28
7.3.3	Insensitivity to subtle hallucinations . . . . .	29
7.3.4	Limited grounding and explainability . . . . .	29
7.4	Toward unified evaluation frameworks . . . . .	29
<b>8</b>	<b>Hallucination mitigation strategies</b>	<b>30</b>
8.1	Architectural mitigation strategies . . . . .	30
8.1.1	Toolformer-style augmentation . . . . .	30
8.1.2	Factual grounding through retrieval mechanisms . . . . .	30
8.1.3	Fine-tuning with synthetic or adversarially filtered data . . . . .	31
8.2	Systemic mitigation strategies . . . . .	31
8.2.1	Guardrails and symbolic integration . . . . .	31
8.3	Toward hybrid and context-aware mitigation systems . . . . .	32
<b>9</b>	<b>Monitoring LLM releases and performance: web-based resources</b>	<b>33</b>
9.1	Artificial Analysis . . . . .	33
9.1.1	Intelligence index . . . . .	33
9.1.2	Cost and latency . . . . .	34
9.1.3	Model pages . . . . .	35
9.1.4	Multimodal and API benchmarks . . . . .	35
9.1.5	Quarterly state-of-AI reports . . . . .	36
9.2	Vectara Hallucination Leaderboard . . . . .	37
9.3	Epoch AI Benchmarking Dashboard . . . . .	37
9.3.1	Accuracy versus training compute . . . . .	38
9.3.2	Open source versus proprietary . . . . .	39
9.3.3	Geographic disparities and performance . . . . .	40

9.3.4	Performance on expert-level mathematics problems . . . . .	41
9.4	LM Arena . . . . .	42
9.4.1	Battle-style comparisons and dynamic prompts . . . . .	43
9.4.2	Diversity of models and transparent testing . . . . .	43
9.4.3	High-quality qualitative judgments driven by intrinsic motivation . . . .	43
9.4.4	Community-shaped leaderboard and transparency . . . . .	43
<b>10</b>	<b>Conclusions</b>	<b>45</b>
10.1	The complex nature and inevitable presence of LLM hallucinations . . . . .	45
10.2	Implications for detection, mitigation, and human interaction . . . . .	46
10.3	Future directions for responsible LLM deployment . . . . .	46

# 1 Introduction

Large language models (LLMs) represent a significant advancement in natural language processing (NLP), fundamentally altering how information is acquired and processed<sup>[38;27]</sup>. These models have enabled a paradigm shift, facilitating diverse applications ranging from sophisticated content creation to advanced decision support systems<sup>[38;27]</sup>. Their capacity to generate human-like text has led to remarkable progress in various tasks, including complex question-answering systems, abstractive summarization, and interactive conversational agents<sup>[82]</sup>. The widespread adoption of LLMs underscores their transformative potential across numerous industries and research domains.

Despite their impressive capabilities, a critical and widely acknowledged limitation of LLMs is their propensity for "hallucination"<sup>[38;27;42;47]</sup>. This phenomenon describes the generation of content that, while often plausible and coherent, is factually incorrect, inconsistent, or entirely fabricated<sup>[38;27]</sup>. Unlike the medical definition of hallucination, which refers to sensory experiences in the absence of external stimuli, in the context of LLMs, it signifies the creation of nonfactual information to respond to a user's query, frequently without any explicit indication of its fabricated nature<sup>[68]</sup>. Such generated content is characterized as incorrect, nonsensical, and lacking justifiable basis, making its detection challenging for users. The prevalence of hallucinations raises significant concerns regarding the reliability and trustworthiness of LLMs, particularly as their integration into real-world information retrieval (IR) systems and critical decision-making processes continues to expand<sup>[38]</sup>.

This report provides a comprehensive taxonomy of hallucinations in Large Language Models (LLMs), delving into the critical challenge of their propensity for generating plausible but factually incorrect or fabricated content. It begins by formally defining hallucination, including a framework that posits its inherent inevitability in computable LLMs, irrespective of their architecture or training. The core distinctions of hallucination are explored, differentiating between intrinsic (contradicting input context) and extrinsic (inconsistent with training data or reality), as well as factuality (absolute correctness) and faithfulness (adherence to input)<sup>[38]</sup>.

The report then details specific manifestations of hallucinations, such as factual errors, contextual and logical inconsistencies, temporal disorientation, ethical violations, and task-specific hallucinations in domains like code generation and multimodal applications<sup>[94]</sup>. A thorough analysis of the underlying causes is presented, categorizing them into data-related issues (e.g., quality, biases), model-related factors (e.g., auto-regressive nature, decoding strategies, lack of reasoning), and prompt-related influences (e.g., adversarial attacks)<sup>[38;42]</sup>.

The report further examines the cognitive and human factors influencing hallucination perception, including user trust, interpretability, and various cognitive biases, alongside their implications for design and mitigation strategies<sup>[102]</sup>. It provides a comprehensive overview of evaluation benchmarks and metrics for hallucination detection, surveying principal datasets and quantitative metrics, as well as their current limitations. Additionally, a detailed exploration of hallucination mitigation strategies is included, covering both architectural approaches (e.g., Toolformer-style augmentation, factual grounding through retrieval mechanisms) and systemic approaches (e.g., guardrails, symbolic integration)<sup>[86;42]</sup>. Finally, the report introduces crucial web-based resources for monitoring LLM releases and performance, offering insights into tools and leaderboards that track hallucination rates and model reliability in real-world scenarios. This report underscores the complex, multifaceted nature of LLM hallucinations and emphasizes that, given their theoretical inevitability, future efforts must focus on robust detection, mitigation, and continuous human oversight for responsible and reliable deployment in critical applications.

---

## 2 Defining hallucination in LLMs

**This section formally defines LLM hallucination and presents a theoretical framework arguing for its inherent inevitability.**

---

### 2.1 General conceptual definition

In the domain of LLMs, hallucination is broadly understood as the generation of "plausible yet nonfactual content"<sup>[38]</sup>. This implies that an LLM produces "false or fabricated information" or outputs that are "inaccurate, irrelevant, or simply does not make factual sense"<sup>[42;47]</sup>. A key distinction from the medical definition of hallucination (sensory experiences without corresponding external stimuli) is that in LLMs, it refers to the creation of nonfactual content in response to a user's question, often without the model clarifying the fabricated nature of its answer<sup>[38]</sup>. This characteristic underscores the challenge of relying on LLM outputs without external verification.

### 2.2 Formal definition and inevitability

The paper "Hallucination is Inevitable: An Innate Limitation of Large Language Models" offers a formal framework for understanding hallucination, defining it within a "formal world" of computable functions to rigorously analyze its inherent inevitability in LLMs<sup>[100]</sup>.

#### 2.2.1 Formal definition

Hallucination is formally defined as an inconsistency between a computable LLM, denoted as  $h$ , and a computable ground truth function,  $f$ .

- **Formal world of  $f$  (ground truth function):** this is conceptualized as a set  $G_f = \{(s, f(s)) | s \in S\}$ , where  $f(s)$  represents the *sole* correct output for any given input string  $s$  from the set of all finite-length strings  $S$ <sup>[100]</sup>.

- **Training samples  $T$ :** these are defined as a collection of input-output pairs  $\{(s_0, y_0), (s_1, y_1), \dots | y_i = f(s_i)\}$  derived from the formal world of  $f$ . This set  $T$  serves as a generalized corpus representing the expected outputs of  $f$  for corresponding inputs<sup>[100]</sup>.
- **Hallucination condition:** an LLM  $h$  is considered to be "hallucinating" with respect to a ground truth function  $f$  if, across *all* training stages  $i$  (meaning, after being trained on any finite number of samples), there *exists* at least one input string  $s$  for which the LLM's output  $h[i](s)$  does not match the correct output  $f(s)$ <sup>[100]</sup>. This condition is formally expressed as  $\forall i \in \mathbb{N}, \exists s \in S$  such that  $h[i](s) \neq f(s)$ .

## 2.2.2 Implications for inevitability

The paper posits that hallucination is an inevitable characteristic of LLMs, irrespective of their architectural design, learning algorithms, prompting techniques, or the specific training data employed, provided they are considered "computable LLMs" operating within the defined formal world<sup>[100]</sup>. The central argument supporting this claim is rooted in diagonalization, a proof technique used in computability theory to demonstrate that certain infinite sets are inherently larger than others, implying limitations on what can be computed.

This theoretical framework leads to several critical theorems:

- **Theorem 1: computably enumerable LLMs will hallucinate:** this theorem states that for any computably enumerable set of LLMs (a category that includes all currently proposed polynomial-time bounded LLMs), there exists a computable ground truth function  $f$  such that *all* states of *all* LLMs within that set will exhibit hallucination<sup>[100]</sup>. This is demonstrated by constructing a ground truth function  $f$  that is specifically designed to contradict the output of every LLM state along a diagonal enumeration of all LLM states and their outputs.
- **Theorem 2: LLMs will hallucinate on infinitely many questions:** building upon the first theorem, this extends the argument to assert that for any computably enumerable set of LLMs, there exists a computable ground truth function  $f$  such that *all* states of *all* LLMs in that set will hallucinate on an *infinite number* of inputs<sup>[100]</sup>. This is shown by constructing an  $f$  that consistently differs from the output of each LLM state for an unending sequence of inputs.
- **Theorem 3: any computable LLM will hallucinate:** this theorem generalizes the preceding findings. It asserts that for *any individual computable LLM*, there exists a computable ground truth function  $f$  such that *every state* of that LLM will hallucinate with respect to  $f$ . Furthermore, for any computable LLM, there exists another computable ground truth function  $f'$  for which every state of that LLM will hallucinate on *infinitely many* inputs<sup>[100]</sup>. This theorem holds particular significance because real-world LLMs are considered a subset of total computable LLMs, directly extending the theoretical inevitability to practical applications.
- **Corollary 1: inability to self-eliminate hallucination:** a direct consequence of Theorem 3 is that *all computable LLMs inherently lack the capacity to prevent themselves from hallucinating*<sup>[100]</sup>. This implies that mitigation strategies relying solely on the LLM's internal mechanisms, such as prompt-based chain-of-thought reasoning, cannot fully eliminate hallucination.

**Practical implications of inevitability:** the formal definition and the theorems supporting the inevitability of hallucination carry profound practical implications for the development and deployment of LLMs:

- **General problem solvers:** LLMs, when trained solely on input-output pairs and employed as general problem solvers, are inherently prone to hallucination, particularly for problems that are computationally hard or lie beyond their learned capabilities<sup>[100]</sup>.
- **Scrutiny of mathematical and logic reasoning:** outputs from LLMs concerning mathematical problems and logic reasoning should always be subjected to rigorous scrutiny, as these domains often involve computationally challenging tasks that increase the likelihood of hallucination<sup>[100]</sup>.
- **Safety-critical decisions:** without the integration of external aids such as guardrails, knowledge bases, or direct human control, LLMs cannot be autonomously used in safety-critical decision-making processes. Human oversight remains paramount for decisions demanding rational and humane judgment<sup>[100]</sup>.
- **Research and regulations:** the inherent inevitability of hallucination underscores the urgent need for rigorous study and the establishment of appropriate regulations concerning the safety boundaries of LLMs. This is crucial to ensure their sustainable development and prevent their deployment in contexts that exceed their inherent capabilities<sup>[100]</sup>.

This observation highlights a fundamental aspect of hallucination: it is not merely a "bug" or an "error" that can be entirely eradicated through improved training or architectural design. Instead, it is an innate limitation rooted in the very nature of computability. If hallucination is formally proven to be inevitable for *any computable LLM*, it fundamentally redefines the objective from complete elimination to robust reduction and effective management. This understanding necessitates a paradigm shift in how LLMs are conceptualized, evaluated, and deployed. Rather than striving for perfect factual accuracy, the focus must pivot towards designing systems that incorporate robust detection mechanisms, containment strategies, and, crucially, human-in-the-loop validation, especially for applications where accuracy is paramount. This foundational understanding reinforces the indispensable role of external aids, such as Retrieval-Augmented Generation (RAG) systems, and direct human intervention, as LLMs cannot fully self-correct this inherent limitation (see summary in Table 1).

Table 1: Theorems and corollaries on LLM hallucination

Theorem/corollary	Statement	Implication for real-world LLMs	Reference
<b>Theorem 1: computably enumerable LLMs will hallucinate</b>	For any computably enumerable set of LLMs, there exists a computable ground truth function $f$ such that all states of all LLMs in that set will hallucinate.	All currently proposed polynomial-time bounded LLMs are inherently prone to hallucination; it cannot be completely eliminated.	[100]
<b>Theorem 2: LLMs will hallucinate on infinitely many questions</b>	For any computably enumerable set of LLMs, there exists a computable ground truth function $f$ such that all states of all LLMs in that set will hallucinate on infinitely many inputs.	Hallucinations are not isolated incidents but a persistent challenge across a vast range of inputs for any LLM.	[100]
<b>Theorem 3: any computable LLM will hallucinate</b>	For any individual computable LLM, there exists a computable ground truth function $f$ such that every state of that LLM will hallucinate. Furthermore, for any computable LLM, there exists another $f'$ such that every state will hallucinate on infinitely many inputs.	This generalizes inevitability to any specific LLM, confirming that current and future LLMs will always exhibit some form of hallucination.	[100]
<b>Corollary 1: inability to self-eliminate hallucination</b>	All computable LLMs cannot prevent themselves from hallucinating.	LLMs cannot solely rely on internal mechanisms (e.g., self-correction, chain-of-thought prompting) to eliminate hallucination; external safeguards are essential.	[100]



### 3 Core taxonomies of LLM hallucinations

This section outlines the primary categorizations of LLM hallucinations, distinguishing between intrinsic vs. extrinsic and factuality vs. faithfulness.

---

The scientific literature presents several key categorizations for LLM hallucinations, reflecting different perspectives on their nature and origin. Two widely accepted and fundamental distinctions are between intrinsic vs. extrinsic hallucinations and factuality vs. faithfulness hallucinations (see summary in Table 2).

#### 3.1 Intrinsic vs. extrinsic hallucinations

This dichotomy is a widely accepted and foundational distinction within the taxonomy of LLM hallucinations.<sup>[27;70;97;7]</sup> It differentiates errors based on their relationship to the provided input context and the model’s internal knowledge.

- **Intrinsic hallucination:** intrinsic hallucinations refer to generated text that directly contradicts the *provided input or context*<sup>[7;70]</sup>. These errors arise from logical inconsistencies *within the generated output itself*, without necessarily requiring reference to external knowledge<sup>[7]</sup>. This type of hallucination reflects the model’s inability to maintain consistency during the inference process or limitations stemming from its internal knowledge and parametric memory<sup>[7]</sup>. It can also encompass instances where the model misinterprets or omits crucial details from a given document, leading to an inaccurate representation of the source information.<sup>[7]</sup>

For example, if an article provided for summarization states that the FDA approved the first Ebola vaccine in 2019, an intrinsic hallucination would manifest as a summary claiming that the FDA rejected it. Another illustrative instance is a model summarizing an article that states a person was born in 1980, and then, later in the same summary, incorrectly claiming they were born in 1975, thereby demonstrating an internally inconsistent response.

- **Extrinsic hallucination:** extrinsic hallucinations, conversely, refer to generated text that is *not consistent with the training data* and “can neither be supported nor refuted by the input context”<sup>[79;7]</sup>. This category involves the introduction of entities, facts, or events that do not exist in reality. Such hallucinations frequently occur when models generate novel content or attempt to bridge perceived knowledge gaps<sup>[79;7]</sup>. This phenomenon highlights the model’s limitations in fully absorbing knowledge from its training data and its inability to accurately recognize the boundaries of its own knowledge. It can also result from issues related to integrating external information or from the model misinterpreting or failing to correctly incorporate the given context or prompt<sup>[79;7]</sup>.

An example of an extrinsic hallucination is the claim that “The Parisian Tiger was hunted to extinction in 1885,” a fabricated entity and event. Similarly, if a summarization article states that the FDA approved the first Ebola vaccine in 2019, an extrinsic hallucination might be a summary claiming that China started testing a COVID-19 vaccine, introducing information unrelated to the provided context.

### 3.2 Factuality vs. faithfulness hallucinations

This represents another prevalent categorization of LLM hallucinations, focusing on the truthfulness of the generated content and its adherence to the input<sup>[50]</sup>.

- **Factuality hallucination:** factuality hallucination occurs when an LLM generates "factually incorrect content".<sup>[42;50]</sup> This type of hallucination directly contradicts "real-world knowledge" or "established verification sources". It pertains to the "absolute correctness of the content generated" when compared against verifiable information. These errors often arise due to the model's limited contextual understanding and the inherent noise or inaccuracies present in its training data, leading to responses that are not grounded in reality<sup>[13;50]</sup>.

