PENINSULA PUBLISHING LLC

EDRAAK
Vol. **(2026)**, 2026, **pp**. 41–51
ISSN: 3078-8412

Research Article

AI Hallucinations in Retrieval-Augmented and Generative Systems: A Rigorous Review of Definitions, Failure Mechanisms, Evaluation, and Mitigation Strategies

Toufik Mzili[1, *], Ilyass Mzili[1], Ahmed Abatal[2], Zahra Oughannou[3], Adarsh Kumar Arya[4], Mohamed Kurdi [5]

[1] *ELITES Lab, Department of Computer Science, University of Chouaib Doukkali, Faculty of Sciences, EL Jadida, Morocco*

[2] *Faculty of Juridical, Economic and Social Sciences, Chouaib Doukkali University, El Jadida, Morocco*

[3] *Innovation in Mathematics and Intelligent Systems (IMIS) Laboratory, Ibn Zohr University, Agadir, Morocco*

[4] *Department of Chemical Engineering, Harcourt Butler Technical UniversityThe institution will open in a new tab, Uttar Pradesh , Kanpur, 208002, India*

[5] *Faculty of Informatics Engineering, Al-Shamal Private University, Idlib, Syrian Arab Republic*

ARTICLEINFO

Article History

Received 20 Dec 2025

Revised: 10 Feb 2026
Accepted 12 Mar 2026
Published 27 Mar 2026

Keywords

AI hallucination,

retrieval-augmented generation,

large language models,

grounding,

hallucination detection,

evaluation metrics,

trustworthy AI,

mitigation strategies,

human-AI trust.

ABSTRACT

The research on AI hallucinations undergoes complete evaluation through this review which studies retrieval-augmented generation (RAG) and large language models (LLMs) and multimodal systems and their applications in healthcare and education and law and cybersecurity and business and tourism. The paper uses the bibliography from the source document as its review material to combine different definitions and show the reasons for hallucinations at both the model stage and the pipeline stage and to evaluate detection methods and assessment approaches and to develop a functional system for mitigation methods. Research shows that user trust along with interface behavior and anthropomorphic design and legal accountability and regulatory oversight now determine how hallucinated output systems impact actual operational systems. Three conclusions emerge. The research identifies hallucination as a group of different errors which produce unsupported content and weak grounding and context conflict and false citations and misleading confidence. The research identifies hallucination as a group of different errors which produce unsupported content and weak grounding and context conflict and false citations and misleading confidence. The RAG system implements security features which prevent particular failure modes but it generates new failure modes through its retrieval quality and evidence selection and grounding fidelity systems. The RAG system removes certain failure types but it produces new problems which impact both retrieval operations and evidence selection and grounding process accuracy. A stack which unites corpus control with retrieval validation and constrained generation and uncertainty quantification and human review stands as the most defensible solution for high-risk operational deployment. The most defensible approach for high-risk deployment requires a combination of corpus control and retrieval validation and constrained generation and uncertainty quantification and human review processes. The paper ends with three main recommendations which include operational metrics and research.

1. INTRODUCTION

The primary reliability issue with Generative AI systems today is the occurrence of hallucinations. A phenomenon described as outputs that look professional but present information without evidence support and show made-up components or incorrect data and other misleading content [1]. The available text database proves that this problem has extended from machine learning to scientific writing along with clinical practice and educational environments and cybersecurity operations and legal systems and marketing activities and tourism management and crisis response operations [2][3].

The scientific evidence shows that hallucination occurs because of more than basic factual mistakes. Different researchers consider hallucination to be an extensive epistemic and sociotechnical problem that causes users to trust fake references together with incorrect source credits and inconsistent information and fake proof sequences and user interface

*Corresponding author email: Mzili.t@ucd.ac.ma
DOI: https://doi.org/10.70470/EDRAAK/2026/004

manipulations that deceive users into trusting unverified responses [4],[5]. The comprehensive assessment stands as a critical step because false positives which users detect during their operations will generate real-world harm.

The article proposes that researchers need to investigate hallucination occurrences in generative systems by studying three different but connected levels of analysis. The first analysis occurs at the model level because probabilistic next-token generation systems select plausible tokens instead of generating factually accurate ones. The second level operates through the pipeline system which uses retrieval and ranking and chunking and prompt construction and tool integration and decoding decisions to produce unfounded outputs [8]. The third is the human-governance level, where trust, anthropomorphism, responsibility attribution, and regulatory context determine whether generated misinformation is ignored, corrected, or acted upon [6], [7].

The paper presents new findings which go beyond the existing understanding that AI systems produce hallucinated information. The review transforms the given bibliography into a scholarly publication focus while achieving four main objectives which include: defining competing definitions and revealing the operational systems which maintain hallucinations in LLM and RAG environments and establishing an evaluation and mitigation method structure and pinpointing areas where research findings remain insufficient and disorganized and lack robust methodologies [9], [10].

2. REVIEW DESIGN AND SCOPE

The current article uses a structured narrative-review approach to organize its content. The review corpus consists of the seventy references provided in the source Word file. Because the input bibliography spans journal papers, conference papers, preprints, essays, book chapters, and commentary pieces, a formal meta-analysis would not be methodologically appropriate [11]. The research literature requires narrative synthesis because it contains conceptual discussions and real-world case studies and domain-based warning studies and technical solution proposals which do not follow a standard experimental format [12].

The review process proceeded in four stages. First, each reference was coded for year, apparent domain, and major thematic focus. The research papers were organized into four main groups which included conceptual papers and technical papers and human-factor studies and application-based research [13]. The study examined repeated statements throughout the research material to identify both shared and conflicting evidence between various information sources [14]. The literature was transformed into a failure analysis framework to their visible symptoms and assessment parameters and protection systems. The design maintains the paper's connection to the provided bibliography while it develops an organized analytical framework [15].

