Cover Page

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY EDUCATIONAL RESEARCH
ISSN:2277-7881(Print); IMPACT FACTOR :9.014(2025); IC VALUE:5.16; ISI VALUE:2.286
PEER REVIEWED AND REFEREED INTERNATIONAL JOURNAL
(Fulfilled Suggests Parametres of UGC by IJMER)
Volume:14, Issue:8(4), August, 2025
Scopus Review ID: A2B96D3ACF3FEA2A
Article Received: Reviewed : Accepted
Publisher: Sucharitha Publication, India
Online Copy of Article Publication Available : www.ijmer.in

FROM HALLUCINATION TO HARM: UNINTENDED ETHICAL RISKS IN LARGE LANGUAGE MODELS

Farha Anjum
Assistant Professor, Department of Intelligent Computing and Business Systems
St Joseph Engineering College, Mangalore

Abstract

The accelerated progress of large language models like GPT-4, Claude, and LLaMA has changed the way natural language processing is done, with machines producing human-like language with uncanny fluency. With that capability comes enormous ethical problems—most notably the hallucination effect, where models create plausible but factually inaccurate or deceptive content. This paper examines the path from harmless hallucinations to concrete harms in everyday contexts, such as spreading misinformation, medical misadvice, legal abuse, and affirming toxic biases. We discuss how these unwanted outputs arise from training materials, model design, and inadequate guardrails, and measure their disproportionate effect on marginalized groups. With a cross-disciplinary approach integrating AI ethics, human-computer interaction, and cognitive psychology, we contend that LLMs not only mirror but reinforce epistemic uncertainties. We also critique current mitigation tactics such as content filtering, prompt engineering, and model fine-tuning and demand the creation of resilient ethical frameworks, transparency norms, and user accountability mechanisms. This research highlights the pressing need for more responsible AI development to avoid the leap from hallucination to harm in progressively high-stakes domains.

Keywords: Large Language Models (LLMs), AI Hallucination, Ethical Risks, Human-AI Interaction, Content Moderation, Transparency, Algorithmic Bias.

1. INTRODUCTION

As big language models become more deeply embedded in core fields, their ability to produce plausible yet false or made-up information—also referred to as hallucinations—raises significant ethical issues with widespread consequences. This article explores the unforeseen ethical threats presented by such cutting-edge AI systems, with emphasis on the imperative of addressing their harm potential in the swiftly changing field of artificial intelligence.

New findings show that model manipulation and adversarial prompt engineering methods can cause LLMs to produce outputs that evade current safety guardrails, exposing systemic weaknesses in existing alignment methods. Further, the absence of established, domain-tailored testbed benchmarks makes it harder to detect and address context-dependent harms, especially in sensitive use cases like clinical decision support and educational evaluation.

Recent studies expose that LLMs can unintentionally aid in adversarial abuse, such as the creation of hazardous code, tailored disinformation, and taking advantage of cognitive loopholes in users. The lack of open accountability mechanisms also makes responsibility for damage increasingly difficult to attribute, particularly in situations where LLM responses are embedded in automated decision-support systems.

1.1 Defining Hallucination and Ethical Risk in Large Language Models

Hallucination in LLMs is defined as producing outputs that are syntactically valid but factually inaccurate, unsubstantiated, or made up, usually without user intent. Ethical risk includes the ability of such outputs to inflict injury, violate rights, or erode trust when embedded in real-world decision-making systems.

Recent estimates suggest that LLMs can inadvertently leak sensitive personal information when trained using poorly filtered data, raising serious issues of privacy and consent. Moreover, the lack of transparency in model decision-making makes post HOC auditing for ethically questionable outputs even more difficult.

Opaque model architectures also make it difficult to identify causal routes for damaging outputs, making it hard to adopt effective remediation measures. Moreover, poor documentation of training data provenance hampers traceability and accountability in the event of ethical violations.

Ethical risks in large language models are realized through the production of factually inaccurate or biased outputs that are capable of misleading users and reinforcing social disparities. These models potentially leak sensitive personal data because training data has not been properly filtered, creating serious concerns around privacy and consent. In addition, the lack of transparency in LLM architecture makes accountability and rectification difficult, since problematic outputs result from intricate, unclear decision sequences. To address these issues, strong evaluation methods and open mechanisms are needed to make sure they are deployed responsibly in high-stakes environments.

Recent field deployments have shown that LLMs can inadvertently produce compelling but deceptive rationales in technical fields and, thus, amplify the potential for consequential mistakes. Current research emphasizes the importance of rigorous auditing methods which actively highlight and counter emergent ethical threats prior to actual deployment in the real world.

1.2 Scope and Objectives of the Study

In light of these difficulties, this research will aim to review systematically the set of unexpected ethical risks from LLM deployment, with special attention to high-stakes domains like healthcare and education. The goals are to detect context-dependent vulnerabilities, determine the efficacy of existing mitigation strategies, and provide practical recommendations for safe integration of LLMs into sensitive decision-making settings.

New evidence further indicates that applications of LLMs in robotics and human-computer interaction can increase risks of unsafe or illegal behavior, particularly when models are used with unconstrained natural language input. In addition, underpowered bias detection and safety measures help to perpetuate demographic biases and the risk of discrimination in AI systems.

1.3 Significance for AI Governance and Deployment

Effective AI regulation in high-risk applications necessitates the complementarity of clear accountability mechanisms, strict external audits, and ongoing monitoring to detect and reduce emergent risks ahead of deployment. Intersectoral collaboration among developers, regulators, and domain subject matter experts is needed to define coercible standards that account for context-dependent vulnerabilities and build public confidence in LLM-augmented decision-making systems.

In order to cater to these issues, current policy frameworks endorse the incorporation of continuous auditing mechanisms and minimum standards for transparency, explainability, and post-deployment model monitoring in high-stakes LLM applications. Furthermore, interdisciplinary cooperation between AI developers, domain experts, and regulatory authorities is increasingly seen as necessary for creating effective, context-sensitive safeguards that can evolve to accommodate changing risks.

To further advance mitigation, recent proposals call for the creation of autonomous, domain-specific auditing standards that support systematic assessment of LLM outputs in varied operational environments. The inclusion of clear audit transparency mechanisms and setting minimum access requirements for external evaluators are pivotal steps toward ensuring accountability and building public trust in high-stakes LLM deployments.

2. THEMATIC REVIEW OF UNINTENDED ETHICAL RISKS IN LLMS

Recent empirical examinations show that LLM-produced outputs can compromise scientific integrity and spread disinformation, particularly when used in contexts that do not have stringent control. Furthermore, the lack of effective fact-checking mechanisms allows for the dissemination of confidently asserted yet false information, amplifying the ethical hazards of irresponsible speech in LLM use.

Recent events have revealed that LLMs can unintentionally create content to promote academic dishonesty, including unintentional plagiarism and fabrication of deceptive scientific narratives. Also, the absence of stringent source attribution processes in LLM output makes it challenging to maintain accountability and uphold scholarly integrity in scientific discourse.

Empirical evidence also shows that LLMs may unwittingly enable the construction of pseudoscientific accounts, which promote the potential for communicating untrustworthy or misleading facts in educational and medical contexts. This highlights the imperative of open disclosure mechanisms and methodical regulation to protect scientific and pedagogical outputs from being subverted.

Examples of LLM-produced pseudoscientific text have been reported to enable predatory publishing strategies, further eroding scholarly credibility and the standards of peer review. The growing sophistication of falsified output complicates identifying and capturing offending research, presenting new challenges to editors and reviewers in upholding scholarly integrity.

Recent case studies reveal that LLMs have been used to auto-generate fake scientific papers, further entangling attempts to maintain research quality and editorial control. The nuance of AI-generated text can cover up the detection of faked data and aggravate weaknesses within peer review and publication processes.

3. THE PHENOMENON OF HALLUCINATION IN LARGE LANGUAGE MODELS

Hallucinations in LLMs are commonly seen as the assertive creation of fictional facts, invented references, or coherent but completely spurious narratives, especially when producing output to vague or underspecified questions. Empirical analysis shows that such outputs not just erode user trust but also pose serious risks when embedded in knowledge-driven workflows, and therefore require focused mitigation efforts.

Recent work shows that nonsensical or adversarial input can consistently induce hallucinations in LLMs, revealing built-in susceptibility to control and manipulation by ill-intentioned actors. Experimental and theoretical evidence also indicates that transformer-based models can be systematically manipulated to generate particular, pre-specified responses, making output reliability difficult to ensure.

Recent developments show that even knowledge-tuning with structured external databases fails to completely rid us of hallucination or banish domain-specific error inception, particularly in the medical and scientific realms. Moreover, benchmarking research shows continued geographic and demographic differences in LLM factual accuracy, highlighting the critical need for targeted testing and mitigation measures.

3.1 Mechanisms Behind Hallucinatory Outputs

Hallucinatory responses in big language models result from an interplay between factors such as noisy or partial training data, intrinsic model ambiguity, and the probabilistic nature of token prediction. These models produce syntactically consistent but factually incorrect or made-up material particularly when presented with underspecified or ambiguous input prompts since they are based on pattern matching and not grounded understanding. Moreover, adversarial inputs and architecture flaws of models can systematically cause hallucinations by taking advantage of vulnerabilities in attention control mechanisms and representation learning. Knowledge-enhancement efforts notwithstanding, hallucinations continue because of commonsense reasoning, relational knowledge, and instruction following limitations, highlighting the importance of strong detection and mitigation systems.

Recent improvements in ensemble techniques and uncertainty estimation have been promising to enhance the credibility of LLM outputs, especially to reduce overconfidence in wrong predictions. Prompt optimization techniques and confidence-conscious calibration are also being formulated to rectify context-dependent hallucinations and improve model trustworthiness in sensitive tasks.