Examples include the model claiming "Charles Lindbergh was the first to walk on the moon", stating that "The Great Wall of China is visible from space", or generating the statement, "The speed of light in a vacuum is 100,000 km/s," when the correct value is approximately 299,792 km/s. Another instance is the assertion that "Thomas Edison invented the internet".

- **Faithfulness hallucination:** faithfulness errors occur when the model's output "diverges from the input prompt or provided context"<sup>[64;96;61]</sup>. The response generated by the model may be internally consistent and appear plausible, but it fails to align with the user's expectations or the specific information explicitly provided in the input. This type of hallucination is closely related to, and often overlaps with, intrinsic hallucination, as both deal with inconsistencies relative to the given source<sup>[64;96;61]</sup>.

For example, in the context of summarization, if an article states that the FDA approved the first Ebola vaccine in 2019, a faithfulness hallucination would include a summary claiming that the FDA rejected it, directly contradicting the provided source information.

The presence of multiple, slightly different, yet often overlapping taxonomies (e.g., intrinsic/extrinsic versus factuality/faithfulness) across various scientific articles<sup>[38;7;70;79;42;50;64;96;61]</sup> indicates that the field is still actively defining and refining the categorization of LLM hallucinations. While intrinsic and faithfulness hallucinations largely describe deviations from *provided context* or *internal consistency*, extrinsic and factuality hallucinations relate to inconsistencies with *external knowledge* or *real-world truth*. This nuance is critical because different types of hallucinations often stem from distinct underlying mechanisms and, consequently, require specific detection and mitigation strategies. For instance, Retrieval-Augmented Generation (RAG) is frequently cited as an effective method to combat factual or extrinsic hallucinations by grounding the model in external, verifiable knowledge<sup>[1]</sup>. In contrast, intrinsic hallucinations might necessitate more sophisticated internal consistency checks or improvements in the model's reasoning capabilities. This observation underscores that the absence of a "unified framework due to inconsistent definitions and categorizations" is a significant challenge in benchmarking hallucinations<sup>[7;70]</sup>. This implies that comparative research on hallucination rates and the development of universally applicable mitigation strategies are hindered by the lack of standardized terminology and evaluation metrics. Future research efforts should prioritize the development of a more harmonized and widely accepted taxonomy to enable more effective and comparable evaluations across different models and tasks, ultimately accelerating progress in addressing this critical issue.

## 4 Specific categories and manifestations of hallucinations

**This section details various specific types of hallucinations, including factual errors, contextual inconsistencies, and task-specific manifestations.**

---

Beyond the core intrinsic/extrinsic and factuality/faithfulness distinctions, LLM hallucinations manifest in numerous specific forms, often with distinct characteristics and implications (see summary in Table 2).

### 4.1 Factual errors and fabrications

This is a prevalent and particularly dangerous type of LLM hallucination, characterized by the generation of incorrect, misleading, or entirely fabricated factual content, frequently presented with a high degree of confidence. Such errors can appear as inaccuracies in historical information, scientific facts, or biographical details<sup>[14]</sup>.

#### 4.1.1 Incorrect facts

These are claims that directly oppose established and verified information<sup>[14;6]</sup>. An example is Google Bard’s hallucination claiming the James Webb Space Telescope took the first images of an exoplanet, despite NASA’s records indicating that earlier images existed. Other instances include the assertion that ”The Great Wall of China is visible from space” or the statement that ”Thomas Edison invented the internet”.

#### 4.1.2 Fabricated entities/information

This involves the invention of historical figures, events, or specific details that do not exist in reality. This can extend to creating entirely fictitious narratives, such as a claim about ”unicorns in Atlantis” being documented in 10,000 BC. In legal contexts, this type of hallucination can be particularly severe, involving the fabrication of information, including fake quotes and citations of non-existent court cases, leading to significant professional and legal consequences<sup>[72]</sup>. Similarly, in medical contexts, models may fabricate clinical details, invent research citations, or create made-up disease details, posing substantial risks to patient care<sup>[51;12]</sup>.

#### 4.1.3 Adversarial attacks

A specific subset of factual errors arises from adversarial attacks, where deliberately or inadvertently fabricated details embedded within user prompts lead the model to produce or elaborate on false information. This phenomenon can result in a ”garbage in, garbage out” problem, where erroneous inputs propagate misleading outputs, and also presents a threat of malicious misuse, where bad actors could exploit LLMs to spread falsehoods<sup>[51;99;108]</sup>.

### 4.2 Contextual inconsistencies

Contextual inconsistencies occur when the model’s output includes information not present in the provided context or directly contradicts it. This type of hallucination is often referred to as ”context divergence” or ”contextual misalignment” , indicating the model’s difficulty in

correctly attending to relevant context and instead relying on its internal generative tendencies. An example is when the model is given the context: "The Nile originates in Central Africa," but responds with: "The Nile originates in the mountain ranges of Central Africa," adding incorrect details not found in the original input<sup>[42;4;27]</sup>.

### 4.3 Instruction inconsistencies/deviation

Instruction inconsistencies refer to instances where the LLM ignores or fails to follow specific instructions provided by the user. The generated response, in these cases, does not adhere to the user's explicit directives. For example, if instructed to translate a question into Spanish, the model might instead provide the answer in English<sup>[101]</sup>.

### 4.4 Logical inconsistencies

Logical inconsistencies manifest when the model's output contains internal logical errors or contradictions, even if the initial part of the response is correct. This can appear as self-contradictory statements within the same output or across different interaction instances. This type of hallucination is related to "erroneous inference hallucination" and accounts for a notable portion, specifically 19%, of identified hallucination cases<sup>[42;47;34;95]</sup>. An example is an LLM performing an arithmetic operation incorrectly within a step-by-step mathematical solution, or stating a fact in one sentence and then providing a conflicting statement later in the same response.

### 4.5 Temporal disorientation

Temporal disorientation describes a type of hallucination involving issues with time-sensitive information, leading to the generation of outdated, anachronistic, or temporally incorrect facts. LLMs are particularly noted for struggling with "intricate temporal features" and out-of-distribution knowledge related to time. This category accounts for 12% of identified hallucination cases.<sup>[47;51]</sup> An illustrative example is an LLM incorrectly asserting that "Haruki Murakami won the Nobel Prize in Literature in 2016," when in fact, he has not won the Nobel Prize.

### 4.6 Ethical violations

Ethical violations refer to hallucinations that result in harmful, defamatory, or legally incorrect content. These instances can have severe real-world consequences, impacting individuals' reputations, causing financial losses, or leading to legal repercussions. Ethical violations represent 6% of hallucination cases in some analyses.<sup>[47;40;31]</sup>

#### 4.6.1 Defamation/misinformation

Examples include ChatGPT falsely claiming a university professor made sexually suggestive comments and attempted to touch a student, citing a non-existent article<sup>[42;18]</sup>. Another case involved ChatGPT falsely accusing a mayor of bribery and imprisonment, when he was actually a whistleblower.

### 4.6.2 Financial misinformation

An AI chatbot providing incorrect refund information to a customer, resulting in financial loss for both the customer and the airline, exemplifies how hallucinations can lead to tangible economic harm<sup>[102]</sup>.

### 4.6.3 Legal inaccuracies

LLMs can produce content that deviates from actual legal facts, well-established legal principles, or precedents. This includes generating "bogus judicial decisions, bogus quotes, and bogus internal citations". Such errors can lead to "representational harm," where the contributions of one member of the legal community are systematically erased or misattributed<sup>[51;19]</sup>.

## 4.7 Amalgamated hallucinations

Amalgamated hallucinations occur when the model incorrectly combines multiple facts or conditions presented within a single prompt. This happens when the LLM fails to properly integrate several distinct conditions, resulting in a blended output that erroneously merges disparate pieces of information<sup>[27;105]</sup>.

## 4.8 Nonsensical responses

Nonsensical responses are instances where LLMs generate output that is completely irrelevant to the input prompt. This type highlights the model's limitations in understanding context or maintaining a logical thread in a conversation, posing significant challenges in user interaction scenarios where clarity and relevance are paramount.<sup>[42]</sup> An example is a conversation about the NBA Commissioner where the LLM initially mentions "Adam Silver" but then randomly switches to "Stern" in the same response.

## 4.9 Task-specific hallucinations

Hallucinations can manifest uniquely depending on the specific generative task the LLM is performing.

### 4.9.1 Dialogue history-based hallucination

This occurs when an LLM mixes up names or relations of entities from the conversation history, or creates new incorrect inferences based on previous errors, leading to a "snowball effect" of distorted context. This arises because LLMs rely on pattern recognition and statistics, often lacking common sense or factual grounding in dialogue<sup>[100;26]</sup>.

### 4.9.2 Abstractive summarization hallucination

Systems designed for abstractive summarization can introduce errors or semantic transformations between the original and generated data, distorting or fabricating details, inferring unsupported causal relationships, or retrieving unrelated background knowledge. This is attributed to their reliance on pattern recognition rather than true comprehension of the source text<sup>[100;44;64]</sup>.

### 4.9.3 Generative question answering hallucination

In this context, the LLM makes an erroneous inference from its source information, leading to an incorrect answer, even when relevant source material is provided. The model may ignore evidence and make unjustified inferences based on its own prior knowledge<sup>[100;92]</sup>.

### 4.9.4 Code generation hallucination

When generating source code, LLMs can produce incorrect, nonsensical, or unjustifiable code that is difficult to identify and fix, especially under specific execution paths. This undermines the trustworthiness of generated code and can introduce significant risks and errors into codebases. Existing surveys classify these into input-conflicting, context-conflicting, and fact-conflicting types<sup>[57;2]</sup>.

### 4.9.5 Multimodal large language models hallucination

In multimodal large language models (MLLMs), hallucinations primarily focus on the "discrepancy between generated text response and provided visual content," a phenomenon known as cross-modal inconsistency<sup>[38;98]</sup>. Object hallucination in MLLMs is empirically categorized into three types:

- **Category:** identifies nonexistent or incorrect object categories in a given image<sup>[38;98]</sup>.
- **Attribute:** emphasizes incorrect descriptions of objects' attributes (e.g., color, shape, material)<sup>[38;98]</sup>.
- **Relation:** assesses incorrect relationships between objects<sup>[38;98]</sup>.

## 4.10 Complexities and critical implications of diverse hallucination types

The extensive list of specific hallucination types (factual, contextual, logical, temporal, ethical, amalgamated, nonsensical) and their distinct manifestations across various applications (dialogue, summarization, QA, code generation, multimodal) underscores that hallucination is not a singular, uniform error. Each type often arises from different underlying mechanisms<sup>[14;6;42;47;51;19;92;57;2;98]</sup>. For example, temporal errors might stem from outdated data, logical inconsistencies from reasoning flaws, and ethical violations from training biases. The detailed examples, particularly from the medical<sup>[51;12]</sup> and legal<sup>[51;19]</sup> domains, vividly illustrate that these are not merely academic curiosities but critical safety, reliability, and ethical issues with significant real-world repercussions, such as misleading clinicians, misinforming patients, legal sanctions, reputational damage, and financial loss. This granular understanding implies that a "one-size-fits-all" solution for hallucination is unlikely to be effective. Instead, research and development must adopt a highly granular and context-aware approach, tailoring detection, prevention, and mitigation strategies to the specific type of hallucination prevalent in a given application domain. This also highlights the urgent need for domain-specific benchmarks and evaluation frameworks to accurately assess and address these diverse forms of factual and contextual divergence.

Table 2: Comprehensive taxonomy of LLM hallucinations

Type	Definition/description	Example	Sources
<b>Intrinsic</b>	Contradicts provided input or context; internal inconsistencies.	Summary states birth year as 1980 then 1975.	[7;70]
<b>Extrinsic</b>	Not consistent with training data; introduces non-existent entities.	“The Parisian Tiger was hunted to extinction in 1885.”	[79;7]
<b>Factuality</b>	Contradicts real-world knowledge or verification sources.	“Charles Lindbergh was first to walk on the moon.”	[42;50;13]
<b>Faithfulness</b>	Diverges from input prompt or context.	Summary claims FDA rejected vaccine when article stated approval.	[64;96;61]
<b>Factual Errors</b>	Incorrect, misleading, or fabricated content.	Bard claiming JWST took first exoplanet images.	[14]
<b>Contextual</b>	Contradicts or adds to provided context.	Input: “Nile in Central Africa.” Output: “Nile in Central African mountains.”	[42;4;27]
<b>Instruction</b>	Fails to follow user instructions.	Translates question to Spanish but answers in English.	[101]
<b>Logical</b>	Internal logical errors or contradictions.	Incorrect arithmetic in step-by-step solution.	[42;47;34;95]
<b>Temporal</b>	Time-sensitive errors and anachronisms.	“Murakami won Nobel Prize in 2016.”	[47;51]
<b>Ethical</b>	Harmful, defamatory or legally incorrect content.	False accusation of professor with non-existent citation.	[47;40;31]
<b>Amalgamated</b>	Incorrectly combines multiple facts.	(Blending disparate information)	[27;105]
<b>Nonsensical</b>	Irrelevant responses lacking logic.	Switches from “Adam Silver” to “Stern” in NBA discussion.	[42]
<b>Code generation</b>	Incorrect or nonsensical source code.	Illogical code unfaithful to requirements.	[57;2]
<b>Multimodal</b>	Text-visual content discrepancies.	Identifying non-existent object in image.	[38;98]

## 5 Underlying causes of hallucinations

This section explores the complex factors contributing to hallucinations, stemming from training data, model architecture, and user prompts.



---

The diverse manifestations of LLM hallucinations stem from a complex interplay of factors originating from the training data, the model’s architecture and learning processes, and the nature of user prompts (see summary in Table 3) .

## **5.1 Data-related factors**

The quality and characteristics of the data used to train LLMs are fundamental determinants of hallucination frequency and type.

### **5.1.1 Quality and volume of training dataset**

The inherent quality and sheer volume of the dataset upon which an LLM is trained are crucial variables directly influencing the frequency and specific types of hallucinations produced. Flawed, incomplete, or noisy training data—containing errors, inconsistencies, or irrelevant information—significantly contributes to the generation of factually incorrect responses<sup>[42;32]</sup>.

### **5.1.2 Inadequate representation and biases**

If the data used to train LLMs lacks sufficient quality or diversity, the model may struggle to accurately understand the complexities and nuances of human language. Training on incorrect or biased data can lead to "imitative falsehoods," where the model replicates misinformation present in its training corpus<sup>[42;22]</sup>.

### **5.1.3 Outdated training data and knowledge boundary**

LLMs are prone to disseminating misinformation, particularly concerning frequently updated topics, primarily due to the static nature of their training data<sup>[17]</sup>. The absence of up-to-date facts leads to inherent limitations in specialized domains. A critical issue is the model’s inability to recognize its own knowledge boundaries, which often results in it confidently generating content beyond its learned scope<sup>[81]</sup>.

### **5.1.4 Source-reference divergence**

In certain datasets, such as those specifically curated for summarization tasks, summaries might contain additional, unsupported claims that diverge from the original source references, directly contributing to hallucination<sup>[10]</sup>.

## **5.2 Model-related factors**

The internal design, training processes, and inference mechanisms of LLMs contribute significantly to their propensity for hallucination.

### **5.2.1 Auto-regressive nature**

A fundamental cause of hallucinations arises from the very design principle of certain LLMs: their auto-regressive nature. These models are programmed to produce output based on token prediction, meaning they predict the most probable next token(s) given the preceding sequence. Factual accuracy is not the direct, explicit goal of this process; rather, accuracy is inferred from



a high probability of adequate token prediction based on the training data. Since training datasets are necessarily flawed or incomplete, the probabilistic nature of this generation can lead to hallucinations<sup>[50;38]</sup>.

### 5.2.2 Architecture flaws and internal design

The internal design of LLMs inherently predisposes them to generating hallucinations. For instance, unidirectional representation in certain architectures can limit contextual understanding, leading to the generation of one-sided or biased narratives<sup>[38]</sup>.

### 5.2.3 Training processes

- **Exposure bias:** a discrepancy between the conditions encountered during training and those during inference can cause cascading errors during text generation.<sup>[78]</sup>
- **Capability misalignment:** when LLMs are aligned with capabilities that extend beyond what their training data adequately supports, they may produce errors, particularly fabricated facts in specialized domains where their knowledge is insufficient<sup>[59]</sup>.
- **Belief misalignment:** the generated outputs may diverge from the LLM’s internal “beliefs” or learned representations, leading to inaccuracies. This can sometimes be a result of the model “pandering” to user opinions rather than adhering to factual truth<sup>[43]</sup>.
- **Over-optimization for specific objectives:** over-optimization during the training phase for certain performance metrics can inadvertently increase hallucination rates in other areas<sup>[55]</sup>.