The main restriction needs to be directly stated. The corpus contains extensive content but it lacks consistent empirical information throughout its material. The sources include experimental data from some sources and others present structured evaluations but other sources contain reflective essays and position papers and conceptual critical reviews. The review therefore does not treat every source as evidentially equivalent [16]. The research employs technical surveys and empirical investigations and domain analyses to establish robust descriptive evidence while commentary pieces function to document ongoing debates about definitions and ethical problems and management effects [17]

The review foundation becomes clear through Table I which displays the composition of the corpus together with its analytical dimensions and the boundaries which control the review process.

TABLE I. REVIEW PROTOCOL AND CORPUS PROFILE.

Element	Description	Evidence Basis	Interpretive Note
Corpus size	70 references from the supplied bibliography	Direct count from source file	Sufficient for broad conceptual synthesis, but heterogeneous in method
Time span	2023–2026, with strongest concentration in 2024–2025	Bibliographic coding	Indicates rapid growth and recency of concern
Main source types	Journal articles, conference papers, preprints, essays, book chapters	Manual classification	Suggests uneven empirical maturity across the field
Core analytical axes	Definition, causation, evaluation, mitigation, human factors, governance, domain applications	Thematic clustering	Used to structure the review and avoid purely descriptive listing

As Table I indicates, the review corpus is broad enough to support conceptual synthesis, yet heterogeneous enough to require caution when comparing empirical claims across sources.

3. BIBLIOGRAPHIC STRUCTURE OF THE CORPUS

Before starting your main synthesis work you need to establish the characteristics which define the bibliography as a whole. The publication pattern shows a sharp increase after 2023, reflecting the wider diffusion of generative AI systems into public, academic, and professional settings [18]. The corpus contains studies from various academic fields because its core technical articles exist alongside research about ethical concerns and trust problems and legal responsibility and educational

uses and clinical practices and particular risks in industry sectors. The field faces obstacles in developing its terminology because its concepts range across multiple aspects which include unreliable results as a major reliability issue [19], [20]

3.1 Publication trend

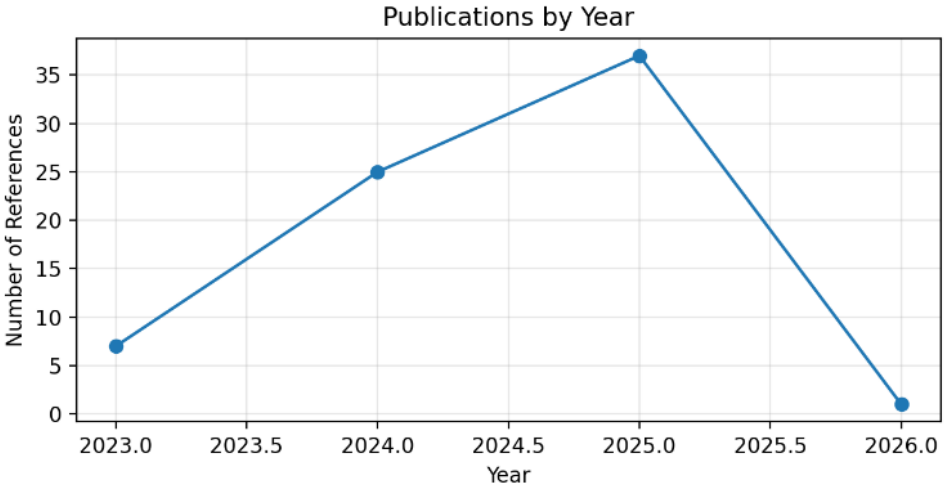

Fig. 1. Publications in the supplied corpus by year.

Figure 1 is used here to visualize the temporal growth of the literature included in the review, particularly the concentration of publications in the most recent years.

The trend shown in Figure 1 supports the claim that hallucination research accelerated sharply after 2023, reflecting the rapid public and institutional diffusion of generative AI systems.

3.2 Domain distribution

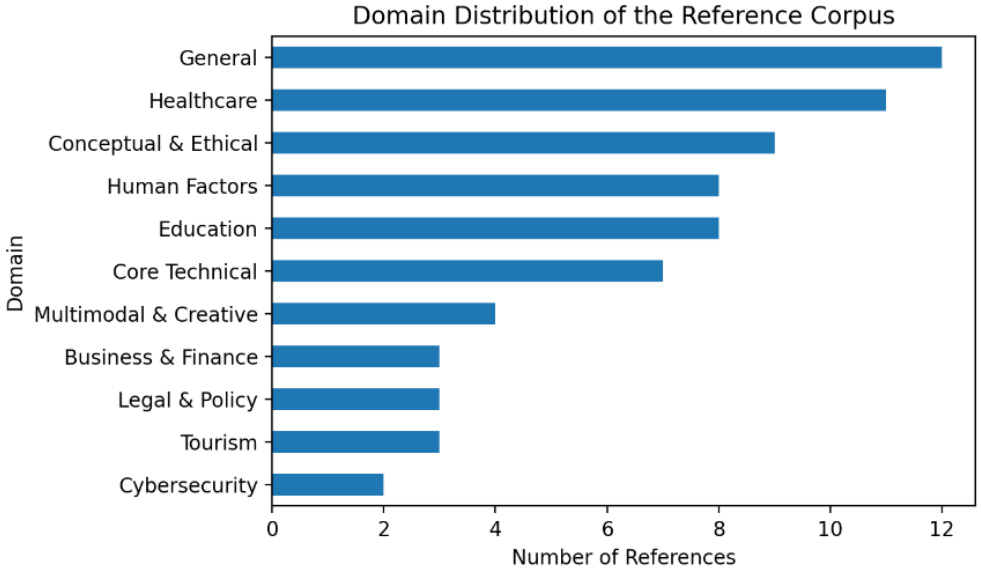

Fig. 2. Domain distribution of the supplied reference corpus.

Figure 2 shows how the reviewed studies are distributed across application domains, making it easier to see that hallucination research is no longer confined to core computer-science venues.