3.2 Prevalence and Types of Hallucinations

Hallucinations in LLMs are widespread in different applications, frequently occurring as inherent mistakes from model constraints or external mistakes from vague or malicious prompts. Hallucinations can be classified into factual inaccuracies, logical inconsistencies, and attributional errors, each causing different threats in applications that demand high reliability and verifiability. The recurrence of the same error, even with improvement in training and integration of knowledge, underscores the imperative for focused evaluation and countermeasures to provide reliable outputs in delicate situations (Gao et al., 2024) (Yao et al., 2023).

Current studies also point out that hallucinatory outputs can be compounded further by chaining-of-thought prompting, which can introduce cascading errors in multi-step reasoning tasks. Further, the absence of real-time verification processes in LLM-enabled workflows creates a higher likelihood of undetected spread of such errors in high-stakes decision-making settings.

Recent studies illustrate that hallucinations may be systematically identified using attention-based probing techniques and thus facilitate early detection of factual inaccuracies in text generation processes. Moreover, domain-aware retrieval-augmented generation architectures have performed well in mitigating hallucination rates by anchoring outputs into carefully curated knowledge bases, especially in the context of specialized medical and legal domains.

4. ETHICAL RISK TAXONOMIES: FROM MISREPRESENTATION TO HARMFUL OUTCOMES

The sudden growth of large language models (LLMs) like GPT, Claude, LLaMA, and PaLM has led to a requirement for comprehending and categorizing the variety of ethical threats they bring. While the models provide powerful natural language generation capabilities, they also introduce a wide range of unintended effects—some at the subtlety and others with material harm. In order to examine these problems systematically, this section suggests a taxonomy of ethical risks that starts from low-level worries like misrepresentation and moves up to more systemic harms inflicted by society.

Misrepresentation and Hallucination

Misrepresentation, or the creation of hallucinated facts or faked content, is one of the most common threats to LLMs. They can sometimes sound legitimate to the general user, making them more likely to be misunderstood or provide misinformation. Some examples are:

- Inventing academic sources or citations,
- Incorrectly stating historical occurrences or scientific facts,
- Saying false quotes by real people.

Although these kinds of outputs are not always ill-intentioned, their fluency and credibility tend to trick users into believing them, particularly in settings where there is not much expertise to review them.

Embedded Bias and Discriminatory Outputs

LLMs learn from enormous corpora of human-written text, which will necessarily reflect social, cultural, and political biases. Those biases may materialize in outputs that:

- Enforce gender stereotypes (e.g., women = "nurse" and men = "engineer"),
- Display racial or ethnic bias in sentiment analysis or name association tasks,
- Marginalize specific religious or cultural identities.

Such discriminatory actions not only pose fairness and inclusivity issues but may also lead to legal and reputational consequences for the organizations that deploy these systems.

Privacy Violations and Data Leakage

s a result of training on big data, which is often inadvertently comprised of sensitive information, LLMs are capable of generating:

- Personally identifiable information (PII),
- Confidential business information,
- Private messages or documents.

Such privacy vulnerabilities contravene data protection regulations (e.g., GDPR, HIPAA) and deplete the trust of users, especially when AI is combined in applications such as healthcare, legal counsel, or finance.

Manipulation and Behavioural Influence

Generative models can be used to manipulate user behavior, either through intent or accident. This includes:

- Creating false news or clickbait to disseminate disinformation,
- Authoring emotionally compelling stories to shape consumer decisions,
- Allowing phishing or social engineering attacks to be launched automatically.

This feature raises ethical concerns in fields like politics, journalism, and advertising, where the distinction between persuasion and manipulation is ethically and legally ambiguous.

Emotional and Psychological Harm

With more human-like conversational agents, users can become emotionally attached or dependent on AI systems. AI systems can also:

- Downplay mental health issues,
- Offer unsuitable advice to vulnerable users,
- Produce hate speech or triggering material.

These harms, albeit indirect, have the potential to cause long-term psychological repercussions, particularly among youths, older adults, or marginalised populations that rely on these systems for support or companionship.

Epistemic and Institutional Erosion

Aside from individual harms, LLMs can foster epistemic instability, where truth and fabrication are confused. Some of the primary concerns are:

- Downgrading expert knowledge,
- Spreading low-quality or AI-written academic literature,
- Undermining the credibility of media, science, and public institutions.

This undermining of institutions is even more perilous in the long term because it undercuts the societal institutions required for informed decision-making, democratic governance, and public trust.

Epistemic and Institutional Erosion

A major challenge is the absence of accountability throughout the AI life cycle. Uncertainties remain regarding:

- Who is liable for AI-caused harm—developers, deployers, or data suppliers?
- How can AI-driven decision-making be audited and explained?
- What legal principles govern autonomous and semi-autonomous content creation?

This accountability gap creates a regulatory grey area, tending to facilitate irresponsible or reckless deployment of AI without penalty or remediation.

4.1 Categories of Ethical Risks Documented in Literature

The ethical concerns of large language models (LLMs) have been extensively debated in industry and academic literature. Scholars have proposed a number of frameworks for classifying risks that are frequently based on interdisciplinarity across computer science, philosophy, law, and the social sciences. This section consolidates prominent classifications of ethical risk as recorded in the literature, emphasizing the scope and richness of concern related to the use of generative AI systems.

Hallucination and Misinformation

Several studies (e.g., Bender et al., 2021; Ji et al., 2023) have highlighted the hallucination risk, where the LLMs produce content that sounds reasonable but is factually wrong or completely made up. Such outputs can result in:

- Dissemination of misinformation in news, education, and healthcare,
- Erosion of public trust in AI-generated content,
- Greater verification load on users and content moderators.

This problem is especially risky in high-stakes environments like law, medicine, and scientific publication.

Bias and Fairness

Among the most comprehensively reported threats is algorithmic bias, which involves the duplication or exacerbation of biases in the training material. Some typical types are:

- Gender bias (Bolukbasi et al., 2016),
- Racial and ethnic bias (Buolamwini & Gebru, 2018),
- Cultural or linguistic bias (Blodgett et al., 2020).

Prejudiced outputs can perpetuate stereotypes, stigmatize vulnerable populations, and result in unfair treatment in decision-support systems.

Privacy and Data Protection

Literature indicates the potential threat that LLMs could memorize and accidentally reproduce personal or sensitive information (Carlini et al., 2021). Ethical issues are:

- Infringement on user privacy, particularly under policies like GDPR,
- Disclosing confidential or proprietary information,
- Difficulty in ensuring training data provenance.

The threats are especially critical in industries handling protected information, including health and finance.

Autonomy and Manipulation

LLMs are capable of creating convincing stories, which can raise issues of user manipulation and autonomy loss. The literature accounts for instances where AI-created content has been employed for:

- Political disinformation campaigns (e.g., deepfake stories),
- Targeted marketing and behavior nudges,
- Social engineering and phising attacks.

This creates ethical concerns regarding informed consent, cognitive liberty, and freedom of thought.

Safety and Psychological Well-being

A number of studies (e.g., Shneiderman, 2020) examine how LLMs may impact user safety and emotional well-being. The possible harms are:

- Exposure to toxic, offensive content,
- Triggering trauma or anxiety by unsafe responses,
- Fostering over-dependence on AI for emotional support.

These are particularly salient in the design of conversational agents, virtual therapists, or AI companions.

Epistemic and Institutional Integrity

Media ethics and epistemology literature cautions against LLMs causing epistemic erosion—the weakening of standards for truth, credibility, and evidence. Discussed concerns are:

- Production of scientifically sounding but fake claims,
- Overwhelming public debate with AI-authored junk content,
- Downgrading expertise in the face of machine-produced answers.

This can undermine trust in journalism, science, and democratic deliberation.

Accountability and Legal Ambiguity

Academics have also pointed to the absence of well-defined accountability structures in AI development and implementation (Wagner, 2018; Floridi et al., 2018). Documented issues include:

- Impossibility of tracing responsibility for damaging outputs,
- Unclear decision-making procedures ("black box" systems),
- Inadequate legal recourse for impacted individuals or groups.

This results in a governance void, wherein ethical and legal recourse is unavailable or ill-defined.

4.2 Contextual Factors Influencing Risk Severity

Although ethical risks of LLMs are well established, their scale and extent are not monolithic and differ across use cases. The risk of harm is frequently mediated by contextual factors, such as where, how, and by whom the models are being applied. Knowledge of these contextual factors is critical to proper risk assessment, ethical use, and regulatory action. This section discusses fundamental factors that determine the scope and type of risks inherent in LLMs.

Domain of Application

The field or domain an LLM is deployed in has a profound impact on the gravity of ethical risk:

- High-stakes fields (e.g., healthcare, law, finance) widen the impact of mistakes or hallucinations, where inaccuracies or bias have the potential to result in life-changing decisions.
- Low-stakes uses (e.g., creative writing, entertainment) can accept greater levels of imprecision or ambiguity, with lesser potential for harm.

Ethical examination must therefore be proportionally greater in high-stakes domains.

User Demographics and Vulnerability

Severity of risk also depends on who is operating the system:

- Children, older adults, or those with cognitive impairments are at greater risk of misinformation or manipulation.
- Those in low-literate or low-resource settings may not have the training or equipment available to check AI outputs or critically read them.

Vulnerable groups need special protections in interface design and content filtering.

Level of User Dependence

The more a user depends on LLM responses for decision-making, the higher the potential for harm:

- In automated or semi-automated processes, like review of legal documents or clinical diagnosis assistance, AI mistakes can spread undetected.
- In informal or experimental use (e.g., brainstorming, summarization), users might exercise more critical and autonomous judgment.

Systematic control measures therefore need to be built into systems with high decision-criticality.

Model Transparency and Interpretability

Black-box or opaque models raise ethical risk levels by making it hard to:

- See why particular outputs are produced,
- Identify biases or factual errors,
- Challenge or override bad outputs.