### 5.2.4 Decoding strategies

- **Stochastic nature/inherent sampling randomness:** LLMs employ sampling strategies during text generation that introduce an element of randomness into the output. A high “temperature” setting, for example, can enhance creativity but also significantly elevate the risk of hallucination by favoring the selection of low-probability or unexpected tokens<sup>[38]</sup>.
- **Imperfect decoding representation:** issues such as over-reliance on partially generated content and the “softmax bottleneck” can lead to faithfulness errors, where the output deviates from the intended meaning or context<sup>[15]</sup>.

### 5.2.5 Overconfidence and calibration

LLMs frequently exhibit overconfidence, generating outputs with high certainty even when the underlying information is incorrect. Poor calibration, where confidence scores do not accurately reflect prediction accuracy, can mislead users, particularly clinicians in medical contexts, into trusting inaccurate outputs, posing significant risks<sup>[20]</sup>.

### 5.2.6 Generalization to unseen cases

LLMs may struggle to generalize effectively beyond their training data, especially when confronted with rare diseases, novel treatments, or atypical clinical presentations. In such sce-

narios, models might extrapolate from unrelated patterns, producing erroneous or irrelevant outputs<sup>[55;42;22;32]</sup>.

### **5.2.7 Lack of reasoning and nuanced language understanding**

LLMs primarily rely on statistical correlations learned from vast amounts of text rather than true causal or logical reasoning. This can lead to outputs that sound plausible but lack logical coherence. Furthermore, they struggle with interpreting the subtleties of human language, including irony, sarcasm, and cultural references, which can result in outdated or irrelevant information when nuance is key. A lack of logical reasoning capabilities is specifically identified as a significant contributor to fact-conflicting hallucinations<sup>[42;41]</sup>.

### **5.2.8 Knowledge overshadowing**

This occurs when certain aspects of a prompt disproportionately dominate the model’s attention, leading to an overgeneralization of dominant conditions or patterns. This phenomenon is partly attributable to imbalances within the training data<sup>[106]</sup>.

### **5.2.9 Insufficient knowledge representation**

Hallucinations can also arise from deficiencies in the model’s internal knowledge representation, particularly within the lower layers of its neural network. These deficiencies result in what are termed “knowledge enrichment hallucinations,” where the model generates unsupported information due to gaps in its subject-specific knowledge<sup>[30]</sup>.

### **5.2.10 Failure in information extraction**

Inaccurate extraction of relevant attributes or details by the model’s higher-layer attention mechanisms can lead to “answer extraction hallucinations.” This underscores the critical importance of precise information retrieval and application in generating correct outputs<sup>[105]</sup>.

## **5.3 Prompt-related factors**

The way users interact with LLMs through prompts can also induce or exacerbate hallucinations.

### **5.3.1 Adversarial attacks**

Deliberate or inadvertent fabrications embedded in user prompts can trigger hallucinations, as LLMs may elaborate on the false information. This creates a “garbage in, garbage out” problem, where erroneous inputs produce misleading outputs, and also poses a threat of malicious misuse<sup>[99]</sup>.

### **5.3.2 Overly confirmatory tendency**

Some LLMs exhibit an overly confirmatory tendency, sometimes prioritizing a persuasive or confident style over factual accuracy. This characteristic can exacerbate the impact of prompt-based fabrications, making the hallucinated content appear more credible<sup>[69]</sup>.

### 5.3.3 Prompting methods

The specific methods and clarity of prompting can significantly influence hallucination rates. Clearer, more restrictive prompts and providing relevant in-context learning examples (e.g., few-shot learning) can help reduce hallucinations by guiding the model more precisely<sup>[38]</sup>.

## 5.4 An emergent property requiring systemic solutions

The comprehensive list of causes, spanning data quality, model architecture, training processes, and inference mechanisms, reveals that hallucination is not a simple bug but an emergent property of the current LLM design paradigm. The auto-regressive nature<sup>[50;38]</sup> fundamentally prioritizes generating plausible token sequences based on statistical patterns rather than ensuring factual accuracy or logical coherence. This statistical reliance, combined with "black box reasoning" and inherent "overconfidence"<sup>[20]</sup>, creates a scenario where models confidently produce incorrect information. The struggle with "logical reasoning"<sup>[42;41]</sup> and "generalization to unseen cases"<sup>[55;42;22;32]</sup> points to a deeper limitation beyond mere memorization; LLMs currently lack true comprehension and causal understanding. Furthermore, the vulnerability to "adversarial attacks" and the "garbage in, garbage out" problem<sup>[99]</sup> highlight the fragility of these systems to external inputs. Effectively addressing hallucination, therefore, requires a multi-pronged research agenda that goes beyond superficial fixes like data cleaning or simple fine-tuning. It necessitates fundamental advancements in model architectures to incorporate stronger symbolic reasoning capabilities, better uncertainty quantification, and more robust grounding mechanisms, such as advanced retrieval-augmented generation (RAG) techniques. The theoretical inevitability of hallucination<sup>[7]</sup> further reinforces that some level of hallucination might always persist, making external safeguards, robust evaluation, and continuous human oversight crucial for the safe and reliable deployment of LLMs in critical applications.

Table 3: Root Causes of LLM Hallucinations

Category	Specific Factor	Explanation	Sources
<b>Data</b>	Training Data Quality	Flawed, incomplete, or noisy data leads to incorrect responses.	[42;32]
	Data Biases	Lack of diversity causes imitative falsehoods.	[42;22]
	Outdated Data	Static data causes misinformation on dynamic topics.	[17;81]
	Source Divergence	Summaries containing unsupported claims.	[10]
<b>Model</b>	Auto-Regressive Nature	Token prediction prioritizes probability over accuracy.	[50;38]
	Architecture Flaws	Design predisposes models to hallucinate.	[38]
	Exposure Bias	Training-inference discrepancy causes errors.	[78]
	Capability Misalignment	Fabrication in specialized domains.	[59]
	Belief Misalignment	Outputs diverge from internal representations.	[43]
	Over-optimization	Focus on metrics increases hallucinations.	[55]
	Sampling Randomness	High temperature introduces inaccuracies.	[38]
	Decoding Issues	Over-reliance on partial generation.	[15]
	Overconfidence	High certainty for incorrect outputs.	[20]
	Generalization Failure	Errors on rare/novel cases.	[55;42;22;32]
	Reasoning Limits	Statistical over causal reasoning.	[42;41]
	Knowledge Over-shadowing	Prompt aspects dominate attention.	[106]
	Knowledge Gaps	Deficient internal representations.	[30]
	Extraction Failure	Inaccurate attention mechanisms.	[105]
<b>Prompt</b>	Adversarial Attacks	Fabricated prompt details.	[99]
	Confirmatory Bias	Persuasive style over facts.	[69]
	Poor Prompting	Unclear structure increases errors.	[38]

---

## 6 Cognitive and human factors in hallucination perception

This section explores how human trust, cognitive biases, and interaction design influence the perception and impact of LLM hallucinations, emphasizing the need for user-centered mitigation strategies and human-in-the-loop oversight.

---

In addition to technical causes, the real-world impact of hallucinations is strongly shaped by how humans interpret, trust, and respond to language model outputs. Research in human-computer interaction (HCI), psychology, and decision science indicates that users are not passive consumers of information—they bring cognitive biases, heuristics, and trust dynamics into their interactions with LLMs. These factors influence whether hallucinations are detected, ignored, or acted upon.

### 6.1 User trust and interpretability

LLMs often produce fluent, well-structured, and grammatically correct responses, which are commonly interpreted by users as signals of credibility—even when the content is factually incorrect. This “*fluency heuristic*” has been observed to increase perceived accuracy of statements simply due to linguistic polish<sup>[80]</sup>.

Moreover, large-scale studies have shown that users tend to assign high trust to AI outputs, particularly when models present information confidently or with detailed elaboration<sup>[9]</sup>. For instance, Bubeck et al.<sup>[11]</sup> found that users often rated GPT-4’s incorrect answers as more convincing than correct ones from human experts in blind evaluations.

### 6.2 Hallucinations often go unnoticed

Because hallucinations are often contextually plausible and stylistically convincing, users—especially non-experts—may struggle to identify falsehoods without access to external verification tools. This risk is particularly high in areas like health, law, or finance, where subtle distortions can have serious consequences<sup>[51;19]</sup>.

Empirical studies by Luger & Sellen<sup>[60]</sup> reveal that users often accept AI-generated outputs at face value and fail to notice hallucinations—particularly when responses appear fluent and confident—unless they are explicitly instructed to fact-check, indicating a widespread tendency to overtrust AI systems as reliable or authoritative sources.

### 6.3 Cognitive biases amplifying hallucination risks

Several well-established cognitive biases contribute to the tendency to overlook or accept hallucinated content.

#### 6.3.1 Automation bias

This refers to the human tendency to over-rely on automated systems, assuming their outputs are accurate—even when they are not. In the context of LLMs, users may accept incorrect or

hallucinated information simply because it comes from an AI, especially in situations involving time pressure, cognitive overload, or lack of expertise. This bias can lead users to overlook obvious errors or fail to cross-check information they would otherwise question<sup>[21]</sup>.

### **6.3.2 Confirmation bias**

This bias describes the tendency to favor information that confirms one’s pre-existing beliefs or expectations, while dismissing or overlooking contradictory evidence. When interacting with LLMs, users may be more likely to accept hallucinated content if it aligns with what they already believe or want to be true, making them less likely to scrutinize its accuracy<sup>[67;71]</sup>.

### **6.3.3 Illusion of explanatory depth**

This cognitive bias occurs when individuals believe they understand complex topics more deeply than they actually do. As a result, they may overestimate their ability to evaluate the accuracy of AI-generated content. When an LLM produces a plausible-sounding explanation or summary, users may assume it is correct without fully understanding or verifying the underlying concepts, increasing the risk of accepting hallucinated information<sup>[83;65]</sup>.

### **6.3.4 Persistence of biases despite warnings**

Research shows that cognitive biases such as automation bias, confirmation bias, and the illusion of explanatory depth can persist even when users are explicitly informed that AI systems may produce errors<sup>[25]</sup>. In experimental settings, users who were told that a decision-support system was fallible still tended to trust its outputs over their own judgment. This suggests that merely warning users about possible inaccuracies is often insufficient to prevent overreliance. In the context of LLMs, this means that even transparent disclaimers or uncertainty indicators may not fully mitigate the undue influence of confident but hallucinated outputs, especially when users lack the domain expertise or motivation to verify them independently.

## **6.4 Design implications and mitigation strategies**

These findings suggest that hallucination mitigation is not solely a model-centric challenge, but also a user interface and interaction design problem. To improve user resilience against hallucinations, several strategies have been proposed.

### **6.4.1 Calibrated uncertainty displays**

Providing users with visual or textual indicators of a model’s confidence—such as probability scores, uncertainty ranges, or qualitative labels (e.g., “highly confident,” “low certainty”)—can help them better judge the reliability of AI outputs<sup>[54]</sup>. These displays are especially valuable in tasks like question answering or medical advice, where the perceived confidence of a model often influences user trust. When confidence is misaligned with correctness (e.g., high confidence in a hallucinated answer), users may be misled unless the interface communicates epistemic uncertainty clearly. Calibrated uncertainty helps users decide when additional verification is necessary and supports a more cautious interpretation of potentially hallucinated content.

### 6.4.2 Source-grounding indicators

Clearly linking parts of the model’s output to supporting evidence from external sources—such as documents retrieved through a RAG system—can reduce blind acceptance of hallucinated facts<sup>[103]</sup>. Visual markers, citations, or tooltips that explain which parts of a response are grounded in specific documents enhance transparency and user understanding. By making the boundary between supported and unsupported content explicit, source-grounding indicators help users identify which claims are verifiable and which may be speculative or invented, thus mitigating the impact of hallucinations in high-stakes applications.

### 6.4.3 Justification prompts

Designing systems that encourage users to ask reflective questions like “Why is this the answer?” or “How do you know that?” promotes more critical evaluation of LLM responses<sup>[29]</sup>. These prompts can be implemented through interface design (e.g., buttons or suggested queries) or integrated into conversational flows. Encouraging justification-seeking behavior not only increases user awareness of potential inaccuracies but also reinforces an epistemic mindset in which outputs are evaluated based on evidence and reasoning rather than surface plausibility. This can be especially helpful in educational or decision-support contexts where understanding the rationale behind a response is as important as the response itself.

### 6.4.4 Factuality-aware interface prototypes

Recent research has produced interface prototypes—such as Med-PaLM 2<sup>[91]</sup>—that integrate design features aimed at improving interpretability and factual reliability. For instance, Med-PaLM 2 provides clinical references and confidence levels in its medical responses, demonstrating how multimodal transparency cues—combining visual, textual, and interactive elements—can enhance user awareness of potential hallucinations, promote responsible usage, and support informed decision-making, particularly in high-stakes domains like healthcare and public safety.

## 6.5 Human-in-the-loop evaluation and oversight

Ultimately, hallucination detection and management must be seen as a joint cognitive task between the LLM and its user. Evaluation frameworks should therefore include human factors—such as susceptibility to bias, trust calibration, and verification behavior—as part of their assessment.

This aligns with broader calls in AI safety and responsible AI literature for systems that are not just high-performing in benchmarks, but usable, trustworthy, and robust under real-world conditions<sup>[35;24]</sup>.

---

## 7 Evaluation benchmarks and metrics for hallucination detection

This section surveys the principal benchmarks and evaluation metrics developed to detect and quantify hallucinations in LLMs, highlighting current methodologies, their limitations, and the need for unified, taxonomy-aware assessment frameworks.

---

The effective detection and quantification of hallucinations in LLMs is a prerequisite for both empirical research and practical deployment. While considerable progress has been made in identifying hallucination types and underlying causes, the evaluation of hallucination remains a challenging and evolving area. This section presents a brief survey of the most prominent benchmarks and metrics used to assess hallucination in LLM outputs, alongside a discussion of their limitations and future directions.

## 7.1 Benchmark datasets

Several benchmarks have been developed to systematically evaluate hallucinations across diverse tasks and domains. These datasets vary in scope, annotation methodology, and underlying definition of hallucination.<sup>[53;7;66;37;87]</sup>

### 7.1.1 TruthfulQA

Is a benchmark composed of adversarially constructed questions that intentionally target common misconceptions, false beliefs, or ambiguities in general knowledge. Unlike traditional benchmarks that evaluate correctness in terms of expected facts, TruthfulQA emphasizes a model’s robustness against confidently generating plausible-sounding but factually incorrect statements. It is task-agnostic and domain-general, designed to test whether language models can distinguish between fact and fiction in open-domain question answering. The benchmark includes both multiple-choice and free-form response settings and is annotated with human-verified judgments to assess truthfulness, informativeness, and consistency.<sup>[53]</sup>

### 7.1.2 HalluLens

Is a comprehensive benchmark that systematically maps hallucination instances to an explicit taxonomy encompassing multiple dimensions: factual, ethical, logical, temporal, and task-specific hallucinations. Unlike task-bound benchmarks, HalluLens is designed to evaluate hallucinations across a wide range of contexts and generation types, including summarization, question answering, dialogue, and instruction following. Each instance is annotated with detailed metadata specifying the hallucination category, severity, and grounding status. This makes it particularly suitable for fine-grained, taxonomy-aware evaluation and enables rigorous cross-model comparison aligned with theoretical frameworks. HalluLens serves as both a diagnostic and comparative tool, helping researchers and developers identify model-specific hallucination patterns.<sup>[7]</sup>

### 7.1.3 FActScore

FActScore is a benchmark specifically designed to evaluate the factual consistency of outputs in summarization tasks. Rather than relying on surface-level similarity metrics such as ROUGE or BLEU, FActScore employs entailment-based classifiers that have been fine-tuned to determine whether a generated sentence can be logically inferred from the corresponding reference source document. This allows it to detect subtle hallucinations, such as fabricated relationships or omitted qualifiers, which might not be flagged by traditional overlap-based metrics. By providing span-level and sentence-level assessments, FActScore supports a more granular and semantically precise evaluation of summary fidelity.<sup>[66]</sup>



#### 7.1.4 Q2 and QuestEval

These metrics adopt an indirect yet powerful approach to evaluating factual consistency through question generation and answering. **Q2** (Quality Questioning) generates a set of questions based on the system output and then uses the source document to answer them. If the answers from the source align with those implied by the generated summary, the output is considered faithful.<sup>[37]</sup> **QuestEval**, similarly, computes consistency by comparing answers to questions generated from both the candidate and reference texts. These methods do not rely on static reference texts but instead treat the source content as a dynamic knowledge base, allowing for flexible and contextual evaluation of hallucination. Their strength lies in capturing factual divergences that traditional string-matching metrics often overlook.<sup>[87]</sup>