Figure 2 confirms that the corpus is multidisciplinary, with technical studies appearing alongside work in healthcare, education, law, tourism, and other applied settings.

3.3 Theme distribution

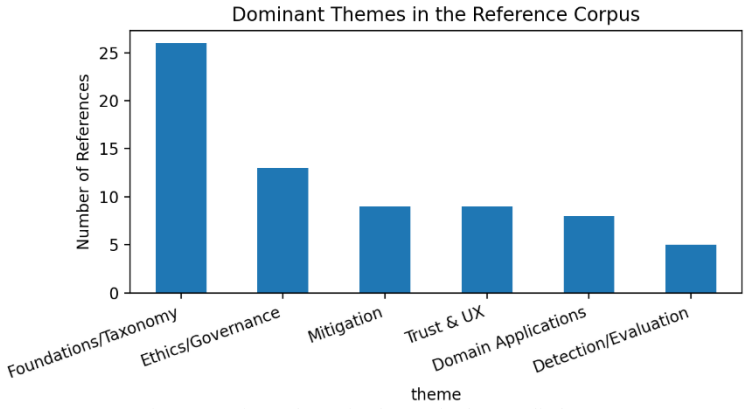

Fig. 3. Dominant thematic clusters in the supplied corpus.

Figure 3 highlights the dominant thematic clusters in the corpus and helps explain why the review is organized around definition, causation, evaluation, mitigation, and governance.

As Figure 3 suggests, the literature is clustered around a small set of recurring concerns, especially definitions, evaluation, mitigation, and governance, rather than around a single unified theory.

4. CONCEPTUALIZING HALLUCINATION

The academic community continues to face an unresolvable problem because of varying definitions between different research publications. The term hallucination receives different definitions from various authors who restrict it to false information and references but others include unsupported reasoning and contextual inconsistencies and fictional examples and perceptual inaccuracies and outputs that appear semantically correct but lack external validation [8], [21]. The definition differences between these terms create more than language management problems because they influence how benchmarks get developed and how reports are made and how mitigation strategies appear to perform.

Multiple academic articles challenge metaphors through their theoretical work. Some researchers believe that using the term 'hallucination' to describe statistical systems creates human characteristics which might prevent us from understanding how probabilistic generators create convincing yet unsubstantiated content [22]. The term remains in use because it effectively identifies the difference between basic language skills and the evidence needed to support statements when addressing public and policy audiences [23], [24]. The disagreement exists but research shows that unauthorized outputs which appear to have authorization stand as the main operational concern in this field.

The most effective analytical synthesis defines hallucination as content which a system generates with certain semantic confidence but does not have sufficient backing from its corresponding evidence base or environmental context or operational task rules or original source information. The extended definition includes five common subtypes which appear throughout the corpus including false information and false citations or citations and evidence-conflict and reasoning distortion and multimodal misinterpretation [25], [26]. The research method prevents the identification of a singular subsequent effect because it stops the detection of individual signs which would appear during the illness progression [27]. The review adopts the complete definition because it supports both LLM-only operations and retrieval-augmented systems. The system demonstrates knowledge through its communication but its supporting evidence contains no substance and lacks necessary information and shows conflicting data which creates the core issue. A claim which lacks any supporting evidence will not gain credibility through the use of sophisticated language.

Based on the definitional discussion above, Table II organizes the major hallucination subtypes into an operational typology that can support later analysis of causes and controls.

TABLE II. WORKING TYPOLOGY OF HALLUCINATION IN GENERATIVE AND RAG SYSTEMS.

Subtype	Operational description	Typical manifestation	Illustrative sources
fabricated information	Answer contains unsupported entities, values, or events	Invented statistics, names, mechanisms, or causal claims	[8], [18], [26], [67], [69]
fabricated citations	System invents or corrupts source details	Non-existent citations, false authors, wrong volumes/years	[7], [22], [53], [60], [61]
evidence contradiction	Retrieved or supplied evidence contradicts answer	Answer ignores source text or overstates evidence	[4], [26], [39], [63], [68]
reasoning distortion	Intermediate explanation is internally coherent but not warranted	False chains of reasoning, pseudo-logic, unjustified certainty	[19], [40], [56], [70]
multimodal misinterpretation	Model misreads images/visual cues or adds non-existent details	Visual misclassification, surreal additions, false scene details	[12], [44], [63]

Table II makes clear that hallucination is best treated as a family of failures rather than a single phenomenon, since false information, grounding conflicts, reference errors, and multimodal distortions do not arise or behave in identical ways.

5. WHY HALLUCINATIONS PERSIST IN GENERATIVE AND RAG SYSTEMS

Research literature shows that autoregressive language models achieve their primary function by maximizing token likelihood instead of determining the meaning of statements. The training process enables them to follow existing patterns but it does not ensure the production of verified propositions. A model generates text which appears correct at first glance because it matches answer distribution patterns but it does not verify factual accuracy or maintain causal relationships or authenticate source materials [28], [29], [30].

Training data also matter. The pretraining process for large datasets introduces various issues because the data contains errors and conflicting information and multiple copies of the same content and information that has become obsolete and unresolved contradictory elements. The model parameters used for data compression preserve essential patterns but they fail to create a standardized system for tracing data origins. The structure enables the creation of outputs which seem reliable but their origin cannot be verified through trustworthy evidence [31], [32], [35]

The system becomes more unstable because the process of prompting and decoding adds new elements. The output contains more unsupported information because users submit diverse prompts which lack explicit task instructions and their creative methods become more powerful. Research studies and analysis articles show that prompt engineering as a mitigation strategy produces some benefits but does not fully resolve the issue because ineffective retrieval systems and model states remain problematic despite improved prompt quality [33], [34], [36].