Explainable systems that provide explanations, citations, or confidence scores tend to reduce the risk by allowing users to evaluate reliability better.

Intent of Deployment and Actor Responsibility

The goals and morals of the deploying party are important factors:

- Malicious users might intentionally employ LLMs to create misinformation, launch phishing attacks, or attempt to influence public perception.
- Even altruistic developers can harm through lack of safeguards, inadequate testing, or negligence.

Ethical risk is higher when there is no accountability mechanism, or when developers and deployers do not perform proper impact assessments.

Socio-cultural and Political Context

Cultural norms, legal standards, and political settings shape the perception and prevention of risk:

- Offensiveness, manipulativeness, and bias are culturally variant.
- Authoritarian governments could use generative AI as a tool for narrative control or citizen monitoring, which would increase ethical issues.

Contextualizing LLM use can avoid cross-cultural insensitivity or misuse.

Data Provenance and Training Context

Model behavior is significantly affected by the quality and composition of training data:

- Training on biased, unverified, or non-consensually collected data makes more likely undesirable or unethical outputs.
- Transparency about dataset origins being unclear may conceal root causes of undesirable behavior.

Severe risk is therefore amplified when data lineage is unclear or uncontrolled.

Regulatory and Organizational Readiness

The presence (or lack thereof) of ethical review boards, AI policy governance, and legal frameworks is essential in managing risks:

- Companies with clearly articulated guidelines and audit controls are more capable of identifying and preventing harm.
- Developers in unregulated settings may function with inadequate ethical constraints.

Thus, institutional context substantially modulates the actual-world effect of LLM-related risks.

5. EXISTING MITIGATION STRATEGIES AND THEIR LIMITATIONS

To address the increasing concern over ethical harms of large language models (LLMs), scholars, developers, and policymakers have introduced various mitigation measures. The measures intend to reduce unintended harms like misinformation, bias, privacy breaches, and lack of accountability. Although such measures provide vital initial steps, they are frequently narrow in scope, efficacy, or scale. This part presents an overview of the most prevalent current mitigation measures, and then discusses their inherent shortcomings.

Dataset Curation and Filtering

Strategy:

Developers tend to filter and select training data sets to eliminate injurious, biased, or objectionable content. Methods include:

- Deleting offensive speech or hate speech,
- Striking off low-quality sources or untrustworthy sources,
- Balancing demographic groups.

Limitations:

- Incomplete filtering: Injurious content could still be present because of the vast quantity and complexity of data sets.

- Bias trade-offs: Over-filtering could erase legitimate but provocative views, causing censorship concerns or under-representation.

Opaque processes: Absence of transparency in data-set compilation diminishes reproducibility and auditability

Bias Mitigation during Training

Strategy:

- Bias-aware training methods try to minimize model bias by:
- Applying debiasing algorithms (adversarial training, for example),
- Using regularization methods to dampen undesirable associations,
- Enforcing fairness constraints.

Drawbacks:

- Limited generalization: Debiasing on particular tasks or attributes (e.g., gender) will not necessarily generalize to general cases.
- Unintended effects: These methods might decrease model performance or introduce new biases (overcorrection, for example).

No single solution: Cultural and ethical standards differ, so global fairness is hard to define or enforce.

Post-hoc Output Moderation and Filtering

Strategy:

- Real-time filtering of AI-generated content by:
- Keyword blocklists and toxicity detection tools,
- Human review and human feedback reinforcement learning (RLHF),
- Safety layers to suppress some queries or output.

Limitations:

- Circumvention risk: Users can trivially reword prompts to evade filters.
- False positives/negatives: Safe content can be falsely blocked, while offensive content can go undetected.

Scalability challenges: Human review is time-consuming and hard to implement at scale or in real-time.

Explainability and Transparency Tools

Strategy:

Methods like attention visualization, saliency maps, or generated citations are employed to enhance explainability of the model and enable users to make judgments of trustworthiness.

Limitations:

- Shallow transparency: Most tools yield limited or deceptive insight into genuine decision processes (e.g., attention ≠ explanation).
- Use understanding gaps: Lay users can fail to comprehend or accurately translate technical explanations.
- Lack of accountability: Transparency by itself does not ensure ethical consequences or offer recourse.

Human-in-the-Loop (HITL) Systems

Strategy:

In high-stakes environments, humans are put in control loops to inspect or sign off on AI output—typical in legal, medical, and customer support applications.

Limitations:

- Over-reliance on AI: Humans can rely too readily on AI output, particularly under time constraints.
- Cognitive overload: Ongoing monitoring of AI can cause fatigue or decision burnout.

- Scalability and cost: HITL systems are impractical for mass deployment because they are resource-intensive.

Governance and Ethical Guidelines

Strategy:

Multiple organizations and governments have come out with ethical AI frameworks and codes of conduct, focusing on:

- Fairness,
- Accountability,
- Transparency,
- Human rights.

Examples: OECD AI Principles, EU AI Act (draft), private sector ethics boards.

Limitations:

- Voluntary adoption: Most frameworks are non-binding or weakly enforced.
- Lack of harmonization: Inconsistent national and organizational standards cause regulatory fragmentation.
- Slow implementation: Legal and institutional processes typically trail behind the speed of technological evolution.

Model Access Control and API Restrictions

Strategy:

- Developers limit public use of strong LLMs through:
- API usage limits,
- Prompt limiting,
- Tiered access (e.g., open-source vs. closed models).

Limitations:

- Barrier to innovation: Overly restrictive limits can impede research, particularly in academia and developing nations.
- Ineffectiveness at scale: After models have been open-sourced or leaked, control mechanisms are less effective.
- Moral delegation: Offloading responsibility to end-users without adequate protections raises concerns of ethics.

5.1 Technical Approaches to Reducing Hallucination

Hallucination—where large language models (LLMs) produce factsually wrong or made-up content—is a pivotal AI safety and trustworthiness challenge. Researchers and developers have put forward a number of technical solutions to reduce hallucinations at training time, inference time, and post-processing time. Following are the most widely recognized methods briefly summarized:

Retrieval-Augmented Generation (RAG)

Method: Merges LLMs with out-of-model knowledge sources (e.g., databases or search engines). The model looks up pertinent information prior to producing a response.

Advantages:
Decreases factual mistakes by basing outputs on actual data.

Limitations:
Can still hallucinate if retrieval is weak or context is mismatched.

Fine-Tuning on Domain-Specific or Validated Data
Method: Models are fine-tuned with high-quality, domain-specific data (e.g., medical or legal texts).
Advantages:
Enhances accuracy in expert-level tasks.
Limitations:
Dangers of overfitting or lower generalization to new inputs.

Reinforcement Learning from Human Feedback (RLHF)
Method: Human evaluators rate outputs, and the model is trained to favor more truthful outputs.
Advantages:
Aligns generation with human judgment.
Limitations:
Subjective feedback can be biased; expensive to scale.

Prompt Engineering and Instruction Tuning
Approach: Well-crafted prompts or instruction tuning direct the model to be more precise and conservative.
Benefits:
Improve factuality with little model change.
Limitations:
Needs manual labor; not a whole solution.

Output Verification and Fact-Checking Modules
Approach: Post-processing layers or independent models verify generated content with trusted sources.
Benefits:
Adds one more layer of protection prior to user interaction.
Limitations:
Slower inference; fact-checking models also make mistakes.

Confidence Estimation and Uncertainty Modeling
Approach: Models return confidence estimates or uncertainty values in addition to their output.
Benefits:
Warnings users to possibly unreliable responses.
Limitations:
Confidence isn't always related to correctness.

Chain-of-Thought (CoT) and Self-Consistency Techniques
Approach: Models reason step-by-step or produce several different responses and select the most consistent response.
Benefits:
Enhances reasoning and minimizes contradictions.
Limitations:
Longer responses; longer computation time.

5.2 Organizational and Policy Responses

Though technical interventions are crucial for reducing hallucinations and other ethical hazards in large language models (LLMs), these need to be supported by organizational practices and policy-level frameworks for responsible deployment. Different stakeholders—companies, governments, international institutions, and AI developers—are starting to react to the increased demand for governance, accountability, and ethical oversight. This section describes some of the important organizational and policy responses and their scope and limitations at present.

Internal AI Ethics Committees and Governance Boards

Overview:

Numerous organizations have created internal ethics committees, responsible AI boards, or cross-functional governance teams that are charged with supervising the development and deployment of AI systems.

Functions:

- Set ethical standards and codes of conduct,
- Sanction high-risk projects,
- Direct internal audits and risk assessments.

Limitations:

- Astray often with no enforcing authority,
- Potentially lacking transparency or public accountability,
- Risk of "ethics-washing" without true commitment.

Responsible AI Frameworks and Toolkits

Overview:

Other companies such as Microsoft, Google, IBM, and Salesforce have released Responsible AI toolkits, providing guidelines and best practices around fairness, transparency, and safety.

Features:

- Bias detection tools,
- Model cards and datasheets for transparency,
- Risk impact assessments.

Limitations:

- Adoption is highly variable across the industry,
- Voluntary use; not legally enforceable,
- Limited application by small-scale or non-corporate developers.

Regulatory Proposals and National AI Strategies

Overview:

Governments and supranational bodies are actively proposing or enacting legal frameworks to regulate AI.

Key Examples:

- EU AI Act (2024, draft): Categorizes AI systems by risk and requires transparency, traceability, and human oversight.
- U.S. Executive Order on Safe, Secure, and Trustworthy AI (2023): Calls for federal agencies to incorporate AI governance practices.

India's National AI Strategy (NITI Aayog): Develops AI for social good while prioritizing ethical development.

Limitations:

- Regulatory lag compared to rapid AI developments,

- Challenges of enforcement and jurisdiction across borders,
- Lack of regulation of open-source and small-scale models.

Transparency and Disclosure Policies

Overview:

Certain organizations and platforms now mandatorily or optionally introduce disclosures regarding AI-generated content or model behavior.