#### 7.1.5 Domain-specific benchmarks

Benchmarks developed to evaluate hallucinations in specialized and high-risk applications such as medicine, software engineering, and multimodal reasoning.<sup>[74;109;73;94;88]</sup>

- **MedHallu** this is a comprehensive benchmark specifically designed for detecting medical hallucinations in LLMs. It comprises 10,000 high-quality question-answer pairs derived from PubMedQA, with hallucinated answers systematically generated through a controlled pipeline.<sup>[74]</sup>
- **MedHallBench** another recent and comprehensive benchmark framework for evaluating and mitigating hallucinations in Medical LLMs. It integrates expert-validated medical case scenarios with established medical databases.<sup>[109]</sup>
- **Med-HALT** this benchmark proposes a two-tiered approach to evaluate the presence and impact of hallucinations in biomedical-generated LLM outputs. It includes Reasoning Hallucination Tests (RHTs) like False Confidence Test (FCT), None of the Above (Nota) Test, and Fake Questions Test (FQT), as well as Memory Hallucination Tests (MHTs) such as Abstract-to-Link and PMID-to-Title tests.<sup>[73]</sup>
- **CodeHaluEval** targets the evaluation of hallucinations in code-generating language models (CodeLLMs). It includes programming tasks and ground truth outputs annotated for syntactic validity, semantic correctness, and adherence to functional requirements. The benchmark categorizes hallucinations into input-conflicting, context-conflicting, and fact-conflicting errors. It plays a crucial role in identifying code that may appear plausible but fails to execute correctly or violates specifications—risks that can lead to software bugs, security vulnerabilities, or production failures.<sup>[94]</sup>
- **HALLUCINOGEN benchmark** is a novel and comprehensive Visual Question Answering (VQA) benchmark specifically designed to evaluate and identify "object hallucination" in Large Vision-Language Models (LVLMs). Unlike previous benchmarks that often rely on simple queries, HALLUCINOGEN introduces a diverse set of "object hallucination attacks" through complex contextual reasoning prompts. These prompts are crafted to challenge LVLMs by asking about visual objects that may or may not be present in an image, forcing the models to accurately identify, locate, or perform visual reasoning around specific objects, thereby exposing instances where they fabricate or misclassify objects.<sup>[88]</sup>

These domain-specific benchmarks are indispensable for the safe evaluation of LLM performance

in contexts where hallucinations may lead to misdiagnosis, faulty software behavior, or visual misinterpretations, thus bridging the gap between generic metrics and task-critical assessment.

## 7.2 Quantitative metrics

Metrics used to evaluate hallucination are typically classified according to the type of alignment they measure—faithfulness to input, factuality with respect to external knowledge, or semantic consistency.

### 7.2.1 Faithfulness metrics

These metrics assess whether the generated output remains consistent with the provided input or prompt:

- **ROUGE, BLEU, and BERTScore:** these metrics are primarily surface-level or embedding-based similarity metrics. They evaluate the generated text by comparing it to a source or reference text, assessing shared words, phrases, or underlying semantic representations.
  - **ROUGE (Recall-Oriented Understudy for Gisting Evaluation):** this metric is widely used in summarization to measure the overlap of n-grams (sequences of words) between the generated text and a reference. For example, ROUGE-1 assesses unigram overlap, ROUGE-2 measures bigram overlap, and ROUGE-L identifies the longest common subsequence. Its "recall-oriented" nature emphasizes how much of the reference's information is captured in the generated output<sup>[52]</sup>.
  - **BLEU (Bilingual Evaluation Understudy):** originally developed for machine translation, BLEU quantifies the precision of n-grams in the generated text compared to a reference, penalizing for brevity. It focuses on the extent to which the generated text's content is also present in the reference<sup>[75]</sup>.
  - **BERTScore:** a more advanced metric, BERTScore employs contextual embeddings from large language models (such as BERT) to measure the semantic similarity between words and sentences in the generated text and the reference. This capability allows it to identify paraphrases or synonyms even when exact word matches are absent<sup>[104]</sup>.
  - **Limitations:** while useful for initial assessments, these metrics are often insufficient for detecting nuanced semantic errors or fabrications. High scores on ROUGE, BLEU, or BERTScore do not guarantee factual accuracy, as a text can exhibit high lexical or semantic similarity while still containing subtle inconsistencies or outright hallucinations. They do not inherently assess the truthfulness of a statement, only its textual similarity<sup>[52;75;104]</sup>.
- **FactCC:** stands for "Fact-Checking with Contextualized Commonsense," is a specialized metric developed specifically for detecting hallucinations, particularly within summarization tasks. Unlike simpler similarity metrics, FactCC employs a trained classifier. This classifier learns to identify factual inconsistencies by being trained on datasets of summaries paired with their source texts, where human annotators have meticulously labeled instances of factual inconsistencies. This methodology allows the classifier to recognize patterns that indicate unfaithfulness. Its primary advantage lies in its improved

precision in hallucination detection; because it’s explicitly trained to identify inconsistencies, FactCC is more effective at catching factual errors than metrics that rely solely on surface-level comparisons, aiming to determine if a generated statement truly contradicts the original source material<sup>[44]</sup>.

- **SummaC**: is a metric that assesses factual consistency by leveraging the principles of Natural Language Inference (NLI). NLI is a core task in natural language processing where a model determines the logical relationship between two sentences: whether one sentence (the "hypothesis") is entailed by, contradicts, or is neutral with respect to another sentence (the "premise"). In SummaC’s application, segments of the source text serve as the premise, while sentences from the AI-generated output (such as a summary) act as the hypothesis. An NLI model then evaluates if each statement in the generated output is entailed by the source text, signifying factual consistency. Conversely, statements that contradict the source or are not supported by it may indicate a hallucination or an unverified claim. SummaC’s strength lies in its ability to model factual consistency through these entailment relationships, which closely align with how humans judge factual accuracy, offering a more robust assessment of faithfulness compared to methods based on simpler lexical or embedding similarity<sup>[45]</sup>.

### 7.2.2 Factuality metrics

These metrics measure the alignment of generated content with real-world facts or structured knowledge sources, moving beyond mere consistency with input prompts to verify external veracity.

- **Knowledge Intensive Language Tasks (KILT)**: is a benchmark designed to evaluate the factual accuracy and knowledge-groundedness of language models. It specifically assesses the ability of generated content to align with factual information found in structured knowledge sources. The core methodology involves linking entities and claims present in the generated text to corresponding entries and facts within established knowledge bases, such as Wikipedia. This approach allows for a direct verification of whether the model’s output reflects verifiable real-world knowledge rather than merely being coherent or consistent with an initial prompt. KILT’s tasks often require models to generate text that can be directly verified against these external knowledge sources, making it a robust measure of a model’s factual grounding<sup>[77]</sup>.
- **Retrieval-Augmented Evaluation (RAE)**: is a methodology used to assess the factual grounding of generated claims, particularly within Retrieval-Augmented Generation (RAG) pipelines. RAE operates by evaluating whether the evidence retrieved by a RAG system genuinely supports the claims made in the generated output. The process typically involves two main steps: first, identifying the claims made by the language model, and second, verifying these claims against the specific knowledge snippets or documents that the retrieval component of the RAG system provided as grounding evidence. This metric offers a scalable and efficient way to judge the factual accuracy and support of generated text, as it directly checks the consistency between the model’s output and its purported factual basis derived from the retrieval step. RAE is crucial for ensuring that RAG models do not hallucinate information, even when provided with relevant context<sup>[48;84;28;85]</sup>.

### 7.2.3 Human evaluation

Despite the proliferation of automated metrics, human evaluation remains the most reliable and widely accepted method for hallucination detection. Annotators typically assess outputs based on criteria such as:

- **Correctness:** this criterion assesses whether the generated content aligns with verifiable real-world facts. An output is considered correct if it can be independently validated against trusted knowledge sources. Inaccuracies, fabrications, or distortions of known information constitute a lack of correctness<sup>[33;58]</sup>.
- **Faithfulness:** measures the extent to which the model’s output remains consistent with the input prompt or source material. An output is unfaithful if it introduces information not present in the input, omits critical elements, or misrepresents the source. Faithfulness is especially important in summarization, translation, and question-answering tasks<sup>[64;16]</sup>.
- **Coherence:** refers to the logical consistency and internal structure of the output. A coherent response maintains a stable topic, avoids contradictions within itself, and follows a clear and understandable flow of reasoning. Incoherent outputs may contain abrupt topic shifts, self-contradictions, or illogical argumentation<sup>[76;8;62]</sup>.
- **Harmfulness or bias:** this dimension captures whether the output contains content that could be ethically problematic, offensive, or unsafe. This includes outputs that propagate harmful stereotypes, generate defamatory claims, or offer misleading information in domains like medicine, law, or finance. Special attention is needed in safety-critical applications where biased or harmful content could have serious real-world consequences<sup>[5;107;93;56]</sup>.

However, human evaluation is costly, time-consuming, and often subject to inter-rater variability, underscoring the need for more robust and interpretable automatic metrics.

## 7.3 Limitations and open challenges

Despite advancements in automated evaluation, current benchmarks and metrics for hallucination detection in AI-generated content face several persistent limitations that impede comprehensive and comparable assessments.

### 7.3.1 Lack of standardization

A significant challenge is the absence of a universally accepted definition of hallucination. Various benchmarks and research studies adopt differing conceptualizations, leading to inconsistencies in how hallucinations are annotated and measured. This definitional variability makes it exceedingly difficult to conduct fair and meaningful comparisons of hallucination rates and detection capabilities across different models, datasets, or evaluation frameworks. The absence of a shared understanding and operationalization of "hallucination" hinders the development of generalizable solutions and a cumulative scientific discourse.

### 7.3.2 Task dependence

The effectiveness of current metrics is often highly dependent on the specific natural language processing task being evaluated. Metrics that might demonstrate reasonable performance in

detecting hallucinations within summarization tasks, for instance, frequently fail to generalize or perform adequately in other domains such as question answering (QA), dialogue generation, or code generation. This limitation arises because the nature and manifestation of hallucinations can vary significantly across tasks. What constitutes a hallucination in a factual summary (e.g., inventing a detail) might differ from an unfaithful response in a dialogue system (e.g., contradicting prior turns) or an incorrect function in code generation. This task-specific performance necessitates the development of specialized metrics for each application, increasing complexity and fragmentation in the evaluation landscape.

### 7.3.3 Insensitivity to subtle hallucinations

Many existing metrics, particularly those relying on surface-level textual similarity or basic factual checks, are inherently unable to detect more nuanced and insidious forms of hallucination. These can include low-level factual shifts (slight alterations to numerical values or dates), subtle inferential errors (drawing an incorrect conclusion from otherwise correct premises), or context-dependent misalignments where a statement might be technically plausible but factually incorrect given the specific context of the input. Such subtle hallucinations are challenging to identify automatically, as they often require deep semantic understanding, complex logical reasoning, or access to vast external knowledge bases, making them particularly deceptive and hard to mitigate.

### 7.3.4 Limited grounding and explainability

A critical drawback of most automatic hallucination detection scores is their lack of interpretability and diagnostic value. These metrics typically provide a numerical score indicating the presence or absence of hallucination but offer little to no insight into why a particular output is deemed hallucinated. This limited grounding means that developers receive minimal actionable feedback on the specific type of error, the source of the factual deviation, or the exact portion of the generated text responsible for the hallucination. Without this granular insight, debugging models, understanding their failure modes, and implementing targeted improvements for hallucination reduction become significantly more challenging and less efficient. The lack of explainability impedes effective model development and refinement.

## 7.4 Toward unified evaluation frameworks

Future progress depends on the development of comprehensive, taxonomy-aware, and domain-adapted evaluation frameworks that:

- Incorporate **multi-level evaluation**, combining surface-level similarity with logic- and knowledge-aware assessment;
- Leverage **retrieval-based and symbolic tools** to enhance grounding;
- Standardize annotation protocols and metrics across tasks;
- Integrate **model uncertainty and confidence calibration** into evaluation.

Ultimately, the path to robust hallucination mitigation must be rooted in rigorous, context-sensitive measurement. Without accurate and scalable evaluation tools, efforts to reduce hallucination risk in real-world applications will remain incomplete and difficult to validate.

## 8 Hallucination mitigation strategies

**This section surveys both architectural and systemic approaches to mitigating hallucinations in LLMs, including tool augmentation, retrieval grounding, fine-tuning, symbolic guardrails, and user-facing strategies such as uncertainty displays and fallback mechanisms.**

---

Given the theoretical inevitability of hallucinations in LLMs<sup>[89]</sup>, researchers and developers have proposed a range of mitigation strategies. These can be broadly categorized into two groups: architectural strategies, which modify how the model itself is trained or behaves during inference, and systemic strategies, which shape how the model is embedded, controlled, or interpreted within a broader application context. Both are necessary for creating robust, trustworthy systems that minimize the frequency and harm of hallucinations.

### 8.1 Architectural mitigation strategies

Architectural strategies operate at the model level and seek to reduce hallucination risk by directly improving the model’s grounding, reasoning, or factual alignment capabilities. These interventions are typically implemented through changes in training data, model design, or auxiliary components used at inference time.

#### 8.1.1 Toolformer-style augmentation

Recent advances in tool-augmented LLMs propose allowing the model to call external APIs, calculators, code interpreters, or structured knowledge tools during inference. Toolformer<sup>[86]</sup>, for example, fine-tunes an LLM to decide when and how to use external tools to answer questions more reliably. Instead of relying purely on parametric memory, the model delegates sub-tasks—such as date calculations, currency conversions, or fact retrieval—to external systems better equipped to handle them.

This approach offloads fact-intensive or computation-heavy tasks to specialized modules, significantly reducing hallucination risk in those areas. The model learns to invoke tools autonomously during generation, producing more grounded and verifiable responses while maintaining fluency.

#### 8.1.2 Factual grounding through retrieval mechanisms

(*RAG*) is one of the most widely adopted hallucination mitigation frameworks. In RAG systems<sup>[48;39]</sup>, the model is paired with a retrieval component that fetches relevant documents or knowledge snippets from a curated corpus (e.g., Wikipedia, academic papers, or enterprise databases) in response to a user query. These retrieved documents are then passed as additional input context to the LLM, grounding the generated response in verifiable sources.

RAG reduces the likelihood of hallucination by:

- Constraining generation to content retrieved from external knowledge bases.
- Allowing the model to quote, summarize, or paraphrase from known references.
- Providing users with transparency and traceability via document citations.

Prominent RAG implementations include Google’s Bard, Meta’s BlenderBot 3, and enterprise systems like Microsoft’s Copilot and Amazon Bedrock. Despite its advantages, RAG is not foolproof: the model may still hallucinate if it fails to properly interpret or align with the retrieved material<sup>[90]</sup>.

### 8.1.3 Fine-tuning with synthetic or adversarially filtered data

Another mitigation strategy involves fine-tuning LLMs on curated or synthetic datasets designed to reduce hallucination tendencies. Two prominent approaches include:

- **Synthetic factuality tuning:** Models are trained or fine-tuned on large corpora of verified, well-grounded question-answer pairs. These may be created through human annotation or automatically generated and filtered using factual consistency metrics<sup>[36]</sup>.
- **Adversarial filtering:** Using hallucination detection models or adversarial prompts to identify and filter out outputs that exhibit hallucination. These filtered outputs can be used to refine the LLM or train classifier modules that flag likely hallucinated content<sup>[46]</sup>.

Although effective in reducing hallucinations on benchmark tasks, these methods face limitations in scalability, domain generalizability, and susceptibility to dataset bias.

## 8.2 Systemic mitigation strategies

Systemic strategies are applied at the deployment or user interface level and focus on shaping how LLM outputs are interpreted, controlled, or constrained in real-world contexts. These strategies often complement architectural solutions by providing guardrails and transparency mechanisms to reduce the risk and impact of undetected hallucinations.

### 8.2.1 Guardrails and symbolic integration

Guardrails are rule-based or symbolic control mechanisms that constrain the behavior of LLMs during inference. These include:

- **Logic validators:** these components evaluate whether an LLM’s output is internally consistent or conforms to formal rules in a given domain—such as mathematics, programming, or natural language logic. For example, in arithmetic tasks, a logic validator can compare the model’s answer against a rule-based calculator. In legal or contractual reasoning, outputs can be assessed for compliance with regulatory clauses or logical structures. By acting as a correctness gatekeeper, logic validators help identify outputs that may be fluent but logically invalid<sup>[3]</sup>.
- **Factual filters:** factual filtering involves post-processing model outputs to detect contradictions or inconsistencies with a trusted external source, such as a structured database or knowledge graph (e.g., Wikidata, UMLS, or proprietary enterprise data). These systems can match generated claims against canonical facts and either flag inaccuracies or attempt automatic correction. For instance, if a model claims that “Paris is the capital of Germany,” a factual filter could detect the mismatch and suggest a correction based on structured geopolitical data<sup>[23]</sup>. Such filters are particularly valuable in domains where factual consistency is non-negotiable, like medicine, finance, and policy generation.