RAG operates as a solution to hallucinations in AI yet the analyzed research demonstrates it operates as a system which moves failure points instead of completely eliminating them. The retrieval process fails because the index lacks essential data and users create incorrect queries and the ranking system operates at a low level and the chunking method breaks down the original context and the final context window reduces the strength of vital evidence. The generator continues to produce fluent text despite it develops incorrect connections with partial and irrelevant and contradictory evidence [37], [38], [39].

There is also a data processing pipeline in AI systems. The operational system depends on data ingestion which leads to indexing and retrieval and reranking and context assembly and generation and citation formatting and interface presentation [40]. A system may retrieve the right evidence but summarize it incorrectly; it may retrieve contradictory evidence and overconfidently collapse uncertainty; or it may produce a weak answer that the interface then frames as authoritative. The current research community supports the idea that hallucination detection should focus on system-level challenges instead of concentrating on model-specific errors [41], [42].

6. DETECTION AND EVALUATION STRATEGIES

Detection and evaluation research has experienced rapid growth because measurement is essential to validate reliability claims. The provided corpus contains taxonomies and classification frameworks together with rate estimation studies and domain-specific assessment papers and general surveys which agree that evaluation needs separate different failure types instead of using a generic accuracy metric [43], [44].

One major evaluation family focuses on grounding. The evaluation process needs to determine if the answer derives its support from the retrieved passages together with the original documents and visual content and established domain regulations. The family requires special attention because RAG and legal and medical and scholarly applications need protection against citation management which create more severe risks than typical factual errors [45]

The evaluation process includes a second family which focuses on calibration and uncertainty assessment. A model which states 'I am not sure' during weak evidence situations operates in a completely different way than a model which tries to improvise with strong confidence. The human-factors literature demonstrates that users will accept or question hallucinated output based on how uncertainty is presented through wording choices and politeness levels and attribution methods and interface design elements [46].

The third evaluation family examines how tasks and harm relate to each other. The main problem in high-stakes environments extends beyond detecting unsupported fragments because these fragments also impact how downstream actions and professional judgments and safety decisions are performed. Healthcare and crisis-response operations and legal research and cybersecurity studies need dedicated methodological approaches because their formal hallucination rates produce distinct real-world risk patterns according to studies [47], [48].

To avoid treating hallucination as a single measurable defect, Table III compares the main evaluation lenses used across the literature and the contexts in which each lens is most appropriate.

TABLE III. MAJOR EVALUATION LENSES FOR HALLUCINATION STUDIES.

Evaluation lens	What it measures	Suitable contexts	Representative sources
Grounding / support	Whether claims are supported by evidence or retrieved passages	RAG, legal, medical, scholarly systems	[7], [24], [42], [53], [60]

Evaluation lens	What it measures	Suitable contexts	Representative sources
Hallucination rate	Frequency of unsupported or fabricated items	Benchmarking and model comparison	[8], [10], [67], [69]
Calibration / uncertainty	Whether confidence matches evidential strength	Decision support, user-facing assistants	[14], [21], [50], [64]
Task harm / consequence	Impact of errors on downstream action or safety	Healthcare, crisis, cybersecurity, law	[4], [17], [37], [45], [48], [64]

The comparison in Table III shows that reliable evaluation depends on matching the metric to the deployment context; a simple hallucination rate is rarely sufficient in high-stakes domains.

6.1 Proposed operational equations for future empirical work

The reviewed corpus uses overlapping but inconsistent metrics which require the following operational equations to serve as a concise evaluation toolkit for upcoming empirical research. The devices function as reporting tools which exist to create order in study comparison without serving as official standards.

The first equation formalizes the overall frequency of generated misinformation in the evaluated outputs.

$$\mathrm{HR} = \frac{\mathrm{H}}{\mathrm{N}}$$

where H is the number of hallucinated outputs or hallucinated claims and N is the total number of evaluated outputs or claims.

This expression provides the most direct estimate of unreliable results frequency and is useful for coarse benchmarking across models, prompts, or datasets.

The second equation measures the proportion of claims that can actually be supported by evidence.

$$\mathrm{ESR} = \frac{\mathrm{C}}{\mathrm{S}}$$

where S is the number of supported claims and C is the total number of claims extracted from a response.

The ESR metric complements raw error counting by emphasizing how much of a response is positively grounded in available evidence rather than merely how much is wrong.

The third equation is designed for retrieval-augmented settings in which answer quality depends on both retrieval and faithful use of evidence.

$$\mathrm{RGR} = \mathrm{Rk} \times \mathrm{Pe} \times \mathrm{Fc}$$

where R_k is retrieval recall at top-k, P_e is evidence precision, and F_c is citation fidelity or support fidelity.

The RGR formulation is especially relevant for RAG pipelines because it combines retrieval quality, evidence precision, and citation fidelity into one interpretable reliability signal.

The fourth equation extends basic counting by incorporating the seriousness of different error types.

$$\mathrm{RWHI} = \mathrm{N}\sum\mathrm{i} = 1\mathrm{mwiei}$$

where e_i is an error event, w_i is the harm weight assigned to that error type, and N is the number of evaluated outputs.

Unlike the previous metrics, the RWHI captures consequence severity by weighting errors according to their practical harm, which makes it more appropriate for high-stakes domains.

7. MITIGATION AND CONTROL STRATEGIES

The corpus contains three distinct layers of mitigation strategies which include pre-generation controls and generation-time controls and post-generation controls. The first layer contains seven processes which include corpus curation and source validation and de-duplication and document segmentation and query reformulation and retrieval improvement and prompt design. The second includes constrained generation, consensus or self-checking procedures, structured reasoning prompts, and mechanisms that force the model to remain close to supplied evidence. The third includes fact-checking, citation verification, contradiction detection, abstention, and human review [49], [50].