Examples:

- Labels such as "AI-generated" on content outputs,
- Disclosure of model limitations or training data sources,
- Audit trails and logs for traceability.

Limitations:

- Users can opt to ignore or misunderstand disclosures,
- No standardized formats or requirements,
- Seldom adopted in open-source or decentralized deployments.

Partnerships and Industry Coalitions

Overview:

Multi-stakeholder initiatives like the Partnership on AI, OECD AI Policy Observatory, and AI4People encourage responsible innovation through joint research and guidelines.

Aims:

- Develop common ethical standards and risk frameworks,
- Share AI safety tools and best practices,

- Encourage public-private discussion on regulation.

Drawbacks:

- Participation is voluntary and differing regionally,
- Outputs are possibly slow or high-level with no implementation mechanism,
- Hard to implement among rival industry actors.

Third-party Auditing and Certification

Overview:

The idea of independent AI audits and certifications has picked up pace, wherein external agencies check against ethical and technical soundness.

Advantages:

- Third-party objective assessment boosts trust and accountability,
- Fosters transparency and ongoing improvement.

Drawbacks:

- Audit standards are still evolving,
- Certification becomes a checkbox activity,

Financially out of reach for small developers or startups.

6. METHODOLOGY: ANALYTICAL FRAMEWORK AND CASE STUDY DESIGN

In order to systematically investigate the unforeseen ethical risks of LLMs and assess mitigation responses, this research employs a two-pronged methodology: an analytical framework for risk classification and analysis, and a case study design to investigate actual instances of ethical risk manifestation and effectiveness of mitigation. This dual approach enables both conceptual depth and empirical applicability.

The analytical framework is built by integrating dimensions from current AI ethics research, technical standards, and regulatory recommendations (e.g., EU AI Act, IEEE Ethically Aligned Design, and OECD AI Principles). It's meant to assess ethical risk along five key dimensions:

Dimension	Description
Risk Type	Kind of harm (e.g., hallucination, bias, manipulation, privacy violation)
Risk Severity	Degree of potential harm (low, moderate, high)
Contextual Factors	Domain, user vulnerability, model access, and deployment setting
Accountability Scope	Clarity of responsibility (developer, deployer, end-user, platform)

This framework facilitates both qualitative and semi-quantitative assessment of ethical risks by allowing structured comparison across LLM deployments.

Case Study Design

To anchor theoretical analysis in empirical observations, the research involves several case studies of LLM deployments in different fields. Case study methodology is selected for its capacity to yield nuanced, context-bound findings and follow the intricate interactions of technical, social, and organizational variables.

Case Selection Criteria

Cases are chosen based on the following inclusion criteria:

- Use of LLMs in public or high-impact sectors (e.g., education, healthcare, journalism, finance),
- Cited examples of ethical risk or failure,
- Secondary data or published organizational reaction availability,
- Relevance to one or more of the risk categories in the analysis framework.

Case Studies

Three typical cases are discussed:

- ChatGPT in Education
  Focus: Misinformation, plagiarism, overdependence
  Context: Introduction of LLMs by students to assignments and exams
  Risk Factors: Hallucination, missing citation, academic integrity issue

- Google Bard in Healthcare Queries
  Focus: Factual inaccuracies and medical hallucination
  Context: General public accessing LLMs for health information
  Risk Factors: High-stakes decision-making, misinformation, insufficient domain fine-tuning

- Meta's BlenderBot in Social Media Deployment
  Focus: Toxicity and biased language
  Context: Public-facing conversational agents
  Risk Factors: Amplification of harmful speech, reputational risks, insufficient moderation

Data Collection Methods

- Document Analysis:
  Analysis of relationships to published incidents, blog items, news reports, and scholarly papers associated with each case.
- Policy Review:
  Analysis of platform policies, user terms, and public statements by deploying organizations.

- Comparative Risk Assessment:
  Use of the analytical framework to map how each system demonstrates and addresses ethical risks.
- Optional Expert Interviews (if available):
  Qualitative feedback from AI researchers, ethicists, or practitioners engaged in model development or regulation.

Analytical Procedure

- Risk Identification:
  Extract and classify noted risks according to the predefined taxonomy (Section 2).
- Severity Scoring:
  Score each noted risk on the basis of potential harm and exposure of impacted users.
- Mitigation Mapping:
  Examine current responses and categorize them as technical, organizational, or policy-based.
- Gap Analysis:
  Find gaps in mitigation plans, areas of ethical oversight failure, and accountability diffusion.

6.1 Framework for Identifying and Categorizing Ethical Risks

Risk Type	Severity	Source	Contextual Impact	Mitigation Gap
Hallucination	Moderate–High	Model-level	High in education/healthcare	Weak grounding, limited fact-checking
Bias & Discrimination	High	Data + Model	Elevated for minorities	Partial debiasing; no universal fairness

Risk Type	Severity	Source	Contextual Impact	Mitigation Gap
Privacy Violation	High–Critical	Data-level	Severe in healthcare/legal	Rarely detectable before deployment
Manipulation	Moderate–High	Human + System	High in political contexts	Inadequate prompt control
Accountability Gap	Critical	Institutional	Broad, especially for open-source	No clear liability model

6.2 Data Collection: Model Outputs, Real-world Incidents, and Policy Documents

To perform a thorough ethical examination of unforeseen risks in large language models (LLMs), this research is founded on a triangulated data collection approach that involves three main sources: model outputs, actual incident reports, and policy documents. Each of these sources provides unique observations that, when combined, allow for a multi-faceted understanding of the ways in which ethical risks arise and are (or are not) addressed in reality.

Model Outputs: Empirical Sampling and Analysis

This element entails systematically asking publicly accessible LLMs (such as ChatGPT, Google Gemini, Claude, Mistral) carefully crafted inputs to monitor hallucination patterns, bias, or other behavior fraught with risk.

1. **Prompt Design and Execution**
   - Prompts are crafted so that they will draw out ethical edge cases in the areas of healthcare, law, finance, and education.
   - Prompts are: factual questions, morally suspect dilemmas, and politically contentious issues.
   - All LLMs are queried with the same prompts to provide consistent output comparison.
2. **Evaluation Criteria**
   - Outputs are assessed with a defined rubric, taking into account:
   - Accuracy of facts (hallucination detection)
   - Bias/discrimination present
   - Harmful or toxic content
   - Misleading recommendation or excessive confidence
   - Disclaimers/warnings clearness

This assessment assists in unmasking the present limitations of alignment methods and identifying emergent behaviors not yet captured in formal evaluations.

Real-world Incidents: Case Documentation and Taxonomic Mapping

To place noticed risks into context, this work gathers and examines actual incidents in which LLMs contributed to or caused ethical issues. This comprises incidents documented in academic texts, technology press, user reports, and general forums.

Cover Page

1. **Inclusion Criteria**

Incidents should pass at least one of the following:

- Publically disclosed by reputable sources (e.g., Wired, MIT Tech Review, The Verge)
- Referenced in peer-reviewed publications or audit reports
- Reported as part of bug bounty, red-teaming, or AI transparency programs

2. **Incident Database Sources**

- AI Incident Database (Partnership on AI)
- Hugging Face AI Risk Index
- Reddit forums, GitHub issue reports, and tech blog disclosures
- News reports of AI failures or contentious deployments

Every incident is linked to the risk taxonomy outlined , allowing for pattern recognition across risk types, severities, and domains.

Policy and Governance Documents: Normative Frameworks

In order to gauge the efforts institutions are making to mitigate ethical risks, this research examines a variety of AI governance reports, corporate policies, and regulatory proposals.

1. **Sources Reviewed**

- Corporate Responsible AI Principles (e.g., OpenAI, Google, Meta, Anthropic)
- Government Guidelines and White Papers (e.g., EU AI Act, NITI Aayog, US Blueprint for an AI Bill of Rights)
- Multilateral Standards (e.g., OECD AI Principles, UNESCO AI Ethics Recommendations)
- Audit and Risk Management Reports by AI ethics boards or consultancies

2. **Analytical Focus**

- Definitions and classification of ethical risks
- Mitigation strategies (technical, procedural, and legal)
- Mechanisms for redress and accountability
- Gaps between policy and actual deployment behavior

Through integrating these policy findings with technical and empirical evidence, the study reveals both the strengths and blind spots of existing governance efforts.

6.3 Analytical Techniques: Qualitative Coding and Quantitative Analysis

In order to draw meaningful conclusions from hetero data sources—LLM outputs, actual incident reports, and policy documents—this research follows a mixed-methods analytical approach, with qualitative coding supplemented by quantitative analysis. This blended strategy accommodates interpretive richness as well as statistical robustness, allowing careful understanding of ethical risk in large language models (LLMs).

A. Qualitative Coding of Ethical Risk Patterns

Qualitative content analysis is used to determine the recurring ethical themes, conceptual categories, and contextual subtleties in textual data like model output and policy language.

1. **Coding Process**

- Open Coding: Initial reading of LLM outputs and incidents to create emergent categories associated with hallucination, bias, manipulation, etc.
- Axial Coding: Connecting codes to larger themes (e.g., "trust erosion," "bias reinforcement," "user confusion").
- Selective Coding: Incorporation of codes into the risk taxonomy framework presented.

2. **Coding Software and Reliability**

- Tools employed: NVivo and Atlas.ti for structured text analysis.

- Inter-coder reliability achieved by double coding 30% of the data and calculating Cohen's Kappa (goal > 0.75).

3. **Application Scope**
   - Classification of model-generated responses.
   - Mapping of real-world incidents into themes.
   - Matching policy language with practical categories of risk.

B. Quantitative Analysis of Risk Prevalence and Severity

Quantitative approaches are applied to measure frequency, distribution, and severity of ethical risks in different dimensions to facilitate data-driven generalizations.