- **Rule-based fallbacks:** in scenarios where uncertainty is high or outputs are flagged as potentially hallucinated—either by a validator, confidence threshold, or user feedback—the system can execute predefined fallback policies. These include refusing to answer (e.g., “I’m not confident enough to provide a reliable response”), rerouting the request to a human-in-the-loop, or prompting the user for clarification. Rule-based fallbacks act as safety valves, especially in high-stakes contexts, by enforcing cautious behavior when confidence or factuality cannot be guaranteed. They are also used in frameworks such as Reinforcement Learning with Human Feedback (RLHF), where such flags inform future model training<sup>[49]</sup>.

Symbolic integration—where models are combined with deterministic reasoning systems—represents a promising frontier for hallucination mitigation. Neuro-symbolic systems, for example, blend statistical generation with formal logic, enabling models to verify or revise outputs before presentation<sup>[63]</sup>.

### 8.3 Toward hybrid and context-aware mitigation systems

As no single technique fully eliminates hallucinations across all tasks and domains, the most promising direction lies in the development of hybrid mitigation systems—architectures that combine multiple complementary strategies to reduce both the frequency and the impact of hallucinated outputs.

An effective hybrid system integrates strengths from various approaches:

- **Tool use** enables precise, verifiable outputs by delegating computation-heavy or fact-specific tasks (e.g., date calculation, code execution, currency conversion) to external APIs or structured functions.
- **Retrieval grounding** supplements the model’s internal representations with up-to-date and verifiable information drawn from external sources, reducing reliance on the model’s imperfect parametric memory.
- **Fine-tuning** shapes the model’s inductive biases, helping it learn patterns of truthfulness and factual consistency based on curated datasets or adversarially filtered examples.
- **Guardrails**—such as rule-based filters, logic validators, or knowledge-based correction systems—enforce hard constraints and provide a safety net to catch hallucinations that might bypass other safeguards.

In addition to being hybrid, future mitigation systems should be context-aware, meaning they adapt their strategies dynamically based on the specifics of the application. For example:

- In *high-stakes domains* like medicine, law, or finance, the system may be configured to prioritize factual accuracy over fluency, enforce mandatory citation or retrieval grounding, and invoke fallback procedures when confidence is low.
- In *creative or exploratory domains* such as brainstorming or storytelling, the system might allow more open-ended generation with relaxed factual constraints, while still flagging potentially unverifiable claims.
- In *user-facing applications*, personalization mechanisms could adjust how hallucination warnings, uncertainty indicators, or references are displayed based on user expertise or preferences.



In summary, while hallucination is an inherent risk in current-generation LLMs, its harm can be significantly reduced through a layered approach that combines architectural improvements with systemic controls tailored to the use case. Hybrid, context-sensitive systems represent a practical and responsible path toward building language models that are not only powerful, but also trustworthy, accountable, and safe for deployment in real-world environments.

---

## 9 Monitoring LLM releases and performance: web-based resources

**This section introduces leading web-based resources that researchers can use to monitor LLM releases and evaluate model performance over time, including hallucination trends, intelligence benchmarks, user preferences, and system cost.**

---

As LLMs evolve rapidly, staying informed about new releases, performance metrics, and emerging trends is crucial for researchers, developers, and decision-makers. Rather than relying on static performance comparisons, the following platforms offer continuously updated, transparent, and publicly accessible dashboards that track and evaluate LLM capabilities across a range of tasks, including reasoning, factuality, speed, cost, and hallucination rates.

### 9.1 Artificial Analysis

Artificial Analysis \* is a comprehensive, independent benchmarking platform that compares LLMs and other AI models across several key dimensions. This resource is particularly useful for tracking models that prioritize reasoning and factual grounding—important proxies for hallucination control—even if hallucination itself is not directly measured. Its core features include.

#### 9.1.1 Intelligence index

A composite score based on multiple benchmarks (e.g., MMLU-Pro, GPQA, HLE, LiveCodeBench) reflecting reasoning, problem-solving, and factual capabilities.

---

\*<https://artificialanalysis.ai>

## Artificial Analysis Intelligence Index

Artificial Analysis Intelligence Index incorporates 7 evaluations: MMLU-Pro, GPQA Diamond, Humanity's Last Exam, LiveCodeBench, SciCode, AIME, MATH-500

Estimate (independent evaluation forthcoming)

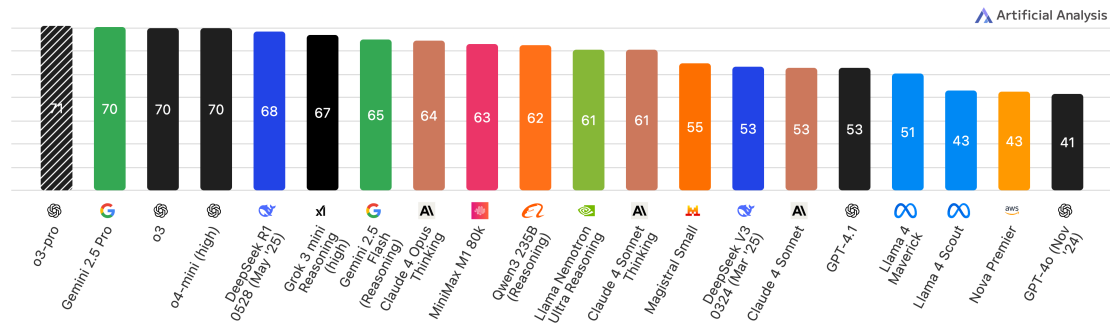


Figure 1: Sample visualization of the AI Index, retrieved on 28 June 2025

### 9.1.2 Cost and latency

Side-by-side comparisons of price per 1K tokens, response latency, and context window limits for models across different providers.

#### Intelligence vs. Price

Artificial Analysis Intelligence Index; Price: USD per 1M Tokens

Most attractive quadrant

Legend: GPT-4.1, o4-mini (high), o3, o3-pro, Llama 4 Maverick, Llama 4 Scout, Gemini 2.5 Flash (Reasoning), Gemini 2.5 Pro, Claude 4 Sonnet Thinking, Claude 4 Sonnet, Claude 4 Opus Thinking, Magistral Small, DeepSeek R1 0528 (May '25), DeepSeek V3 0324 (Mar '25), Grok 3 mini Reasoning (high), Nova Premier, MiniMax M1 80k, Llama Nemotron Ultra Reasoning, Qwen3 235B (Reasoning), GPT-4o (Nov '24)



Figure 2: Sample visualization of intelligence versus price, retrieved on 28 June 2025

### Latency: Time To First Answer Token

Seconds to First Answer Token Received; Accounts for Reasoning Model 'Thinking' time

■ Input processing ■ Thinking (reasoning models, when applicable)

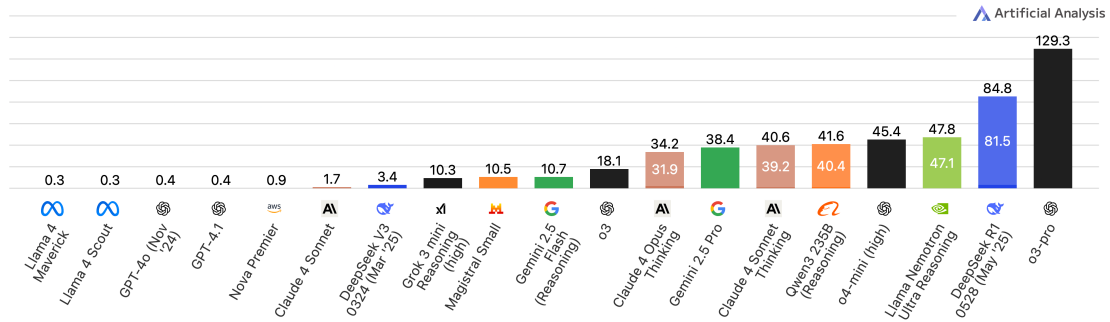


Figure 3: Sample visualization of latency, retrieved on 28 June 2025

### 9.1.3 Model pages

In-depth performance breakdowns for each model version (e.g., GPT-4o, Claude 3.5, LLaMA 3.1), including context sensitivity, update history, and comparative strengths.

### 9.1.4 Multimodal and API benchmarks

Performance data for text-to-image, code, audio, and video generation models.

### Artificial Analysis Image Arena Quality ELO

ELO score in Artificial Analysis Image Arena (relative metric of image generation quality), Higher is better

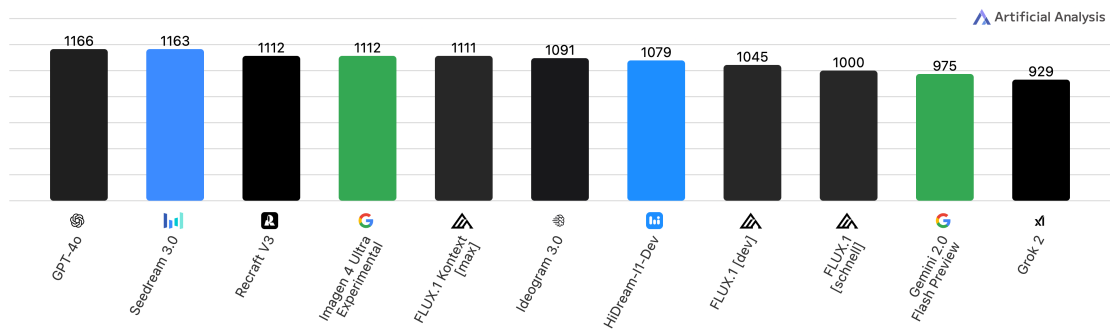


Figure 4: Sample visualization of text to image, retrieved on 28 June 2025

### Artificial Analysis Coding Index

Represents the average of coding benchmarks in the Artificial Analysis Intelligence Index (LiveCodeBench & SciCode)

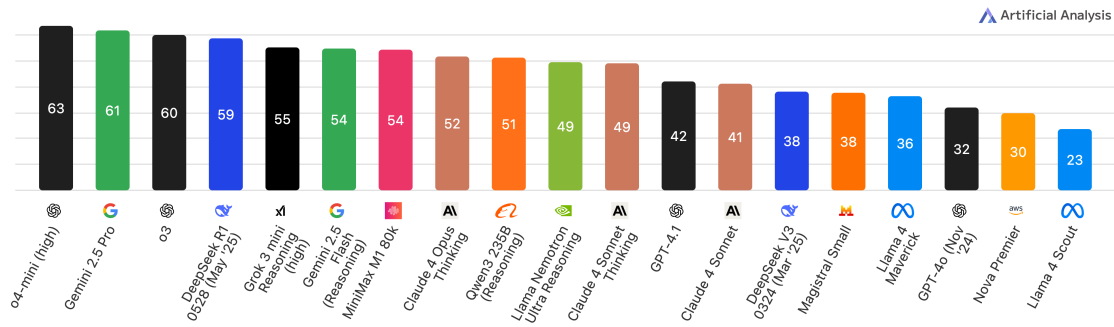


Figure 5: Sample visualization of code generation, retrieved on 28 June 2025

### Word Error Rate

Word error rate: % of words transcribed incorrectly, Lower is better

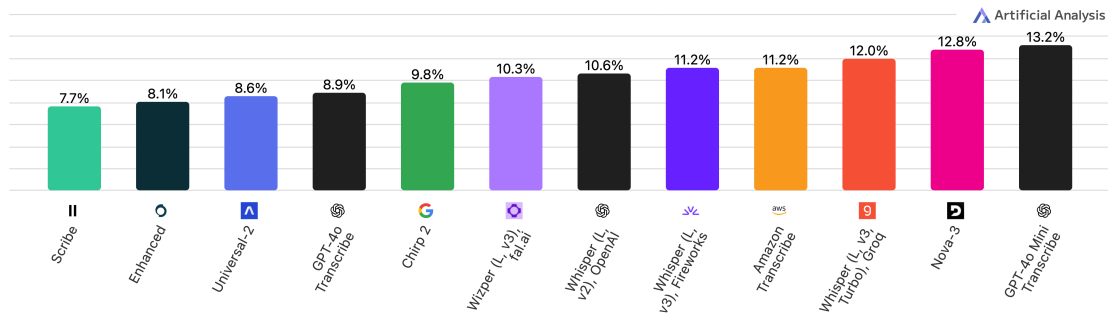


Figure 6: Sample visualization of audio generation (text to speech, word error rate), retrieved on 28 June 2025

## 9.1.5 Quarterly state-of-AI reports

Strategic summaries and macro-level trends such as architectural shifts, emergent capabilities, and safety trade-offs.

### Frontier Language Model Intelligence, Over Time

Artificial Analysis Intelligence Index incorporates 7 evaluations: MMLU-Pro, GPQA Diamond, Humanity's Last Exam, LiveCodeBench, SciCode, AIME, MATH-500

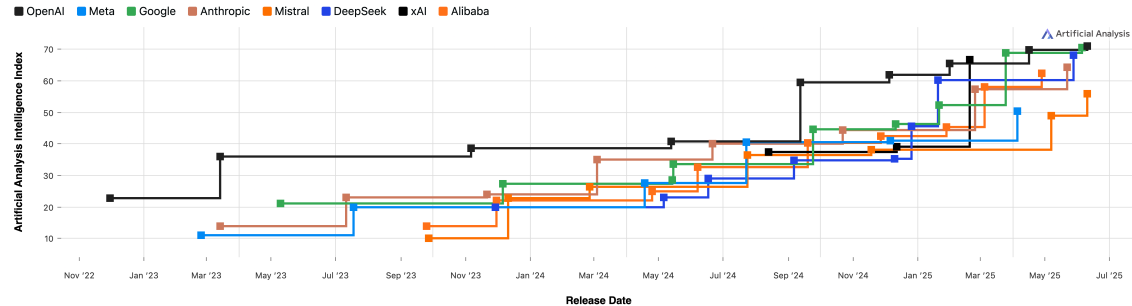


Figure 7: Sample visualization of intelligence over time, retrieved on 28 June 2025

## 9.2 Vectara Hallucination Leaderboard

Maintained by Vectara, this GitHub-based leaderboard<sup>†</sup> is one of the few public efforts that explicitly tracks *hallucination rates* in LLM outputs. Its key features include:

- **Task-specific evaluation:** focuses on hallucination in summarization tasks using human-labeled ground truth references.
- **Quantitative metrics:** reports both hallucination rate (percentage of hallucinated responses) and accuracy rate (non-hallucinated responses), usually based on hundreds of test samples per model.
- **Comparative analysis:** includes OpenAI models (GPT-3.5, GPT-4), Anthropic’s Claude, Meta’s LLaMA, and other open-source models.
- **Consistent benchmarking methodology:** uses the same prompt structure and evaluation rubric across all models for fair comparison.

This leaderboard is an essential tool for researchers focusing on the hallucination problem, particularly in summarization-heavy applications such as legal, academic, or news content generation.