Multiple studies now support layered mitigation methods instead of depending on single-point solutions. The consensus approach helps minimize certain types of false results but it will strengthen existing biases when all possible answers stem from weak sources [51]. The process of prompt engineering provides assistance yet it fails to fix problems which occur during corpus selection and retrieval operations. Human review provides benefits yet its success depends on users who have received proper training and maintain focus and use interfaces which display evidence instead of concealing it with sophisticated language [52]

The corpus delivers its most essential message which states that organizations should develop their mitigation plans according to the threat level present in their specific operational areas. The use of partial fictionalization becomes suitable in low-risk creative assignments when it either meets the requirements or produces better results [53]. The system needs to perform four core functions during legal research and clinical education and crisis-response and cybersecurity operations which include abstention and citation and evidence exposure and uncertainty escalation instead of using deceptive methods to handle missing information [54], [55].

Because no single intervention can fully eliminate hallucination, Table IV presents a layered mitigation stack that links technical controls with organizational oversight.

TABLE IV. LAYERED MITIGATION STACK FOR TRUSTWORTHY DEPLOYMENT.

Layer	Main controls	Strength	Key limitation
Pre-generation	Corpus curation, indexing quality, chunking, query reformulation, reranking	Reduces weak grounding before answer generation	Does not guarantee correct evidence use
Generation-time	Constrained prompting, consensus, self-checking, grounded decoding	Can lower unsupported elaboration	Still depends on evidence quality and calibration
Post-generation	Fact-checking, citation verification, contradiction detection, human review	Catches residual failures and enables escalation	Adds latency, cost, and reviewer burden
Governance layer	Policies, disclosure, audit logs, role-based controls, abstention rules	Aligns system behavior with risk and accountability	Requires organizational discipline, not just model tuning

Table IV demonstrates that trustworthy deployment depends on defense in depth: stronger retrieval and prompting help, but post-generation checks and governance controls remain indispensable.

8. DOMAIN-SPECIFIC FINDINGS

The bibliography shows healthcare as a leading field which maintains a careful tone throughout its literature. The use of misleading explanations together with unsupported recommendations and false citations produces educational distortions which affect medical reasoning and literature review work and educational processes. The regulatory and safety literature requires health applications to meet higher evidence standards and oversight protocols than typical consumer applications do [57], [58].

Research in education reveals that two distinct elements operate within the system. The system enables users to develop content while it provides them with options to check their finished work and find additional information. Students and trainees fail to detect fake content because it appears as well-written text which offers supposed assistance [59]. The research demonstrates this conflict by examining academic writing and university communities and studying teaching approaches and student skills to detect incorrect results [60].

Legal and policy-oriented sources treat hallucination as an accountability problem rather than merely a quality nuisance. The creation of fictitious cases and the presentation of baseless legal arguments together with disputes over data ownership create problems which affect both liability and due diligence requirements and professional conduct violations [61]. The legal literature provides essential insight because it shows how identical technical errors transform into different institutional problems when they enter areas which follow established rules and maintain official records and assign specific operational duties [62].

Cybersecurity and finance and marketing and tourism and crisis-response studies reveal that hallucination effects decision-making abilities through negative impacts on choice accuracy although medical professionals do not need to make life-threatening decisions [63]. The classification process of cybersecurity systems produces incorrect threat detections while these systems generate non-existent attack scenarios; financial and business software applications generate misleading analytical data which appears more reliable than it actually is; tourism and marketing platforms manipulate customer decisions through misleading information; and crisis management systems produce dangerous situations for users who need immediate responses [64], [65].

A small but important subset of the corpus also explores the positive or creative side of hallucination. Artists and visual researchers and creative individuals and specific tourism activities benefit from departing from exact factual accuracy [67]. The literature provides value because it demonstrates that hallucinations can have positive value when the work requires creative thinking instead of absolute precision. The combination of these two systems without defined separation points leads to various problems [66].

To connect the technical discussion with real-world deployment, Table V compares the main risk patterns across domains and the level of safeguard intensity that each domain demands.

TABLE V. DOMAIN-SPECIFIC RISKS AND RELIABILITY EXPECTATIONS.

Domain	Dominant risk	Expected safeguard intensity	Representative sources
Healthcare	poor guidance, false data, misleading educational content	Very high	[1], [17], [24], [45], [48], [54]
Education	Academic misuse, false learning, weak source verification	High	[2], [24], [31], [36], [47]
Law / policy	Fabricated cases, liability, rights violations	Very high	[13], [32], [60], [61]
Cybersecurity	false positive, weak incident reasoning	High	[4]
Business / finance / marketing	cognitive bias	Moderate to high	[29], [51], [65]

Domain	Dominant risk	Expected safeguard intensity	Representative sources
Tourism / consumer contexts	Distorted decision-making through plausible false detail	Moderate	[22], [23], [66]

As shown in Table V, the tolerance for hallucination is not uniform across domains; the closer a system is to safety, legality, or institutional decision-making, the lower the acceptable error margin becomes.

9. HUMAN FACTORS, TRUST, AND GOVERNANCE

Researchers in current literature focus on studying model hallucinations and the human responses that emerge when these systems produce incorrect information. Users experience hallucinations through their interactions with system interfaces which include both visual elements and audio feedback and text-based messages and legal notices and academic references and community-established behavioral standards. Researches on anthropomorphism and fear of AI and trust and politeness and privacy concern and user empowerment and continued-use intention demonstrate that human responses either minimize or maximize the operational effects of model inaccuracies [68].

The interface design establishes an illusion of expertise which affects this situation. The system presents itself as calm and polite yet confident in its statements which helps users trust its output despite insufficient supporting evidence. The style of a response determines which generated misinformation will appear according to multiple sources making reliability an interaction-design issue that goes past modeling challenges [69].

The corpus contains governance discussions which focus on disclosure requirements and traceability systems and role segregation and audit requirements and regulatory compliance monitoring. The academic writing from healthcare and legal fields shows that model output should not function as an independent decision-making tool for institutions. The systems need to show evidence while they keep records and enable users to verify information and they must specify when users must refuse to participate [70], [60].