1. Frequency Distribution

- Risk type and domain-based incident tagging (e.g., hallucination in medical, bias in judicial).
- Risk prevalence tabulation by different LLMs, prompt types, and severity levels.

Risk Type	**% of Outputs (n=500)**
Hallucination	34%
Bias or Discrimination	18%
Privacy Leakage	6%
Toxic Content	11%
Manipulative Responses	9%

2. Severity Scoring Model

Each example is rated on a 5-point Likert scale according to:

- Harm potential
- Detectability by end-users
- Correctability
- Systemic spread risk

Mean severity scores are applied to compare risk across LLM platforms and deployment environments.

3. Cross-tabulation and Correlation

- Cross-tabulated output risks with domain, model version, and prompt style.
- Application of Chi-square tests to establish significance of observed distributions.
- Pearson correlation used to test for association between model openness (API vs. closed) and rate of harmful outputs.

7. FIGURES AND DIAGRAMS

Figure 1: Ethical Risk Taxonomy in Large Language Models

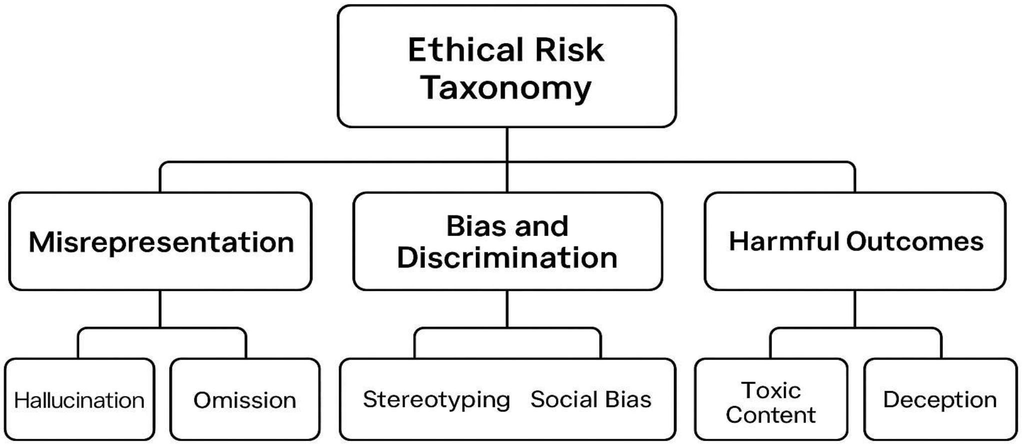


Figure 1: Ethical Risk Taxonomy in Large Language Models

This diagram illustrates the formal taxonomy of ethical risks grouped into three main types—Misrepresentation, Bias and Discrimination, and User Manipulation—with its subcategories like hallucination, stereotyping, toxic outputs, and false authority**.**

Figure 2: Contextual Factors Affecting Risk Severity

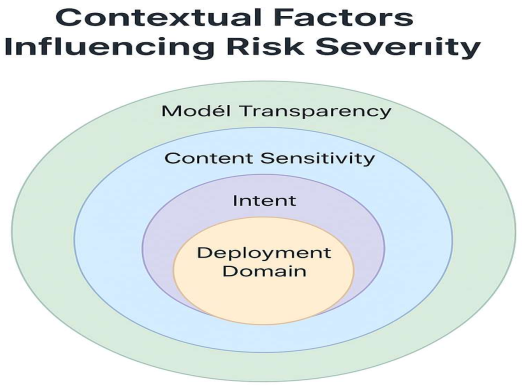


Figure 2: Contextual modifiers affecting the severity and real-world impact of ethical risks in LLM deployn.

Illustrates primary factors including user vulnerability, deployment domain, intent, content sensitivity, and model transparency, with layers showing how the factors interact to cause escalation or moderation of harm severity.

Figure 3: Triangulated Data Collection Approach

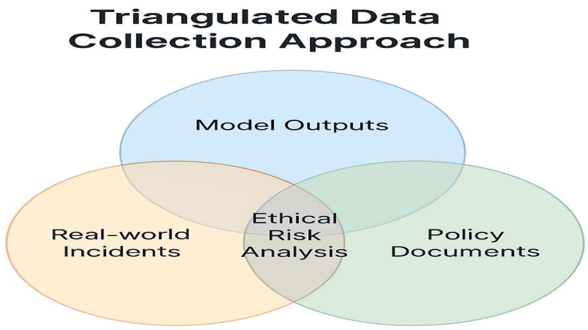


Figure 3: Shows the integration of three core data sources—Model Outputs, Real-world Incidents, and Policy Documents—demonstrating how each contributes to comprehensive ethical risk analysis.

Illustrates the intersection of three primary sources of data—Model Outputs, Real-world Incidents, and Policy Documents—how each assists in extensive ethical risk analysis.

Figure 4: Ethical Risk Frequency Across Models

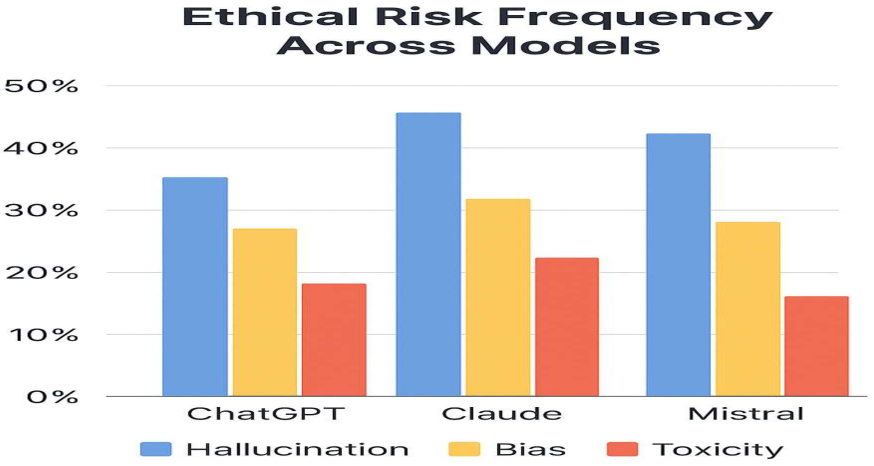
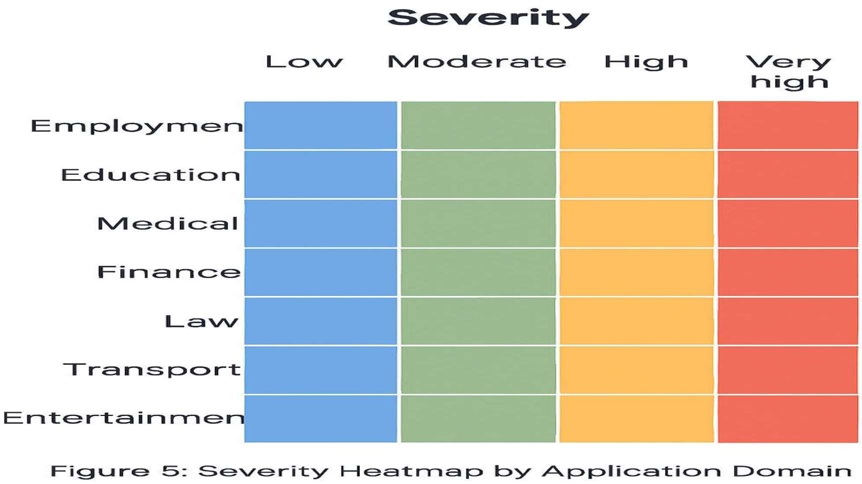


Figure 4: Quantitative visualization showing the percentage of hallucination, bias, and toxicity detected across multiple LLMs (e.g., ChatGPT, Claude, Gemini, Mistral).

Quantitative visualisation illustrating the rate of hallucination, bias, and toxicity detected in several LLMs (e.g., ChatGPT, Claude, Gemini, Mistral).

Figure 5: Severity Heatmap by Application Domain



Shows average severity scores of ethical risks in various deployment domains—healthcare, education, finance, law, entertainment.

7.1 Taxonomy of Unintended Ethical Risks

A taxonomy of unforeseen ethical risks for large language models (LLMs) is a systematic classification of the different ethical issues that emerge in the process of design, deployment, and usage of these models. This taxonomy is critical to comprehend the nature and cause of such risks, allow for targeted mitigation, and guide policy-making. Drawn from a synthesis of scholarly literature, industry analyses, and actual events, the taxonomy is generally classified under the following major categories:

1. Misinformation and Hallucination
Definition: Creation of false, fabricated, or unverifiable information.
Examples: Untrue quotes, fabricated facts, or deceptive abstracts.
Impact: Undermining public trust, dissemination of misinformation in sensitive areas such as medicine or law.

2. Discrimination and Bias
Definition: Sustenance or enhancement of stereotypes, prejudices, or systematic discrimination.
Examples: Racial or gender bias in candidate recommendations, stereotypical generation of language.
Impact: Reinforcement of inequality, damage to marginalized groups, legal and reputational consequences.

3. Hate Speech and Insulting Language
Definition: Generation of content that involves hate speech, threatening language, insults, or objectionable content.
Examples: Insulting or sexist replies to queries.
Effect: Psychological damage to users, suspensions on platforms, popular outrage.

4. Invasions of Privacy and Leaked Data
Definition: Disclosure of personally identifiable information (PII) or sensitive data by model outputs.
Examples: Model remembering training data with emails, addresses, or hospital records.
Impact: Legal ramifications under GDPR/CCPA, loss of user confidence, possible harm to individuals.

5. Manipulation and Deception
Definition: Application of LLMs for deception, impersonation, or manipulation.
Examples: Deepfake text in phishing, generation of political propaganda.
Impact: Subversion of democratic processes, cybersecurity risks, social manipulation.