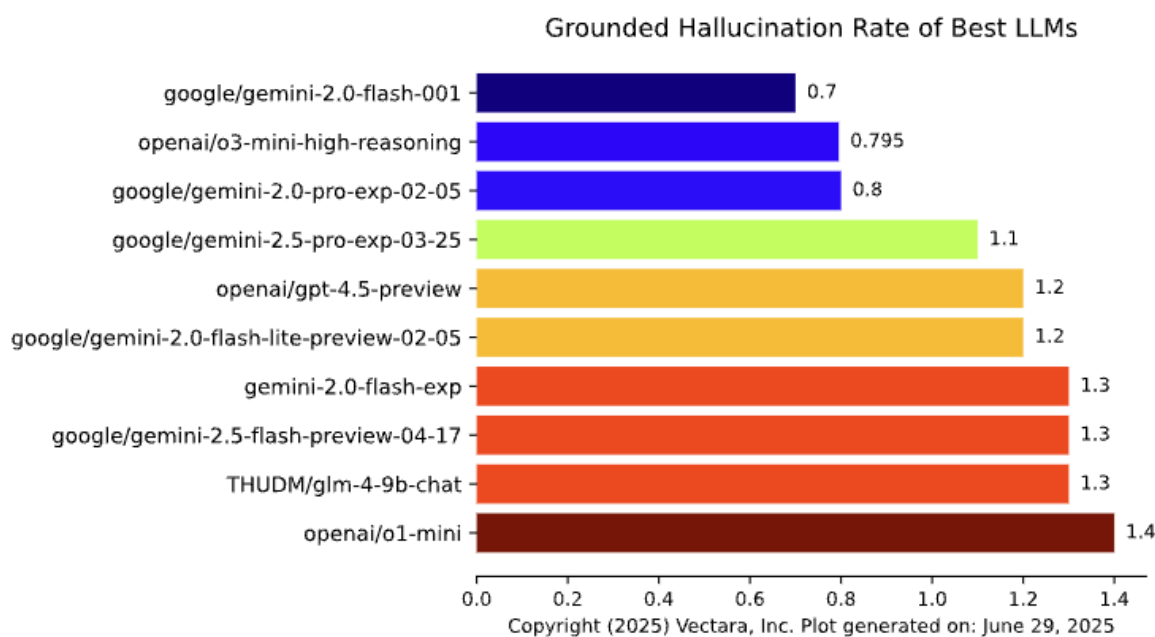


Figure 8: Sample visualization of grounded hallucinations rate using Hughes hallucination evaluation model, retrieved on 29 June 2025

## 9.3 Epoch AI Benchmarking Dashboard

The Epoch AI Benchmarking Dashboard<sup>‡</sup>, a project by the nonprofit research organization Epoch, serves as a valuable resource for understanding the long-term trends in AI capabilities. While it doesn’t directly measure hallucination, the dashboard’s focus on benchmarks such as factual QA and reasoning indirectly reflects a model’s propensity for generating inaccurate

<sup>†</sup><https://github.com/vectara/hallucination-leaderboard>

<sup>‡</sup><https://epoch.ai/data/ai-benchmarking-dashboard>

or fabricated information, making it useful for meta-analyses on the topic. By understanding the factors that contribute to higher accuracy in these benchmarks, we can infer strategies for mitigating hallucinations.

The dashboard offers several key features that contribute to this understanding:

- **Cross-time comparisons:** it meticulously tracks the improvement of leading models over time across various tasks, including MMLU, ARC, and BIG-bench. Improvements in these areas, particularly factual recall and logical consistency, directly correlate with a reduced likelihood of hallucination.
- **Benchmark aggregation:** it combines results from reasoning, language understanding, and coding tasks, enabling a comprehensive analysis of broad capability trends. Enhanced reasoning and language understanding are critical for models to accurately interpret prompts and generate contextually relevant, non-hallucinatory responses.
- **Historical model context:** the dashboard incorporates models dating back to 2018, providing a rich historical perspective on performance scaling and the evolution of AI systems. This allows for observing how changes in model architecture and training have impacted accuracy, offering insights into what contributes to more reliable outputs.
- **Data transparency:** it prioritizes transparency by including information on dataset origins, training scale, and publicly available model sizes. This transparency is crucial for researchers to understand the underlying factors that contribute to model performance and, by extension, to identify potential sources of hallucination.

Beyond these general features, the provided graphs and accompanying text explanations from Epoch AI highlight several insightful trends pertinent to AI performance and, by extension, offer direct implications for combating hallucinations:

### 9.3.1 Accuracy versus training compute

As illustrated in the "Accuracy versus training compute" graph, a clear correlation exists between the estimated training compute (in FLOPs) and the GPQA Diamond and MATH Level 5 accuracies. For GPQA Diamond, models with less than  $10^{24}$  FLOPs often struggle to perform above random chance, sometimes even performing worse due to difficulties in understanding question formatting. This indicates a higher likelihood of generating nonsensical or hallucinated answers. However, beyond this threshold, performance shows a notable increase of approximately 12 percentage points for every 10x increase in compute. Similarly, on MATH Level 5, models with higher compute estimates generally achieve higher scores, with performance increasing around 17 percentage points for every 10x increase in pretraining compute, though this trend appears noisier.

The direct implication for fighting hallucinations is clear: increased training compute leads to more accurate models, which are inherently less prone to hallucinate. Greater computational resources allow models to learn more intricate patterns, better understand factual relationships, and perform more robust reasoning, all of which directly combat the generation of fabricated information. The graph also underscores the impact of algorithmic progress: more recent models like DeepSeek-R1, Phi-4, or Mistral Small 3 surpass older models trained with comparable compute. This suggests that algorithmic advancements, alongside increased compute, are vital for developing models that are more reliably accurate and thus less prone to hallucinate.

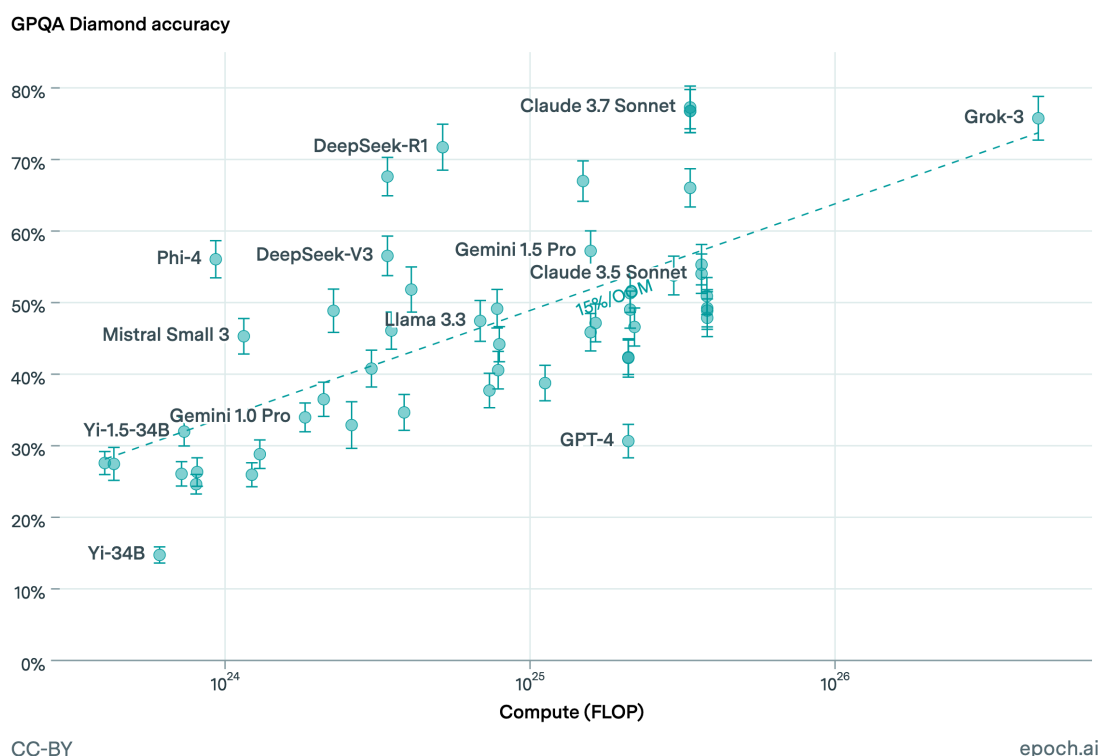


Figure 9: Sample visualization of accuracy versus training compute, retrieved on 29 June 2025

### 9.3.2 Open source versus proprietary

The "Models with downloadable weights vs proprietary" graph demonstrates a discernible performance gap between models with downloadable weights (often open-source) and their top-performing, often proprietary, counterparts. On the GPQA Diamond benchmark, models with downloadable weights tend to lag behind. For instance, in January 2025, OpenAI's o1 outperformed the best downloadable model at the time, Phi-4, by a significant 20 percentage points. A similar disparity was observed on MATH Level 5, where Phi-4 trailed o1 by 29 percentage points. Epoch's analysis further suggests that the best-performing open LLMs lagged the best closed LLMs by a considerable margin, ranging from 6 months on GPQA Diamond to 20 months on MMLU.

This performance gap has significant implications for addressing hallucinations. The limited access to the weights of top-performing models hinders open research into the root causes and mitigation strategies for hallucinations. Researchers cannot directly probe or modify these models to understand why they generate more accurate (and thus less hallucinatory) outputs. However, the release of DeepSeek-R1 in January 2025 marked a notable shift, significantly narrowing this performance gap. On MATH Level 5, DeepSeek-R1 only lagged behind the then-best-performing model, o3-mini, by 2 percentage points. This closing gap is crucial as it suggests that open-source models are catching up in terms of accuracy, which could accelerate community-driven efforts to understand and reduce hallucinations in publicly available models.

## Models with downloadable weights currently lag behind the top-performing models

EPOCH AI

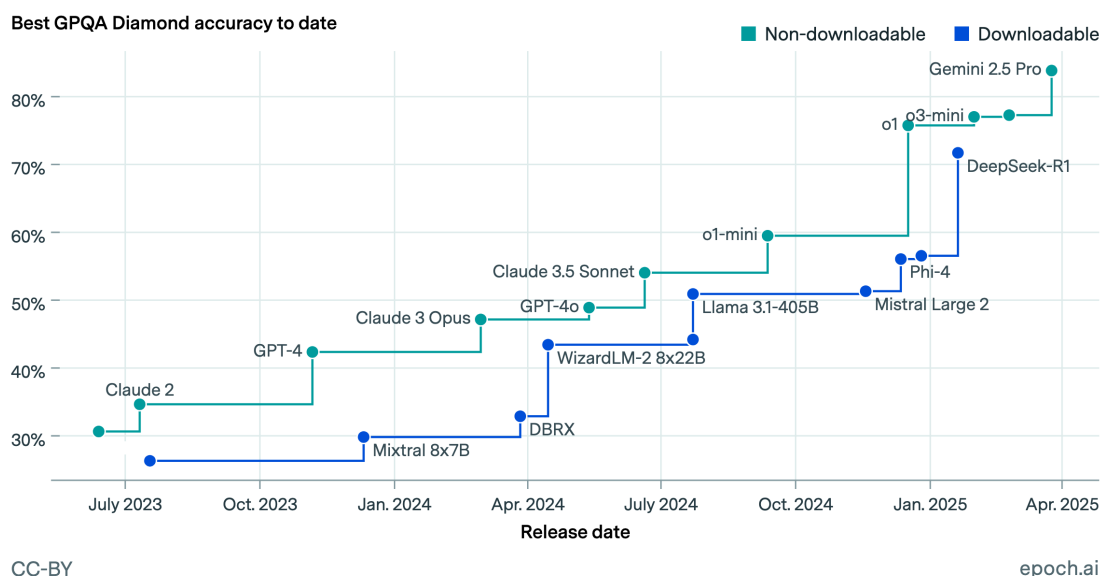


Figure 10: Sample visualization of models with downloadable weights vs proprietary, retrieved on 29 June 2025

### 9.3.3 Geographic disparities and performance

The "US models vs non-US" graph highlights a consistent trend where leading US-developed models exhibit higher accuracies than non-US models on both GPQA Diamond and MATH Level 5 benchmarks. For example, OpenAI's o1 leads on GPQA Diamond, while o3-mini holds the top spot on MATH Level 5, both being US models.

This geographical disparity in performance indirectly impacts the global effort to combat hallucinations. Higher-performing models, regardless of their origin, generally exhibit fewer hallucinations due to their superior understanding and reasoning capabilities. The current dominance of US models suggests that much of the cutting-edge research driving accuracy improvements (and thus hallucination reduction) might be concentrated in US-based organizations. However, with the release of DeepSeek-R1 in January 2025, the performance gap between US and non-US models has substantially reduced. DeepSeek-R1, a non-US model, now trails o3-mini by only 2 percentage points on MATH Level 5 and scores just 4 percentage points lower than o1 on GPQA Diamond. This narrowing gap indicates a more distributed advancement in AI capabilities, which is positive for fostering diverse approaches and collaborations in the fight against hallucinations globally.



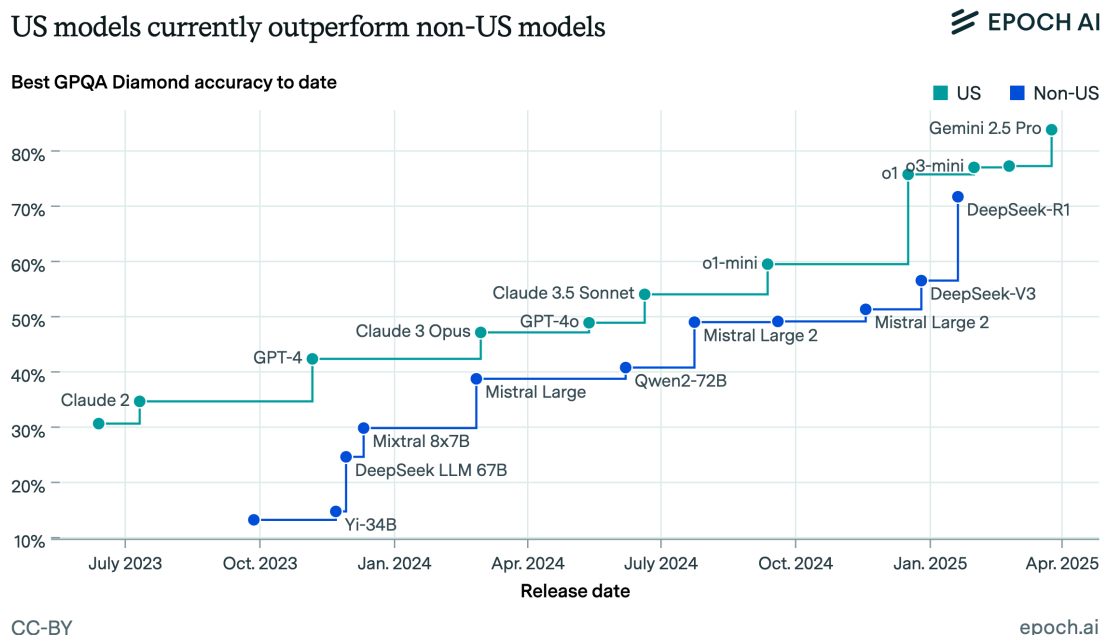


Figure 11: Sample visualization of US models vs non-US, retrieved on 29 June 2025

### 9.3.4 Performance on expert-level mathematics problems

The "models performance on expert-level mathematics problems" graph, focusing on FrontierMath accuracy, provides insights into the capabilities of various AI models in tackling complex mathematical challenges. While the previous graphs mainly used GPQA Diamond, this one provides a different perspective on performance. The data points show the FrontierMath accuracy for different models from various organizations (OpenAI, Anthropic, xAI, Google, Mistral AI, Alibaba, Meta AI, DeepSeek) over time, with error bars indicating the range of results. Notably, models like o4-mini (medium), o3-mini (high), and o1 (high) from OpenAI demonstrate some of the highest accuracies, particularly in the later part of the timeline (late 2024 to early 2025).

This specific benchmark, while not directly tied to hallucination, is crucial for assessing a model's logical reasoning and problem-solving abilities. A high degree of accuracy on expert-level mathematics problems is a strong indicator of a model's foundational understanding and ability to produce precise, non-contradictory outputs. Models that struggle with such tasks are more likely to generate illogical or fabricated results in less constrained domains. Therefore, improvements in FrontierMath accuracy can be seen as a proxy for increased robustness against hallucinations, as it signifies a deeper and more reliable cognitive capacity within the AI model.

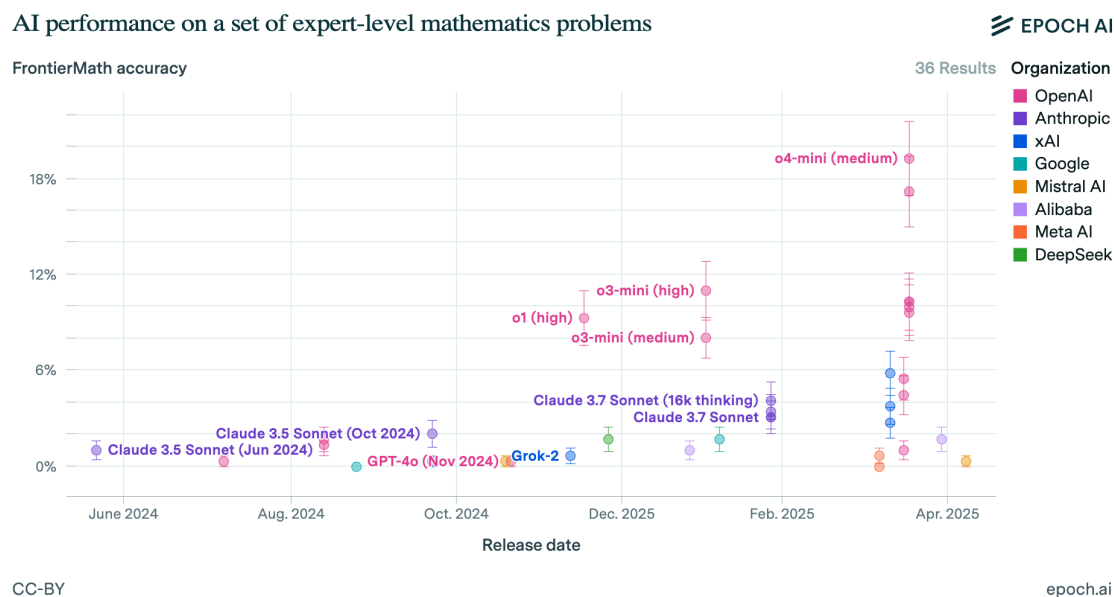


Figure 12: Sample visualization of models performance on expert-level mathematics problems, retrieved on 29 June 2025

## 9.4 LM Arena

Created by researchers from UC Berkeley’s SkyLab, LM Arena (formerly LMSYS Chatbot Arena)<sup>§</sup> is an open platform where anyone can easily access, explore, and interact with the world’s leading AI models. The platform’s foundational values, as expressed by its founders when the first leaderboard was launched in July 2023, are deeply rooted in research: to create a rigorous, reproducible, community-led framework for real-world model evaluation. This commitment to open, community-driven assessment makes LM Arena a uniquely valuable resource for understanding AI model behavior, particularly concerning the critical issue of hallucination. While not explicitly designed as a hallucination detector, its methodology provides crucial qualitative insights into how users perceive model accuracy, helpfulness, and trustworthiness, all of which are intrinsically linked to the presence or absence of factual inaccuracies or fabricated information.