Organizations need to build procedures which will identify bluffing activities when they use models which contain bluffing capabilities. The current situation shows that technology without proper governance becomes an expensive form of unsubstantiated claims.

10. RESEARCH GAPS AND FUTURE DIRECTIONS

The field shows inconsistent methodological approaches despite its fast-growing number of published works. First, benchmark fragmentation persists. Multiple research papers employ various methods to determine error quantities because they use different error definitions and their evaluation tasks contain distinct risk factors. The research community encounters obstacles when trying to match headline results about hallucination reduction across different academic studies [67], [68], [69].

Second, many studies remain too model-centric. The authors present hallucination as a generator issue yet the actual operational errors stem from data processing stages and retrieval system performance and evidence selection methods and interface design and user comprehension of information. RAG studies should therefore report retrieval recall, evidence precision, grounding fidelity, abstention behavior, and harm-sensitive outcomes together rather than in isolation [4], [26].

Research on longitudinal adaptation remains insufficient because most of the existing work fails to meet required quality standards. According to various user studies people develop either increased skepticism or improved skills when using products yet other research indicates users continue to rely on products because of their ease of use and smooth operation. The effectiveness of repeated exposure in enhancing hallucination detection requires additional long-term studies which should take place in natural settings to determine its true impact [49].

Fourth, regulatory and institutional questions are outrunning technical clarity. The number of studies focusing on legal matters and rights-based approaches and integrity protection has increased yet the development of formal standards for error tolerance and domain-specific abstention and evidence presentation remains insufficient [45], [60].

A research agenda which has developed further requires benchmark suites that span multiple layers and domain-weighted harm metrics and mitigation stack comparisons instead of individual techniques and system testing with actual user scenarios instead of benchmark prompts. Research in the future needs to develop particular rules which determine when creative works should include fiction and when vital content requires validation through refusal or verification or escalation procedures [62], [64], [66].

11. CONCLUSION

The research materials reviewed in this section establish that hallucination problem as a major problem which continues to exist after retrieval systems have been implemented. The system faces multiple reliability problems because stochastic generation errors which are worsened by noisy data and retrieval mismatches and insufficient evidence handling and human operators who place excessive trust in the system. The corpus shows that hallucination severity changes between different

fields because creative domains permit more acceptance of these errors than medical settings and legal environments and cybersecurity operations and crisis management work require.

The literature supports its strongest answer through methods which do not depend on naive assumptions to establish a model without hallucinations. The system requires disciplined design which includes curated data and strong retrieval capabilities and explicit evidence display and constrained generation and uncertainty quantification and post-hoc verification and accountable human oversight. The stack operates with more basic functionality than showing an AI system which functions without errors. The system operates through an approach which matches the actual process of building trustworthy systems.

In short, the field needs fewer vague promises and more rigorous measurement. Fluency is cheap. Grounded reliability is the real work.

Funding:

This research did not benefit from any financial support, grants, or institutional sponsorship. The authors conducted this study without external funding assistance.

Conflicts of Interest:

The authors declare that they have no conflicting interests.

Acknowledgment:

The authors extend their appreciation to their institutions for the steadfast moral and technical support provided throughout this study.

References

[1] A. Cheng, V. N. Cheng, S. Eller, D. G. Cheng, and Y. Lin, “Exploring AI hallucinations of ChatGPT,” *Simulation in Healthcare*, 2025, doi: 10.1097/sih.0000000000000877.

[2] A. Cheng, A. W. C. Cheng, and G. Reedy, “Artificial intelligence-assisted academic writing: Recommendations for ethical use,” *Advances in Simulation*, 2025, doi: 10.1186/s41077-025-00350-6.

[3] A. Gadiko, “Understanding and addressing AI hallucinations in healthcare and life sciences,” *International Journal of Health Sciences*, vol. 7, no. 3, pp. 1–11, 2024.

[4] A. K. Sood, S. Z., and E. Hong, “The paradigm of hallucinations in AI-driven cybersecurity systems: Understanding taxonomy, classification outcomes, and mitigations,” *Computers & Electrical Engineering*, 2025, doi: 10.1016/j.compeleceng.2025.110307.

[5] A. Tlili and D. B., “AI hallucinations? What about human hallucination?! Addressing human imperfection for ethical AI,” *International Journal of Interactive Multimedia and Artificial Intelligence*, 2025, doi: 10.9781/ijimai.2025.02.010.

[6] A. Rapp, C. D. L., and L. Di, “How do people react to ChatGPT’s unpredictable behavior? Anthropomorphism, uncanniness, and fear of AI,” *International Journal of Human-Computer Studies*, 2025, doi: 10.1016/j.ijhcs.2025.103471.

[7] A. Adel and N. H. S., “Can generative AI reliably synthesise literature? Exploring hallucination issues in ChatGPT,” *AI & Society*, 2025, doi: 10.1007/s00146-025-02406-7.

[8] A. Jesson *et al.*, “Estimating the hallucination rate of generative AI,” *arXiv*, 2024, doi: 10.48550/arxiv.2406.07457.

[9] A. Mills and N. A., “Are we tripping? The mirage of AI hallucinations,” *SSRN Electronic Journal*, 2025, doi: 10.2139/ssrn.5127162.

[10] A. Jančařík and O. D., “The problem of AI hallucination and how to solve it,” in *Proc. European Conf. e-Learning*, 2024, doi: 10.34190/ecel.23.1.2584.

[11] A. Banafa, “AI hallucinations,” in *River Publishers eBooks*, 2025, doi: 10.1201/9788770046213-5.

[12] B. A. Halperin and S. M. L., “Artificial dreams: Surreal visual storytelling as inquiry into AI hallucination,” in *Proc. Designing Interactive Systems Conf.*, 2024, doi: 10.1145/3643834.3660685.

[13] T. Christakis, “AI hallucinations and data subject rights under the GDPR,” *SSRN Electronic Journal*, 2025, doi: 10.2139/ssrn.5042191.