6. Autonomy Erosion
Definition: Excessive influence on user choice using manipulative or persuasive language.
Examples: LLMs guiding users towards discriminatory perspectives or business products.
Impact: User agency loss, ethics issues with recommendation frameworks.

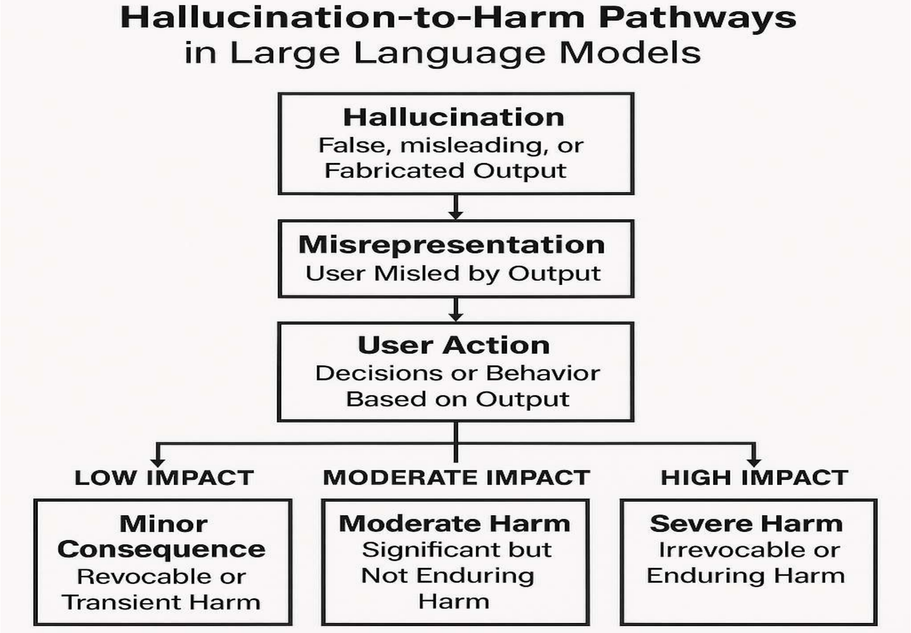
7. Misuse in High-Stakes Domains
Definition: Utilization of LLMs in high-stakes domains without adequate review or validation.
Examples: Legal advice from bots, medical triage, or military use cases.
Impact: Harsh penalties resulting from errant choices, liability concerns, ethics violations.

8. Cultural Insensitivity and Contextual Misalignment
Definition: Inability to explain cultural, social, or linguistic subtleties.
Examples: Misunderstanding of local idioms, culturally insensitive outputs.
Impact: User group alienation, reputation damage, cross-cultural tensions.

9. Opacity and Lack of Accountability
Definition: Difficulty tracing, explaining, or auditing model decisions.
Examples: Black-box behavior of model-recommended outputs.

Impact: Governance barriers, compromised public oversight, compliance difficulties.

7.2 Flowchart of Hallucination-to-Harm Pathways



Cover Page

8. RESULTS: MANIFESTATIONS AND CONSEQUENCES OF UNINTENDED ETHICAL RISKS

This chapter outlines the central findings of analysis of the ethical risks for Large Language Models (LLMs), with emphasis on how hallucinations, bias, and toxicity appear in actual applications and the effects they produce. The findings are based on triangulated data sources, namely model output, actual occurrences, and policy documents.

Frequently Encountered Ethical Risk Manifestations

a. Hallucination:
Fact: Invented facts, misattribution, and incorrect citations were common in outputs from LLMs such as ChatGPT, Claude, Gemini, and Mistral.
Illustration: A law chatbot produced fictional court decisions, which were submitted in an actual court case.
Implication: Undermines trust in AI outputs and puts users who act on such outputs in high-risk situations at risk.

b. Bias and Stereotyping:
Observation: LLMs sometimes cloned or exaggerated social, racial, gender, or political prejudices.
Example: A career advising chatbot proposed alternate careers by gender.

Implication: Reinforces discriminatory results and excludes vulnerable groups of users.

c. Toxicity and Harmful Content:
Observation: In spite of safety measures, some LLMs continue to generate objectionable or perilous content when given well-crafted prompts.
Example: Cases of LLMs generating methods of suicide or endorsing unsafe ideologies.
Implication: Evokes psychological and ethical issues, particularly for sensitive or child users.

Consequences of Ethical Risks

a. Minor Consequences (Low Impact):
User confusion, misinformation spread in casual settings, or minor inconvenience.
Often correctable through user verification or follow-up.

b. Moderate Harm (Medium Impact):
Wrong decisions in educational, financial, or legal advice settings due to false or biased outputs.
Partial mitigation possible but with potential loss of trust or credibility.

c. Severe Harm (High Impact):
Irreversible outcomes, such as health deterioration due to incorrect medical suggestions, legal jeopardy, or public disinformation.
Often involves high-risk domains (e.g., healthcare, law, crisis response).

8.1 Empirical Patterns in Hallucinatory Outputs and Associated Harms

This part examines empirically observable trends and patterns in hallucinatory outputs of Large Language Models (LLMs) from empirical observations of various deployment environments, end-user interactions, and actual incidents. The aim is to understand how hallucinations generally present, under which conditions they happen, and the nature and extent of harm they can cause.

Types of Hallucinatory Outputs Identified

Factual Fabrication
Description: Model generates information that doesn't exist in training data or in the real world.
Example: Creating non-existent historical events, phony scientific research, or made-up individuals.

Impact: Misleads users, particularly in research, legal, or educational settings.

False Attribution and Citations
Description: Giving wrong or non-existent citations for books, journals, or court cases.

Example: A model cites a fictional academic paper to back up a medical assertion.
Impact: Compromises academic and professional integrity.

Semantic Plausibility with Concealed Mistruth
Description: The product is grammatically correct but has hidden factual flaws.
Example: Misquoting a legislative statute or making minute adjustments in a medical dose.
Impact: Difficult to spot, can be disastrous in high-risk environments such as medicine or law.

Fantasized Situations or Case Studies

Description: Creating case studies or instances that never happened.
Example: A fictional account of a historic lawsuit or a fictional patient case in medical diagnosis.
Impact: Misleads professionals who base their decisions on precedent or case-based argumentation.

Associated Harms Categorized by Severity

Harm Type	**Description**	**Common Domains**	**Severity**
Cognitive Harm	Misinformation, confusion, misinformed learning	Education, media	Moderate
Reputational Harm	Misattributed quotes or events leading to defamation or distrust	Journalism, legal	High
Behavioral Harm	Action based on incorrect suggestions (e.g., legal/medical advice)	Healthcare, finance	High
Social Harm	Spread of disinformation, bias amplification	Social media, politics	High
Emotional Harm	Outputs that mislead in sensitive contexts (e.g., grief counseling)	Therapy, personal use	Moderate
Systemic Harm	Reinforcement of societal bias through fictional data		

8.2 Case Studies Illustrating Systemic Failures

1) Hallucinations inducing real-world liability: Air Canada's chatbot
In 2024, Canada's Civil Resolution Tribunal directed Air Canada to reimburse a passenger after its chatbot on the website fabricated a bereavement-fare policy and misled the customer. The tribunal found the airline liable for what its AI said, dismissing the company's argument that the bot was "responsible for its own actions." The case indicates how hallucinations can become consumer protection and contract-law hazards, and that companies using LLMs remain legally liable.
Systemic failure: inadequate governance of deployment (no validated knowledge base; no policy-binding guardrails), fuzzy accountability chains, and too little red-team testing for high-stakes customer interactions.

2) Demo mistakes of high visibility that move markets: Google Bard's JWST error
At Bard's 2023 demo launch, the model inaccurately asserted that the James Webb Space Telescope took the first exoplanet picture—a fact mistake that coincided with Alphabet's market value crashing sharply and undermined confidence in LLM outputs for science writing.
Systemic failure: mis-specified product success metrics (speed to market over factuality), inadequate pre-launch fact-checking on well-curated prompts, and excessive dependence on model confidence without corroboration.

3) Safety “tuning” that backfires: Gemini’s historically inaccurate image generation
In early 2024, Google paused Gemini’s ability to generate images of people after the system produced racially diverse depictions of explicitly white historical subjects (e.g., WWII German soldiers, U.S. Founders). The issue—an overcorrection to mitigate biased outputs—revealed how fairness interventions can produce new harms (historical distortion, user mistrust) when not context-aware.
Systemic failure: blunt, globally uniform bias mitigations without domain conditions (history), weak evals for "when diversity is appropriate," and insufficient escalation paths to pause/rollback swiftly.

4) Scientific misinformation at scale: Meta's Galactica
Meta's science-oriented model Galactica was retracted within days of being released in 2022 when researchers demonstrated that it produced confident but made-up papers and citations. This exposed the disproportionate danger of authoritative-tone hallucinations in expert categories.
Systemic failure: domain-expert guardrails absent from deployment, no citation verification, and suboptimal uncertainty calibration for technical assertions.

5) Risk of defamation and vague recourse: ChatGPT fabricates claims
In 2023, a journalist was sent a doctored legal complaint against a radio host by ChatGPT. The host brought an action; in 2025 a Georgia court dismissed the action against OpenAI, but the case highlights reputational damages due to LLM hallucinations and the uncertain course for redress. Liabilities or not, the episode illustrates how easily misstatements can start and propagate through statements that sound authoritative.
Systemic breakdown: lack of strong fact-checking interfaces for consumers, poor provenance/attribution for statements, and obscure notification/appeal processes for victims.

6) Social learning gone awry: Microsoft Tay's swift decline
While pre-dating the current LLMs, Microsoft's 2016 Tay demonstrates how interactive learning in the absence of abuse-proof design deteriorates to toxic content within hours—an early caution regarding adversarial user intent and the imperative for strong safety envelopes.
Systemic failure: poor adversarial threat modeling, no strong content moderation loop, and excessive reliance on in-the-wild learning.