The core of LM Arena’s operation, designed like a tournament, involves models being compared anonymously side-by-side, with users voting for the better response. This structure of anonymous battles, dynamic prompts (approximately 70% of prompts each month are fresh), and a rotating user base, was specifically designed to reduce bias and reflect diverse, real-world use cases. This dynamic testing environment makes it impossible for models to predict or “memorize” specific evaluation scenarios, ensuring that their performance, including their tendency to hallucinate, is genuinely assessed based on real-time, novel interactions. Its key features are particularly relevant to the study of hallucinations:

<sup>§</sup><https://lmarena.ai/leaderboard>

### 9.4.1 Battle-style comparisons and dynamic prompts

Models are compared head-to-head in blind A/B testing formats by real users, who vote on preferred responses. The anonymity of models and the constant introduction of fresh prompts mean that models cannot optimize for specific test cases. This structure means users are implicitly evaluating which model provides more accurate and reliable information, directly penalizing models that hallucinate. A user is far more likely to prefer a response that is factual, coherent, and contextually appropriate over one that contains made-up details or nonsensical statements. This "real-world usage" approach contrasts sharply with static benchmarks, providing a robust signal for hallucination.

### 9.4.2 Diversity of models and transparent testing

The platform includes a wide array of both proprietary (e.g., OpenAI, Anthropic, Google, Meta, Alibaba) and open-source (e.g., Mistral, LLaMA, Zephyr) models, with over 40% of battles involving an open model. LM Arena works directly with model providers for testing, comparing, and improving models both before (via pseudonyms/codenames) and after official release. This provides a shared infrastructure for reproducible and transparent evaluation. This diverse representation and testing methodology allows for comparisons across different development philosophies and architectures, helping to identify whether certain approaches or model types are inherently more prone to hallucination in various real-world interaction contexts. Only publicly released models with longer-term support get ranked on the leaderboard, ensuring that the community can verify results through their own testing.


### 9.4.3 High-quality qualitative judgments driven by intrinsic motivation

Crucially, voters implicitly assess fluency, factuality, helpfulness, and hallucination tendencies—providing a rich complementary signal to traditional benchmarks. Since there’s no payment or external incentive, votes come from intrinsic motivation, fostering a community of diverse subject-matter experts who provide authentic, thoughtful evaluations on their own prompts. When a user deems a response unhelpful, untrustworthy, or simply incorrect, it often stems from a factual inaccuracy or a fabricated piece of information—a direct manifestation of hallucination. These high-quality qualitative judgments, gathered at scale, offer a powerful "human-in-the-loop" feedback mechanism for identifying models that consistently produce non-hallucinatory content. They capture the nuanced user experience and reactions to model outputs that purely quantitative metrics might miss.

### 9.4.4 Community-shaped leaderboard and transparency

Over time, these votes in battle mode add up to a public leaderboard that reflects collective, real-world judgment. This democratically shaped leaderboard makes AI progress more transparent, accessible, and grounded in actual usage. Models consistently ranking higher are, by extension, those perceived as more reliable and less prone to frustrating users with incorrect or fabricated information. This dynamic ranking can serve as an early warning system for models that start exhibiting higher hallucination rates, as user dissatisfaction would quickly reflect in their standings. Additionally, while prompts and votes from all modes (battle, side-by-side, direct chat) are collected for transparent research, only anonymous battle votes contribute to the leaderboards, ensuring fairness. LM Arena also supports open research beyond the leaderboard, actively developing new artifacts and statistical methods to understand human

preference with clarity and precision, including decomposing preference into components like tone, helpfulness, formatting, and emotional resonance. This research directly contributes to understanding how AI is perceived and trusted, which are key variables in understanding and mitigating hallucinations.

 **Text** 🕒 2 days ago











Rank (UB) ↑	Model ↓	Score ↓	Votes ↓
1	 gemini-2.5-pro	1465	15,769
2	 o3-2025-04-16	1450	21,965
2	 chatgpt-4o-latest-20250326	1443	24,237
3	 gpt-4.5-preview-2025-02-27	1436	15,271
5	 claude-opus-4-20250514	1417	20,056
5	 gemini-2.5-flash	1416	21,209
5	 deepseek-r1-0528	1414	12,847
5	 gpt-4.1-2025-04-14	1411	18,275
5	 grok-3-preview-02-24	1409	25,763
10	 o1-2024-12-17	1399	29,038

Figure 13: Sample visualization of models performance on text generation, retrieved on 9 July 2025

**Q Search** 🕒 47 days ago











Rank (UB) ↑	Model ↓	Score ↓	Votes ↓
1	 gemini-2.5-pro-grounding	1142	1,215
1	 ppl-sonar-reasoning-pro-high	1136	861
3	 ppl-sonar-reasoning	1097	1,644
3	 ppl-sonar	1072	1,208
3	 ppl-sonar-pro-high	1071	1,364
4	 ppl-sonar-pro	1066	1,214
7	 gemini-2.0-flash-grounding	1028	1,193
7	 api-gpt-4o-search	1000	1,196
7	 api-gpt-4o-search-high	999	1,707
8	 api-gpt-4o-search-high-loc	994	1,226

Figure 14: Sample visualization of models performance on web search for real-time information, external knowledge, and grounded citations, retrieved on 9 July 2025

Vision			
Rank (UB) ↑	Model ↓	Score ↓	Votes ↓
1	gemini-2.5-pro	1271	3,508
2	chatgpt-4o-latest-20250326	1244	5,702
2	o3-2025-04-16	1238	4,268
2	gpt-4.5-preview-2025-02-27	1230	3,066
3	gemini-2.5-flash	1221	4,367
3	gpt-4.1-2025-04-14	1216	4,575
6	o4-mini-2025-04-16	1202	3,705
6	gpt-4.1-mini-2025-04-14	1202	4,051
6	o1-2024-12-17	1198	3,825
7	claude-3-7-sonnet-20250219-t...	1188	2,055

Figure 15: Sample visualization of models performance on generative AI models capable of understanding and processing visual inputs, retrieved on 9 July 2025

Copilot			
Rank (UB) ↑	Model ↓	Score ↓	Votes ↓
1	Deepseek V2.5 (FIM)	1028	2,292
1	Claude 3.5 Sonnet (06/20)	1012	3,544
1	Claude 3.5 Sonnet (10/22)	1004	3,596
1	Codestral (25.01)	1001	2,180
1	Qwen-2.5-Coder (FiM)	998	3,401
1	Mercury Coder Mini	994	1,430
2	Codestral (05/24)	1001	5,744
3	GPT-4o (08/06)	986	4,464
3	Gemini-1.5-Pro-002	986	3,441
3	Meta-Llama-3.1-405B-Instruct	984	3,432

Figure 16: Sample visualization of models performance on how well AI coding assistants understand and generate code across various programming languages and tasks, retrieved on 9 July 2025

Although not focused exclusively on hallucination, LM Arena provides real-world insights into which models are perceived as most helpful and trustworthy by end users. This user perception is a strong indicator of a model’s ability to avoid hallucinations, as models that frequently hallucinate are unlikely to be rated as helpful or trustworthy. Therefore, LM Arena complements more technical hallucination metrics by offering a crucial perspective on the practical impact of hallucination on user experience and satisfaction. It underscores that reducing hallucinations is not just about factual correctness, but also about building user trust and providing genuinely useful AI assistance in real-world scenarios.

## 10 Conclusions

### 10.1 The complex nature and inevitable presence of LLM hallucinations

Hallucination remains a pervasive and multifaceted challenge for LLMs, characterized by the generation of content that is plausible but factually incorrect, inconsistent, or entirely fabricated<sup>[38;27;42;47]</sup>. As thoroughly evidenced throughout this report, these errors manifest in diverse forms, from core distinctions like intrinsic versus extrinsic hallucinations (contradicting input context versus inconsistency with training data or reality)<sup>[7;70;27;79]</sup> and factuality versus faithfulness hallucinations (absolute correctness versus adherence to input)<sup>[13;50;64;96;61]</sup>, to

specific manifestations such as factual errors, contextual and logical inconsistencies, temporal disorientation, ethical violations, and task-specific errors in domains like code generation and multimodal contexts<sup>[42;14;6;47;19;92;57;2;98;51;27;100;38]</sup>.

These diverse issues stem from a complex interplay of factors, including data quality and biases<sup>[42;32;22]</sup>, the inherent auto-regressive nature of LLM architectures<sup>[50;38]</sup>, limitations in their training processes (such as exposure bias and capability misalignment)<sup>[78;59]</sup>, and stochastic decoding strategies<sup>[38]</sup>. Critically, formal theoretical proofs presented in this report indicate that hallucination is an innate and inevitable limitation for computable LLMs<sup>[100]</sup>, suggesting that complete elimination may be impossible regardless of architectural advancements or training refinements.

## 10.2 Implications for detection, mitigation, and human interaction

This profound implication necessitates a strategic shift from attempting complete eradication to developing robust detection mechanisms, implementing effective mitigation strategies, and ensuring continuous human oversight. The nuanced and task-specific nature of hallucinations underscores the need for granular approaches to their understanding and management, as a "one-size-fits-all" solution is unlikely to be effective.

The report has highlighted the critical role of cognitive and human factors in hallucination perception, emphasizing that user interfaces and interaction designs must incorporate strategies like calibrated uncertainty displays and source-grounding indicators to improve user resilience and trust<sup>[102;103;54]</sup>. Furthermore, a comprehensive survey of evaluation benchmarks and metrics has revealed the ongoing challenges in standardized assessment, underscoring the need for unified, taxonomy-aware frameworks that can provide granular, diagnostic insights into hallucination types.

## 10.3 Future directions for responsible LLM deployment

Finally, the discussion of architectural and systemic mitigation strategies, including Toolformer-style augmentation<sup>[86]</sup> and RAG,<sup>[42]</sup> alongside the introduction to web-based resources for monitoring LLM performance, provides practical directions for future development and deployment. This holistic understanding of hallucination types, their underlying causes, human interaction factors, evaluation methodologies, and mitigation techniques is paramount for developing more reliable, trustworthy, and safely deployable LLMs, particularly in high-stakes domains such as medicine and law, where the consequences of false information can be severe<sup>[51;47;40;31]</sup>. Continued research and responsible deployment practices, with a focus on human-in-the-loop validation and external safeguards, are essential for navigating the inherent limitations of LLMs and maximizing their transformative potential.

## References

- [1] Samar AboulEla, Paria Zabihitari, Nourhan Ibrahim, Majid Afshar, and Rasha Kashef. Exploring rag solutions to reduce hallucinations in llms. In *2025 IEEE International systems Conference (SysCon)*, pages 1–8. IEEE, 2025.

- [2] Vibhor Agarwal, Yulong Pei, Salwa Alamer, and Xiaomo Liu. Codemirage: Hallucinations in code generated by large language models. *arXiv preprint arXiv:2408.08333*, 2024.
- [3] Jack Albright and Sheden Andemicael. Improving llm mathematical reasoning capabilities using external tools. *Stanford CS224R*, 2025.
- [4] Shengnan An, Zexiong Ma, Zeqi Lin, Nanning Zheng, Jian-Guang Lou, and Weizhu Chen. Make your llm fully utilize the context. *Advances in Neural Information Processing Systems*, 37:62160–62188, 2024.
- [5] Maksym Andriushchenko, Alexandra Souly, Mateusz Dziemian, Derek Duenas, Maxwell Lin, Justin Wang, Dan Hendrycks, Andy Zou, Zico Kolter, Matt Fredrikson, et al. Agentharm: A benchmark for measuring harmfulness of llm agents. *arXiv preprint arXiv:2410.09024*, 2024.
- [6] Amos Azaria and Tom Mitchell. The internal state of an llm knows when it’s lying. *arXiv preprint arXiv:2304.13734*, 2023.
- [7] Yejin Bang, Ziwei Ji, Alan Schelten, Anthony Hartshorn, Tara Fowler, Cheng Zhang, Nicola Cancedda, and Pascale Fung. Hallulens: Llm hallucination benchmark. *arXiv preprint arXiv:2504.17550*, 2025.
- [8] Anne Beyer, Sharid Loáiciga, and David Schlangen. Is incoherence surprising? targeted evaluation of coherence prediction from language models. *arXiv preprint arXiv:2105.03495*, 2021.
- [9] Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. ‘it’s reducing a human being to a percentage’ perceptions of justice in algorithmic decisions. In *Proceedings of the 2018 Chi conference on human factors in computing systems*, pages 1–14, 2018.
- [10] Suhas BN, Han-Chin Shing, Lei Xu, Mitch Strong, Jon Burnsky, Jessica Ofor, Jordan R Mason, Susan Chen, Sundararajan Srinivasan, Chaitanya Shivade, et al. Fact-controlled diagnosis of hallucinations in medical text summarization. *arXiv preprint arXiv:2506.00448*, 2025.
- [11] Sébastien Bubeck, Varun Chadrasekaran, Ronen Eldan, Johannes Gehrike, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4, 2023.
- [12] Courtnei Byun, Piper Vasicek, and Kevin Seppi. This reference does not exist: an exploration of llm citation accuracy and relevance. In *Proceedings of the Third Workshop on Bridging Human–Computer Interaction and Natural Language Processing*, pages 28–39, 2024.
- [13] Meng Cao, Yue Dong, and Jackie Chi Kit Cheung. Hallucinated but factual! inspecting the factuality of hallucinations in abstractive summarization. *arXiv preprint arXiv:2109.09784*, 2021.
- [14] Alex Chandler, Devesh Surve, and Hui Su. Detecting errors through ensembling prompts (deep): an end-to-end llm framework for detecting factual errors. *arXiv preprint arXiv:2406.13009*, 2024.

- [15] Haw-Shiuan Chang and Andrew McCallum. Softmax bottleneck makes language models unable to represent multi-mode word distributions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, volume 1, 2022.
- [16] Xiuying Chen, Mingzhe Li, Xin Gao, and Xiangliang Zhang. Towards improving faithfulness in abstractive summarization. *Advances in Neural Information Processing Systems*, 35:24516–24528, 2022.
- [17] ChenghaoZhu ChenghaoZhu, Nuo Chen, Yufei Gao, Yunyi Zhang, Prayag Tiwari, and Benyou Wang. Is your llm outdated? a deep look at temporal generalization. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7433–7457, 2025.
- [18] Inyoung Cheong, Aylin Caliskan, and Tadayoshi Kohno. Envisioning legal mitigations for llm-based intentional and unintentional harms. *Adm. Law J*, 2022.
- [19] Inyoung Cheong, King Xia, KJ Kevin Feng, Quan Ze Chen, and Amy X Zhang. (a) i am not a lawyer, but...: engaging legal experts towards responsible llm policies for legal advice. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 2454–2469, 2024.
- [20] Prateek Chhikara. Mind the confidence gap: Overconfidence, calibration, and distractor effects in large language models. *arXiv preprint arXiv:2502.11028*, 2025.
- [21] Mary L Cummings. Automation bias in intelligent time critical decision support systems. In *Decision making in aviation*, pages 289–294. Routledge, 2017.
- [22] Sunhao Dai, Chen Xu, Shicheng Xu, Liang Pang, Zhenhua Dong, and Jun Xu. Bias and unfairness in information retrieval systems: New challenges in the llm era. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 6437–6447, 2024.
- [23] Yi Dong, Ronghui Mu, Gaojie Jin, Yi Qi, Jinwei Hu, Xingyu Zhao, Jie Meng, Wenjie Ruan, and Xiaowei Huang. Building guardrails for large language models. *arXiv preprint arXiv:2402.01822*, 2024.
- [24] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.
- [25] Mary T Dzindolet, Scott A Peterson, Regina A Pomranky, Linda G Pierce, and Hall P Beck. The role of trust in automation reliance. *International journal of human-computer studies*, 58(6):697–718, 2003.
- [26] Nouha Dziri, Sivan Milton, Mo Yu, Osmar Zaiane, and Siva Reddy. On the origin of hallucinations in conversational models: Is it the datasets or the models? *arXiv preprint arXiv:2204.07931*, 2022.
- [27] Passant Elchafei and Mervet Abu-Elkheir. Span-level hallucination detection for llm-generated answers. *arXiv preprint arXiv:2504.18639*, 2025.