[14] C. Jacob, P. K., and M. Bastos, “The chat-chamber effect: Trusting the AI hallucination,” *Big Data & Society*, 2025, doi: 10.1177/20539517241306345.

[15] C. Jacob, P. K., and M. Bastos, “The chat-chamber effect: Trusting the AI hallucination,” *SSRN Electronic Journal*, 2024, doi: 10.2139/ssrn.5033125.

[16] C. Lee, J. K., and J. S. Lim, “Generative AI risks and resilience: User adaptation to hallucination and privacy challenges,” *Telematics and Informatics Reports*, 2025, doi: 10.1016/j.teler.2025.100221.

[17] V. Geroimenko, “Generative AI hallucinations in healthcare,” in *Springer Series on Cultural Computing*, 2025, doi: 10.1007/978-3-031-86551-0_17.

[18] S. Greengard, “Shining a light on AI hallucinations,” *Communications of the ACM*, 2025, doi: 10.1145/3715691.

[19] O. H. Hamid, “Beyond probabilities: LLMs and AI hallucination,” 2024, doi: 10.1109/cogsima61085.2024.10553755.

[20] H. Namazi and M. H. R., “Philosophy of medicine meets AI hallucination and AI drift,” *Journal of Medical Ethics and History of Medicine*, 2025, doi: 10.18502/jmehm.v18i2.18812.

[21] H. Kim and S. W. Lee, "Sorry, it's my fault: Politeness, attribution, and anthropomorphism in managing generative AI hallucinations," *International Journal of Information Management*, vol. 86, p. 102996, 2026, doi: 10.1016/j.ijinfomgt.2025.102996.

[22] İ. Önder and S. M., "How AI hallucinations threaten research integrity in tourism," *Annals of Tourism Research*, 2025.

[23] J. Christensen, J. M. H., and P. Wilson, "Understanding the role and impact of generative AI hallucination in tourism decision-making," *Current Issues in Tourism*, 2024, doi: 10.1080/13683500.2023.2300032.

[24] J. Zhou *et al.*, "Integrating AI into clinical education: Evaluating trainees' ability to distinguish AI-generated hallucinations," *BMC Medical Education*, 2025, doi: 10.1186/s12909-025-06916-2.

[25] J. Slater and J. H., "Another reason to call bullshit on AI 'hallucinations'," *AI & Society*, 2025, doi: 10.1007/s00146-025-02346-2.

[26] N. Jones, "AI hallucinations can't be stopped — but these techniques can limit their damage," *Nature*, 2025, doi: 10.1038/d41586-025-00068-5.

[27] J. S. Lim, D. S., C. Lee, J. K., and J. Zhang, "User empowerment, AI hallucination, and privacy concerns in generative AI adoption," *Journal of Broadcasting & Electronic Media*, 2025, doi: 10.1080/08838151.2025.2487679.

[28] J. Dumit and A. R., "AI hallucinations are a feature of LLM design, not a bug," *Nature*, 2025, doi: 10.1038/d41586-025-00662-7.

[29] S. Joshi, "Comprehensive review of AI hallucinations: Impacts and mitigation strategies," *International Journal of Computer Applications Technology and Research*, 2025.

[30] J. P. Nambiar and A. G. S., "Orchestrating consensus strategies to counter AI hallucination in generative chatbots," 2023, doi: 10.1109/ccem60455.2023.00030.

[31] H. Kamel, "Understanding the impact of AI hallucinations on the university community," *Cybrarians Journal*, 2024, doi: 10.70000/cj.2024.73.622.

[32] A. M. Khawaldeh, "Generative AI hallucinations and legal liability in Jordanian civil courts," *International Journal for the Semiotics of Law*, 2024, doi: 10.1007/s11196-024-10199-z.

[33] H. Kim, "Investigating the effects of generative-AI responses on user experience after hallucination," 2024, doi: 10.20319/icssh.2024.92101.

[34] M. R. King, "An update on AI hallucinations: Not as bad as you remember," *Cellular and Molecular Bioengineering*, 2025, doi: 10.1007/s12195-025-00874-x.

[35] P. Kollias, "Nostophiliac AI: Artificial collective memories and hallucinations," *Memory Studies Review*, 2024, doi: 10.1163/29498902-202400014.

[36] R. Lane, "Mitigating generative AI hallucinations in geographical education," *International Research in Geographical and Environmental Education*, 2025, doi: 10.1080/10382046.2025.2555185.

[37] L. J. Jin, Z. S., and S. B. N. Alhur, "Determinants and effects of AI hallucination exposure on adoption in healthcare," *Information Development*, 2025, doi: 10.1177/02666669251340954.

[38] M. Ashwin, S. J., and S. K. Ganga Prasad, "AI hallucinations and ethical dilemmas in anesthesia research," *Journal of Anaesthesiology Clinical Pharmacology*, 2025, doi: 10.4103/joacp.joacp_56_25.

[39] L. Meng, "Architecting trustworthy LLMs: A unified framework for mitigating hallucination," *Journal of Computer Science and Frontier Technologies*, 2025, doi: 10.63313/jcsft.9019.

[40] M. T. Hicks, J. Humphries, and J. Slater, "ChatGPT is bullshit," *Ethics and Information Technology*, 2024, doi: 10.1007/s10676-024-09775-5.

[41] M. Hegazy, "Evolution of AI role in architectural design: Between parametric exploration and hallucination," *MSA Engineering Journal*, 2023, doi: 10.21608/msaeng.2023.291873.

[42] M. Z. Ahmad, I. Y., and T. D. Roy, "Creating trustworthy LLMs: Dealing with hallucinations in healthcare AI," *Preprints*, 2023, doi: 10.20944/preprints202310.1662.v1.

[43] N. Maleki, B. P., and K. Dutta, "AI hallucinations: A misnomer worth clarifying," 2024, doi: 10.1109/cai59869.2024.00033.