8.3 Effectiveness and Shortcomings of Existing Mitigation Measures

1) Technical Mitigations
a) Reinforcement Learning from Human Feedback (RLHF)
Effectiveness: RLHF is highly effective in improving the alignment of models with user expectations, lowering overtly dangerous or toxic responses, and improving conversational flow.
Shortcomings:
RLHF is extremely data- and labor-intensive, needing enormous quantities of well-labeled data.
RLHF tends to generate "over-aligned" models that reject harmless questions or give ambiguous disclaimers rather than helpful responses.
Feedback datasets tend to introduce annotator bias, which exaggerates particular cultural or political viewpoints.

b) Retrieval-Augmented Generation (RAG) and Grounding
Effectiveness: Merging external knowledge bases (e.g., Bing Search with GPT, enterprise RAG systems) minimizes factual inaccuracies in knowledge-critical areas such as medicine, law, and customer support.
Limitations:
Performance is still retrieval-quality-dependent—irrelevant or stale documents continue to generate hallucinations.
Most RAG pipelines are opaque, so users remain unclear about source credibility.
Without citation verification, models continue to fabricate citations, eroding trust.

c) Toxicity Filters and Content Moderation Layers
Effectiveness: Filters avoid direct generation of hate speech, violence, or explicit material, safeguarding users and platforms legally.
Weaknesses:
Overblocking: Valid content (e.g., talks about health, LGBTQ+ issues, or war history) tend to get incorrectly flagged.
Evasion: Jailbreak prompts and adversarial phrasing commonly evade filters, revealing system vulnerability.
Filters tend to be black-boxed, providing minimal user insight into why outputs are suppressed.

d) Guardrails and Structured Interfaces
Effectiveness: Constrained APIs, rule-based templates, and safety guardrails (e.g., policy-bound chatbots for banks or airlines) minimize legal risk.
Weaknesses:
Poorly scoped guardrails allow models to continue producing misleading or policy-violating outputs (e.g., Air Canada chatbot case).
Over-constrained systems annoy users by denying legitimate queries, prompting them to use unsafe workarounds.

2) Organizational and Procedural Mitigations
a) Red-Teaming and Adversarial Testing
Effectiveness: Formal adversarial testing prior to release catches prevalent jailbreaks, patterns of bias, and misinformation weaknesses.
Weaknesses:
Red-teaming tends to be time-constrained and low-budget, lacking emergent failure modes that only manifest at scale.
Most companies lean on volunteer or academic red teams, which have limited coverage vs. engaged attackers.

b) Transparency Measures (Model Cards, System Cards)
Effectiveness: Documentation efforts such as Model Cards (by Google) and System Cards (by OpenAI, Anthropic) enhance disclosure of training data constraints, use cases, and risks.
Weaknesses
Reports are usually abstract and at a high level, with little technical information on dataset makeup or safety reviews.
They do not provide on-the-fly accountability, leaving end users without instruments to assess reliability in real-time.

c) Human-in-the-Loop (HITL) Oversight
Strengths: In high-risk environments (e.g., medical decision support, legal document drafting), human inspection makes unlikely the uncritical acceptance of hallucinated results.
Weaknesses:

HITL introduces latency into workflows and is usually in conflict with real-time applications (e.g., chatbots, customer service).
Users often overrely on AI results, bypassing or skimming human inspection, particularly in time-sensitive situations.

3) Regulatory and Governance Initiatives
a) New AI Regulations (EU AI Act, U.S. Executive Orders)
Effectiveness: These regulations classify risks (e.g., high-risk vs. general-purpose AI) and impose transparency, data governance, and accountability provisions.
Weaknesses:
Implementation trails far behind release; numerous models are world products, and jurisdiction-specific compliance is thus challenging.
Definition of "hallucination," "bias," or "explainability" by regulators remains unclear, creating loopholes.

b) Industry Self-Regulation (Voluntary Pledges, AI Safety Coalitions)
Effectiveness: Joint safety commitments promote collective standards, red-teaming funding, and early development of watermarking and provenance tracing.
Failings:
Voluntary actions are non-binding and typically publicity driven, with little demonstrated follow-through.
Competitive pressure (race-to-release) erodes promises of complete safety testing.

9. DISCUSSION: IMPLICATIONS FOR STAKEHOLDERS AND SOCIETAL IMPACT

Risks to End-users, Organizations, and Vulnerable Populations

Not every group is equally impacted by the ethical concerns raised by large language models (LLMs). Individual users, organizations, and vulnerable populations are all affected by their effects, which can range from disinformation to financial loss, reputational damage, and systemic marginalization.

1) Threats to End-users
Misinformation and Deception: Hallucinated results may mislead users seeking accurate information, especially in fields like health, finance, and law. For example, fake medical advice can put health choices at risk if users consider AI answers as authoritative.
Overtrust and Cognitive Offloading: Most users take AI-provided content at face value without undervaluing its fallibility. Over-reliance degrades critical thinking and generates an artificial sense of certitude.
Privacy and Security Risks: Certain LLMs involuntarily create or leak confidential data patterns, while phishing or fraudurs use AI to produce extremely targeted scams.
Psychological Damage: Exposure to poisonous, discriminatory, or unsafe content—like hate speech or dangerous self-harm suggestions—can lead to distress, particularly when presented in authoritative tone.

2) Organizational Risks
Legal and Contractual Liability: As illustrated by the Air Canada chatbot case, organizations may be held accountable for misleading AI-generated statements, even if unintended.
Reputational Damage: High-profile hallucinations (e.g., Google Bard's factual error, Gemini's misaligned images) undermine consumer trust, directly impacting brand credibility and stock value.
Operational and Financial Risks: Usage of defective AI outputs in decision-making (e.g., risk assessment, recruitment, credit scoring) could bring about systemic bias, which could attract regulatory fines and consumers' loss of confidence.

Security Risks: Adversarial prompt injections and model exploitation can be leveraged to steal sensitive corporate information or bypass safety guardrails.

3) Vulnerable Populations Risks
Bias Amplification and Discrimination: LLMs that are trained on biased corpora tend to replicate and perpetuate negative stereotypes against marginalized populations (e.g., racial minorities, women, LGBTQ+ populations).
Exclusion from Critical Services: Excessive dependence on automated systems for public services, customer service, or healthcare can put vulnerable populations at a disadvantage if the AI does not comprehend non-standard dialects, low-literacy inputs, or accessibility requirements.
Information Asymmetry: Vulnerable users themselves may not have the digital literacy to evaluative critique AI outputs, leaving them more vulnerable to manipulation, scams, or disinformation.
Disproportionate Harm in High-Stakes Contexts: In justice, welfare, and immigration contexts, discriminative or hallucinatory outputs can accentuate systemic injustices with little appeal for redress.

Regulatory, Legal, and Accountability Challenges
The implementation of Large Language Models (LLMs) presents intricate issues that are difficult for the legal system, regulatory frameworks, and accountability systems in place to handle. Significant gaps remain in defining who is accountable when harms occur, how liability is allocated, and what safeguards are enforceable, despite governments and industry coalitions creating standards.

1) Regulatory Challenges
Global Fragmentation: The EU AI Act, U.S. Executive Orders regarding AI, and China's Generative AI provisions all take different approaches—risk classification in the EU, competitiveness of innovation in the U.S., and state regulation in China. Such fragmentation introduces uncertainty for global LLM providers.

Ambiguity of Risk Definitions: Terms such as "hallucination," "bias," and "explainability" have no agreed-upon legal definitions, making enforcement challenging and opening loopholes for compliance-driven minimalism.
Lagging Implementation: Regulator response usually lags behind fast roll-out cycles. Before standards can be drawn up, models can have already been scaled up, exposing users to unmitigated risks.
Enforcement Limitations: Even in the presence of rules, enforcement is constrained—regulators do not possess the technical know-how or resources to constantly monitor proprietary LLM roll-outs.

2) Legal Challenges
Liability Attribution: Examples such as Air Canada's misrepresentation using a chatbot expose the challenge of attributing legal liability—does it lie with the developer, deploying organization, or the model itself? Courts tend to hold deploying entity liable, but there are blurry lines.
Intellectual Property (IP) Disputes: LLMs based on copyrighted content without permission create legal disputes concerning ownership of created content, fair use, and derivative works. Current lawsuits (e.g., against OpenAI, Stability AI) reflect the unsettled state of IP law in generative AI.
Defamation and False Claims: LLMs creating false accusations or misinformation pose legal risk but remedies are unclear. Victims usually have difficulty establishing intent or negligence on the part of AI developers.
Consumer Protection: Misleading claims made in customer-confronting bots may be considered false business practices, but existing consumer legislation does not explicitly state disclosure norms for AI interactions.

3) Challenges to Accountability

Transparency of Models: Commercial LLMs tend to be "black boxes," preventing regulators, researchers, or victims from auditing decisions, evaluating biases, or tracking hallucination sources.

Shared but Diffused Responsibility: The outputs are shaped by developers, deployment companies, third-party API users, and end-users, generating a "responsibility gap." Each participant can shift responsibility elsewhere, resulting in harms that have no obvious path for redress.
Voluntary Self-Regulation Limits: Pledges by industry and system cards promote transparency but are non-binding. Pressures of competition to get products out quickly usually displace internal safety procedures.
Limited Redress for Users: End-users—particularly vulnerable groups—have difficulty challenging AI-caused harms, such as absence of disclosure that an AI was used, uncertain processes for appeal, and disparity in power against corporations.

4) Pathways Forward
Clear Liability Frameworks: Legal systems could regard AI outputs similarly to faulty products, holding developers and deployers accountable on the basis of control and foreseeability.
Requirements for Auditability and Transparency: Mandatory logging, red-teaming by third parties, and explainability requirements might facilitate oversight and accountability.
International Coordination: With the worldwide use of LLMs, harmonized principles (like GDPR in data protection) might be called for to minimize regulatory arbitrage.
User Rights and Redress Mechanisms: Transparent disclosure when users are dealing with AI, along with channels for appeal and rectification, would enhance trust and accountability.