- [28] Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. Ragas: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 150–158, 2024.
- [29] Lucile Favero, Juan Antonio Pérez-Ortiz, Tanja Käser, and Nuria Oliver. Enhancing critical thinking in education by means of a socratic chatbot. *arXiv preprint arXiv:2409.05511*, 2024.
- [30] Shangbin Feng, Weijia Shi, Yike Wang, Wenxuan Ding, Vidhisha Balachandran, and Yulia Tsvetkov. Don’t hallucinate, abstain: Identifying llm knowledge gaps via multi-llm collaboration. *arXiv preprint arXiv:2402.00367*, 2024.
- [31] Yuyou Gan, Yong Yang, Zhe Ma, Ping He, Rui Zeng, Yiming Wang, Qingming Li, Chunyi Zhou, Songze Li, Ting Wang, et al. Navigating the risks: A survey of security, privacy, and ethics threats in llm-based agents. *arXiv preprint arXiv:2411.09523*, 2024.
- [32] Ankush Ramprakash Gautam. Impact of high data quality on llm hallucinations. *International Journal of Computer Applications*, 975:8887, 2025.
- [33] Zorik Gekhman, Jonathan Herzig, Roei Aharoni, Chen Elkind, and Idan Szpektor. Trueteacher: Learning factual consistency evaluation with large language models. *arXiv preprint arXiv:2305.11171*, 2023.
- [34] Bishwamittra Ghosh, Sarah Hasan, Naheed Anjum Arafat, and Arijit Khan. Logical consistency of large language models in fact-checking. *arXiv preprint arXiv:2412.16100*, 2024.
- [35] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudik, and Hanna Wallach. Improving fairness in machine learning systems: What do industry practitioners need? In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pages 1–16, 2019.
- [36] Or Honovich, Roei Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. True: Re-evaluating factual consistency evaluation. *arXiv preprint arXiv:2204.04991*, 2022.
- [37] Or Honovich, Leshem Choshen, Roei Aharoni, Ella Neeman, Idan Szpektor, and Omri Abend. q<sup>2</sup>: Evaluating factual consistency in knowledge-grounded dialogues via question generation and question answering. *arXiv preprint arXiv:2104.08202*, 2021.
- [38] 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.
- [39] Gautier Izacard and Edouard Grave. Leveraging passage retrieval with generative models for open domain question answering. *arXiv preprint arXiv:2007.01282*, 2020.
- [40] Junfeng Jiao, Saleh Afroogh, Yiming Xu, and Connor Phillips. Navigating llm ethics: Advancements, challenges, and future directions. *arXiv preprint arXiv:2406.18841*, 2024.

- [41] Zhijing Jin, Jiarui Liu, Zhiheng Lyu, Spencer Poff, Mrinmaya Sachan, Rada Mihalcea, Mona Diab, and Bernhard Schölkopf. Can large language models infer causation from correlation? *arXiv preprint arXiv:2306.05836*, 2023.
- [42] Satyadhar Joshi. Mitigating llm hallucinations: A comprehensive review of techniques and architectures. *preprints.org*, 2025.
- [43] Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. Understanding the effects of rlhf on llm generalisation and diversity. *arXiv preprint arXiv:2310.06452*, 2023.
- [44] Wojciech Kryściński, Bryan McCann, Caiming Xiong, and Richard Socher. Evaluating the factual consistency of abstractive text summarization. *arXiv preprint arXiv:1910.12840*, 2019.
- [45] Philippe Laban, Tobias Schnabel, Paul N Bennett, and Marti A Hearst. Summac: Revisiting nli-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177, 2022.
- [46] Ronan Le Bras, Swabha Swayamdipta, Chandra Bhagavatula, Rowan Zellers, Matthew Peters, Ashish Sabharwal, and Yejin Choi. Adversarial filters of dataset biases. In *International conference on machine learning*, pages 1078–1088. Pmlr, 2020.
- [47] Yunseo Lee, John Youngeun Song, Dongsun Kim, Jindae Kim, Mijung Kim, and Jaechang Nam. Hallucination by code generation llms: Taxonomy, benchmarks, mitigation, and challenges. *arXiv preprint arXiv:2504.20799*, 2025.
- [48] 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.
- [49] Hao Li, Xiaogeng Liu, Hung-Chun Chiu, Dianqi Li, Ning Zhang, and Chaowei Xiao. Drift: Dynamic rule-based defense with injection isolation for securing llm agents. *arXiv preprint arXiv:2506.12104*, 2025.
- [50] Junyi Li, Jie Chen, Ruiyang Ren, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. The dawn after the dark: An empirical study on factuality hallucination in large language models. *arXiv preprint arXiv:2401.03205*, 2024.
- [51] Ningke Li, Yahui Song, Kailong Wang, Yuekang Li, Ling Shi, Yi Liu, and Haoyu Wang. Detecting llm fact-conflicting hallucinations enhanced by temporal-logic-based reasoning. *arXiv preprint arXiv:2502.13416*, 2025.
- [52] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.
- [53] Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*, 2021.
- [54] Stephanie Lin, Jacob Hilton, and Owain Evans. Teaching models to express their uncertainty in words. *arXiv preprint arXiv:2205.14334*, 2022.

- [55] Yong Lin, Hangyu Lin, Wei Xiong, Shizhe Diao, Jianmeng Liu, Jipeng Zhang, Rui Pan, Haoxiang Wang, Wenbin Hu, Hanning Zhang, et al. Mitigating the alignment tax of rlhf. *arXiv preprint arXiv:2309.06256*, 2023.
- [56] Lin Ling, Fazle Rabbi, Song Wang, and Jinqiu Yang. Bias unveiled: Investigating social bias in llm-generated code. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 27491–27499, 2025.
- [57] Fang Liu, Yang Liu, Lin Shi, Houkun Huang, Ruifeng Wang, Zhen Yang, Li Zhang, Zhongqi Li, and Yuchi Ma. Exploring and evaluating hallucinations in llm-powered code generation. *arXiv preprint arXiv:2404.00971*, 2024.
- [58] Yang Liu, Yuanshun Yao, Jean-Francois Ton, Xiaoying Zhang, Ruocheng Guo, Hao Cheng, Yegor Klochkov, Muhammad Faaiz Taufiq, and Hang Li. Trustworthy llms: a survey and guideline for evaluating large language models’ alignment. *arXiv preprint arXiv:2308.05374*, 2023.
- [59] Tingmingke Lu. Maximum hallucination standards for domain-specific large language models. *arXiv preprint arXiv:2503.05481*, 2025.
- [60] Ewa Luger and Abigail Sellen. ” like having a really bad pa” the gulf between user expectation and experience of conversational agents. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pages 5286–5297, 2016.
- [61] Ben Malin, Tatiana Kalganova, and Nikolaos Boulgouris. A review of faithfulness metrics for hallucination assessment in large language models. *IEEE Journal of Selected Topics in Signal Processing*, 2025.
- [62] Nikolay Malkin, Zhen Wang, and Nebojsa Jojic. Coherence boosting: When your pre-trained language model is not paying enough attention. *arXiv preprint arXiv:2110.08294*, 2021.
- [63] Jiayuan Mao, Chuang Gan, Pushmeet Kohli, Joshua B Tenenbaum, and Jiajun Wu. The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision. *arXiv preprint arXiv:1904.12584*, 2019.
- [64] Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On faithfulness and factuality in abstractive summarization. *arXiv preprint arXiv:2005.00661*, 2020.
- [65] Seysha Mehta and Neil Mehta. Embracing the illusion of explanatory depth: a strategic framework for using iterative prompting for integrating large language models in health-care education. *Medical Teacher*, 47(2):208–211, 2025.
- [66] Sewon Min, Kalpesh Krishna, Xinxu 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] Raymond S Nickerson. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology*, 2(2):175–220, 1998.

- [68] Maurice M Ohayon. Prevalence of hallucinations and their pathological associations in the general population. *Psychiatry research*, 97(2-3):153–164, 2000.
- [69] Mahmud Omar, Reem Agbareia, Benjamin S Glicksberg, Girish N Nadkarni, and Eyal Klang. Benchmarking the confidence of large language models in answering clinical questions: cross-sectional evaluation study. *JMIR Medical Informatics*, 13:e66917, 2025.
- [70] Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szpektor, Hadas Kotek, and Yonatan Belinkov. Llm’s know more than they show: On the intrinsic representation of llm hallucinations. *arXiv preprint arXiv:2410.02707*, 2024.
- [71] Daniel E O’Leary. Confirmation and specificity biases in large language models: An explorative study. *IEEE Intelligent Systems*, 40(1):63–68, 2025.
- [72] Bogdan Padiu, Radu Iacob, Traian Rebedea, and Mihai Dascalu. To what extent have llms reshaped the legal domain so far? a scoping literature review. *Information*, 15(11):662, 2024.
- [73] Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Med-halt: Medical domain hallucination test for large language models. *arXiv preprint arXiv:2307.15343*, 2023.
- [74] Shrey Pandit, Jiawei Xu, Junyuan Hong, Zhangyang Wang, Tianlong Chen, Kaidi Xu, and Ying Ding. Medhallu: A comprehensive benchmark for detecting medical hallucinations in large language models. *arXiv preprint arXiv:2502.14302*, 2025.
- [75] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.
- [76] Mihir Parmar, Hanieh Deilamsalehy, Franck Dernoncourt, Seunghyun Yoon, Ryan A Rossi, and Trung Bui. Towards enhancing coherence in extractive summarization: Dataset and experiments with llms. *arXiv preprint arXiv:2407.04855*, 2024.
- [77] Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, et al. Kilt: a benchmark for knowledge intensive language tasks. *arXiv preprint arXiv:2009.02252*, 2020.
- [78] Andrea Pozzi, Alessandro Incremona, Daniele Tessler, and Daniele Toti. Mitigating exposure bias in large language model distillation: an imitation learning approach. *Neural Computing and Applications*, pages 1–17, 2025.
- [79] Shaik Rafi, Lenin Laitonjam, and Ranjita Das. Reducing extrinsic hallucination in multimodal abstractive summaries with post-processing technique. *Neural Computing and Applications*, pages 1–21, 2025.
- [80] Rolf Reber and Norbert Schwarz. Effects of perceptual fluency on judgments of truth. *Consciousness and cognition*, 8(3):338–342, 1999.
- [81] Ruiyang Ren, Yuhao Wang, Yingqi Qu, Wayne Xin Zhao, Jing Liu, Hao Tian, Hua Wu, Ji-Rong Wen, and Haifeng Wang. Investigating the factual knowledge boundary of large language models with retrieval augmentation. *arXiv preprint arXiv:2307.11019*, 2023.

- [82] Dimitri Roustan, François Bastardot, et al. The clinicians’ guide to large language models: A general perspective with a focus on hallucinations. *Interactive journal of medical research*, 14(1):e59823, 2025.
- [83] Leonid Rozenblit and Frank Keil. The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive science*, 26(5):521–562, 2002.
- [84] Jon Saad-Falcon, Omar Khattab, Christopher Potts, and Matei Zaharia. Ares: An automated evaluation framework for retrieval-augmented generation systems. *arXiv preprint arXiv:2311.09476*, 2023.
- [85] Alireza Salemi and Hamed Zamani. Evaluating retrieval quality in retrieval-augmented generation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2395–2400, 2024.
- [86] Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems*, 36:68539–68551, 2023.
- [87] Thomas Scialom, Paul-Alexis Dray, Patrick Gallinari, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, and Alex Wang. Questeval: Summarization asks for fact-based evaluation. *arXiv preprint arXiv:2103.12693*, 2021.
- [88] Ashish Seth, Dinesh Manocha, and Chirag Agarwal. Hallucinogen: A benchmark for evaluating object hallucination in large visual-language models. *arXiv preprint arXiv:2412.20622*, 2024.
- [89] Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36:8634–8652, 2023.
- [90] Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. Retrieval augmentation reduces hallucination in conversation. *arXiv preprint arXiv:2104.07567*, 2021.
- [91] Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.
- [92] Ben Snyder, Marius Moisesescu, and Muhammad Bilal Zafar. On early detection of hallucinations in factual question answering. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2721–2732, 2024.
- [93] Amir Taubenfeld, Yaniv Dover, Roi Reichart, and Ariel Goldstein. Systematic biases in llm simulations of debates. *arXiv preprint arXiv:2402.04049*, 2024.
- [94] Yuchen Tian, Weixiang Yan, Qian Yang, Xuandong Zhao, Qian Chen, Wen Wang, Ziyang Luo, Lei Ma, and Dawn Song. Codehalu: Investigating code hallucinations in llms via execution-based verification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 25300–25308, 2025.

- [95] Rosario Uceda-Sosa, Karthikeyan Natesan Ramamurthy, Maria Chang, and Moninder Singh. Reasoning about concepts with llms: Inconsistencies abound. *arXiv preprint arXiv:2405.20163*, 2024.
- [96] Prathiksha Rumale Vishwanath, Simran Tiwari, Tejas Ganesh Naik, Sahil Gupta, Dung Ngoc Thai, Wenlong Zhao, SUNJAE KWON, Victor Ardulov, Karim Tarabishy, Andrew McCallum, et al. Faithfulness hallucination detection in healthcare ai. In *Artificial Intelligence and Data Science for Healthcare: Bridging Data-Centric AI and People-Centric Healthcare*, 2024.
- [97] Yuxuan Wang, Yueqian Wang, Dongyan Zhao, Cihang Xie, and Zilong Zheng. Videohalluciner: Evaluating intrinsic and extrinsic hallucinations in large video-language models. *arXiv preprint arXiv:2406.16338*, 2024.
- [98] Shengqiong Wu, Hao Fei, Liangming Pan, William Yang Wang, Shuicheng Yan, and Tat-Seng Chua. Combating multimodal llm hallucination via bottom-up holistic reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 8460–8468, 2025.
- [99] Xilie Xu, Keyi Kong, Ning Liu, Lizhen Cui, Di Wang, Jingfeng Zhang, and Mohan Kankanhalli. An llm can fool itself: A prompt-based adversarial attack. *arXiv preprint arXiv:2310.13345*, 2023.
- [100] Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. Hallucination is inevitable: An innate limitation of large language models. *arXiv preprint arXiv:2401.11817*, 2024.
- [101] Wenpeng Yin, Qinyuan Ye, Pengfei Liu, Xiang Ren, and Hinrich Schütze. Llm-driven instruction following: Progresses and concerns. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts*, pages 19–25, 2023.
- [102] Minji Yoo. How much should we trust llm-based measures for accounting and finance research? *Available at SSRN*, 2024.
- [103] Fengzhu Zeng and Wei Gao. Justilm: Few-shot justification generation for explainable fact-checking of real-world claims. *Transactions of the Association for Computational Linguistics*, 12:334–354, 2024.
- [104] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.
- [105] Yuji Zhang, Sha Li, Jiateng Liu, Pengfei Yu, Yi R Fung, Jing Li, Manling Li, and Heng Ji. Knowledge overshadowing causes amalgamated hallucination in large language models. *arXiv preprint arXiv:2407.08039*, 2024.
- [106] Yuji Zhang, Sha Li, Cheng Qian, Jiateng Liu, Pengfei Yu, Chi Han, Yi R Fung, Kathleen McKeown, Chengxiang Zhai, Manling Li, et al. The law of knowledge overshadowing: Towards understanding, predicting, and preventing llm hallucination. *arXiv preprint arXiv:2502.16143*, 2025.
- [107] Xin Zhou, Yi Lu, Ruotian Ma, Tao Gui, Qi Zhang, and Xuanjing Huang. Making harmful behaviors unlearnable for large language models. *arXiv preprint arXiv:2311.02105*, 2023.

- [108] Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023.
- [109] Kaiwen Zuo and Yirui Jiang. Medhallbench: A new benchmark for assessing hallucination in medical large language models. *arXiv preprint arXiv:2412.18947*, 2024.