[44] M. T. Olson, "Error as insight: AI hallucinations and pedagogical possibilities," *Journal of Visual Literacy*, 2025, doi: 10.1080/1051144x.2025.2569015.

[45] O. Freyer, I. C. W., and J. N. Kather, "Future role of health applications of LLMs depends on safety standards," *The Lancet Digital Health*, 2024, doi: 10.1016/s2589-7500(24)00124-9.

[46] H. Patel, "Bane and boon of hallucinations in generative AI," 2024, doi: 10.36227/techrxiv.171198062.20183635/v1.

[47] R. C. Torres, "AI hallucination in education: College students' use of generative AI," 2025, doi: 10.1109/ic4e65071.2025.11075444.

[48] R. Hatem, B. S., and J. Thornton, "Addressing AI hallucinations and mitigation strategies in healthcare," *Cureus*, 2023, doi: 10.7759/cureus.44720.

[49] R. Massenon, I. G., J. A. Khan, C. A., and A. Alwadain, "'My AI is lying to me': User-reported LLM hallucinations," *Scientific Reports*, 2025, doi: 10.1038/s41598-025-15416-8.

[50] R. Pak, E. R., and A. C. McLaughlin, "Polite AI mitigates user susceptibility to hallucinations," *Ergonomics*, 2024, doi: 10.1080/00140139.2024.2434604.

[51] S. Roychowdhury, "Journey of hallucination-minimized generative AI solutions for financial decision makers," 2024, doi: 10.1145/3616855.3635737.

[52] S. Williamson and V. R. P., "The era of artificial intelligence deception: Unraveling the complexities of false realities and emerging threats of misinformation," *Information*, 2024, doi: 10.3390/info15060299.

[53] S. A. Athaluri, S. V. M., V. S. R. K. Manoj Kesapragada, V. Y., and T. D. S. Dave, "Exploring the boundaries of reality: Investigating the phenomenon of artificial intelligence hallucination in scientific writing through ChatGPT references," *Cureus*, 2023, doi: 10.7759/cureus.37432.

[54] M. Sallam, "ChatGPT utility in healthcare education, research, and practice: Systematic review on the promising perspectives and valid concerns," *Healthcare*, 2023, doi: 10.3390/healthcare11060887.

[55] S.-S. Lee, S. L., and S. Lee, "Big data analysis on AI hallucination: Focusing on LDA topic modeling and sentiment analysis," *Korean Journal of Industry Security*, 2024, doi: 10.33388/kais.2024.14.2.151.

[56] A. Shao, "New sources of inaccuracy? A conceptual framework for studying AI hallucinations," *Harvard Kennedy School Misinformation Review*, 2025, doi: 10.37016/mr-2020-182.

[57] Sofience, "SΔϕ-39 — Fiction, hallucination, and the uninhabited construction: Why AI does not stand on the fictions it generates (v1.0)," *Zenodo*, 2026, doi: 10.5281/zenodo.18901097.

[58] T. Chakraborty and S. M., "The Promethean dilemma of AI at the intersection of hallucination and creativity," *Communications of the ACM*, 2024, doi: 10.1145/3652102.

[59] V. Perov and N. P., "AI hallucinations: Is 'artificial evil' possible?," 2024, doi: 10.1109/usbereit61901.2024.10584048.

[60] V. Magesh, F. S., M. Dahl, M. S., and C. D. Manning, D. E. H., "Hallucination-free? Assessing the reliability of leading AI legal research tools," *Journal of Empirical Legal Studies*, 2025, doi: 10.1111/jels.12413.

[61] V. Magesh, F. S., M. Dahl, M. S., and C. D. Manning, D. E. H., "Hallucination-free? Assessing the reliability of leading AI legal research tools," *arXiv*, 2024, doi: 10.48550/arxiv.2405.20362.

[62] W. Cai and M. G., "Beyond hallucination: Generative AI as a catalyst for human creativity and cognitive evolution," *ICCK Transactions on Emerging Topics in Artificial Intelligence*, 2025, doi: 10.62762/tetai.2025.657559.

[63] W. Xiao, Z. H., L. Gan, W. H., H. Li, Z. Y., F. Shu, H. J., and L. Zhu, "Detecting and mitigating hallucination in large vision language models via fine-grained AI feedback," *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025, doi: 10.1609/aaai.v39i24.34744.

[64] Y. Song, M. C., F. Wan, Z. Y., and J. Jiang, "AI hallucination in crisis self-rescue scenarios: The impact on AI service evaluation and the mitigating effect of human expert advice," *International Journal of Human-Computer Interaction*, 2025, doi: 10.1080/10447318.2025.2483858.

[65] B. Yaprak, "Generative artificial intelligence in marketing: The invisible danger of AI hallucinations," *Journal of Economy Business and Management*, 2024.

[66] Y. Wang and R. G., "The bright side of fictional information: Positive impacts of AI hallucination on tourists' cultural contact," *Journal of Retailing and Consumer Services*, 2025, doi: 10.1016/j.jretconser.2025.104588.

[67] Y. Zhang et al., "Siren's song in the AI ocean: A survey on hallucination in large language models," *arXiv*, 2023, doi: 10.48550/arxiv.2309.01219.

[68] Y. Zhang et al., "Siren's song in the AI ocean: A survey on hallucination in large language models," *Computational Linguistics*, 2025, doi: 10.1162/coli.a.16.

[69] Y. Sun, D. S., and Z. Zhou, "AI hallucination: Towards a comprehensive classification of distorted information in artificial intelligence-generated content," *Humanities and Social Sciences Communications*, 2024, doi: 10.1057/s41599-024-03811-x.

[70] K. Šekrst, "Chinese chat room: AI hallucinations, epistemology and cognition," *Studies in Logic, Grammar and Rhetoric*, 2024, doi: 10.2478/slgr-2024-0029.