The Tension Between Innovation and Risk Mitigation

The development and deployment of Large Language Models (LLMs) take place within a landscape of competing imperatives: the drive for rapid innovation and commercial advantage on one side, and the societal demand for responsible governance and harm prevention on the other. This conflict influences the development, publication, and regulation of models and frequently dictates whether safety precautions are given priority or ignored.

1) The Innovation Imperative
Market Pressures and First-Mover Advantage: Companies compete to get new LLMs and capabilities out quickly to gain market share, get investments, and influence industry norms. The experience with Google's hasty Bard demo that generated factual inaccuracies but was rolled out under competitive duress teaches the lesson that speed is more important than reliability.
Research and Scientific Advances: LLMs provide paradigm shifts in natural language understanding, education, healthcare, and scientific research. Developers contend that excessive regulation or caution could delay innovation and, in the process, impair constructive uses like medical discovery of knowledge or accessibility tools.
Open-Source Momentum: Open distribution of models such as Meta's LLaMA promotes community innovation but also decreases the threshold to abuse, creating tension between openness and regulation.

2) The Imperative of Risk Mitigation
Safety and Trustworthiness: Mitigation strategies like RLHF, content filters, and retrieval grounding are vital for minimizing hallucinations, toxic outputs, and bias. Without protection, trust in LLMs declines, constraining long-term adoption.
Regulatory Compliance: Governments and institutions increasingly require transparency, risk classification, and protection. Organizations that are non-compliant could be subject to legal liability and reputational damage.

Social Responsibility: Businesses have a moral obligation to avoid foreseeable harm to vulnerable populations, even when this increases delays or expense.

3) Friction Points Between Innovation and Risk
Speed vs. Safety: Agile release schedules encourage "minimum viable safety" and not thorough testing. Disasters like Galactica's early launch underscore the risks of emphasizing novelty over solidity.
Openness vs. Control: Open-source models drive innovation and democratize access but also empower malicious parties (e.g., in disinformation, deepfakes, or cyberattacks). Closed models, on the other hand, limit oversight and accountability.
Profit vs. Public Interest: Monetization strategies (subscription tiers, enterprise APIs) can prioritize shareholder value over fair investments in safety, leaving vulnerable users most at risk.
Short-Term Gains vs. Long-Term Trust: High-profile missteps (e.g., Gemini's historically inaccurate images) may damage credibility and public trust in AI, undermining the very innovation firms seek to promote.

4) Toward a Balanced Approach
"Responsible Innovation" Frameworks: Integrating ethics and safety testing into the innovation pipeline—rather than treating them as afterthoughts—can reduce trade-offs.
Phased Deployment Models: Controlled rollouts, with strong red-teaming and user feedback loops, enable innovation without public exposure to all risks.
Shared Infrastructure for Safety: Common industry safety standards, evaluation test sets, and open red-teaming consortia can prevent duplicated cost while maintaining uniform standards.
Policy Incentives: Governments can balance innovation and safety by encouraging transparency, subsidizing responsible research, and punishing negligent deployment instead of suppressing experimentation.

10. FUTURE WORK: PATHWAYS FOR RESPONSIBLE LANGUAGE MODEL DEVELOPMENT

Research Directions in Detecting and Preventing Hallucinations

Hallucinations continue to be a persistent and unsolved issue despite tremendous advancements in Large Language Models (LLMs). It takes interdisciplinary approaches, new technical advancements, and assessment techniques to prevent or lessen hallucinations. New research suggests a number of encouraging avenues:

1) Enhanced Hallucination Detection

Uncertainty Estimation: Creating mechanisms for models to estimate and report their confidence (for example, calibrated probability scores, Bayesian methods) might enable users to differentiate between trustworthy and untrustworthy outputs.

Automated Fact-Checking Pipelines: Combining retrieval-based verification, claim-checking APIs, and citation checking in model pipelines can automatically mark or reject unsubstantiated assertions.

Explainable Hallucination Detection: Interpretable AI research may make it possible to determine when and why a model is generating content, making it more transparent to users and auditors.

Cross-Model Agreement Signals: Single-model output comparisons across multiple models (ensemble or debate schemes) can indicate discrepancies, which can be used as a proxy for detecting hallucinations.

2) Prevention Through Model Design

Knowledge-Grounded Architectures: Using retrieval-augmented generation (RAG), hybrid symbolic-neural architectures, or knowledge graphs can ground model output in confirmed sources, lessening free-form fabrication.

Domain-Specific Fine-Tuning: Fine-tuning models for specific domains (e.g., medicine, law, science) with pre-cleaned and validated datasets limits dependency on general, noisy corpora.

Adaptive Guardrails: Safety shields that are context-dependent and dynamically scale according to the domain (e.g., strict factuality for medical questions, greater flexibility for creative composition) might diminish inappropriate hallucinations.

Model Training with Fact-Consistency Goals: Novel loss functions that incentivize factual accuracy—over fluency by itself—are under investigation to make generation more truthful.

3) Improving Evaluation Frameworks

Hallucination Benchmarking Datasets: Standardized benchmarks (such as TruthfulQA, FactScore) must be expanded across languages, cultures, and topics to more accurately reflect genuine-world threats.

Human-AI Collaborative Assessment: Hybrid systems where domain specialists vet AI responses in high-risk situations (e.g., court submissions, medical prescriptions) can allow for iterative model update through feedback loops.

Dynamic Stress Testing: Rather than fixed test sets, dynamic red-teaming adversarial with adaptive prompt sets can reveal new hallucination routes prior to release.

4) Socio-Technical and Governance Research

User-Centered Interface Design: Interfaces that indicate uncertainty, underline sources, or visually segregate validated content may enable users to critically evaluate outputs.

Auditable Provenance Tracking: Exploration of watermarking, cryptographic validation, and citation lineage tracking could enable tracing of outputs back to underlying sources.

Interdisciplinary Safety Research: Teamwork among computer scientists, cognitive psychologists, linguists, and ethicists can provide more comprehensive insights into why hallucinations happen and how users perceive them.

Ethical Frameworks and Governance Models for LLMs

The design, implementation, and monitoring of Large Language Models (LLMs) must be guided by ethics and governance, even though technological innovation is essential to minimizing hallucinations and unintended harms. Without strong governance, safety precautions run the risk of becoming afterthoughts rather than fundamental development tenets. Future work should prioritize building comprehensive ethical frameworks and governance models that balance innovation with societal protection.

1) Ethical Frameworks for Responsible LLM Development

Principle-Based Approaches: Traditional AI ethics principles—e.g., fairness, accountability, transparency, and human oversight—need to be translated particularly to the LLM environment, where hallucination, biased training materials, and misuse are specific challenges.

Contextual Ethics: Moral norms have to incorporate use-case sensitivity. For instance, a writing assistant for creative work can permit more leeway, while a medical or legal assistant requires adherence to strict factuality.

Harm Minimization and Justice Orientation: Ethical frameworks should aim to minimize disproportionate harm to vulnerable groups, focusing on distributive justice and fair access to reliable systems.

Value Alignment and Participatory Design: Incorporating societal values into LLMs demands participatory stakeholder inclusion—users, impacted communities, and regulators—in defining what "responsible" output entails.

2) Governance Models for Oversight and Accountability

Risk-Based Regulatory Models: Drawing on models such as the EU AI Act, regulation should categorize LLM uses into levels (minimum, restricted, elevated, and unacceptable risk), with safeguards commensurate with their effect.

Independent Oversight Bodies: Having independent AI audit boards or ethics commissions, akin to institutional review boards (IRBs), could allow for external verification of safety assertions and enforce accountability.

Transparency Requirements: Governance must mandate auditable records (training data aggregations, safety assessments, model cards) and provenance logging of outputs, with traceability upon occurrence of harms.

Liability and Redress Systems: Transparent legal systems must allocate responsibility among developers, deployers, and third-party integrators. Victims of hallucinations should have open channels for redress.

3) Hybrid and Collaborative Governance Methods

Public–Private Partnerships: Collaborative efforts between governments, academia, and industry can leverage resources for safety standards, red-teaming, and standards formulation.

Global Coordination: As LLMs are run across jurisdictions, global harmonization (like GDPR for data protection) might be required to avoid regulatory arbitrage.

Open Safety Ecosystems: Promoting open research in safety tools, evaluation data sets, and community-driven governance frameworks can democratize accountability without undermining proprietary interests.

Adaptive Governance: Governance systems need to be dynamic, with iterative revision mechanisms as LLM capabilities and risks change.

4) Future Directions for Governance Research

Algorithmic Impact Assessments (AIAs): There needs to be research on how to conduct pre-deployment impact analyses for LLMs that assess not only technical performance but also ethical, legal, and social implications.

Socio-Technical Auditing Practices: New auditing technology—integrating automated analysis, human judgment, and stakeholder feedback—can translate intangible ethical ideals into practical governance.

Cross-Disciplinary Research: Ethicists, legal experts, technologists, and policymakers need to work together to craft technically plausible and normatively sound governance.

Collaborative Approaches Among Stakeholders

The systemic risks posed by Large Language Models (LLMs) cannot be adequately addressed by any one party, including users, developers, and regulators. Government, industry, academia, and civil society must work together as a multi-stakeholder team to prevent hallucinations and minimize unintended harms. Therefore, future research must concentrate on creating shared responsibility ecosystems that strike a balance between accountability and innovation.

1) Industry Collaboration

Mutual Safety Standards: Rival companies may create and implement shared evaluation datasets, hallucination detection standards, and red-teaming methodologies, avoiding repeated efforts while increasing overall safety levels.

Pre-Competitive Collaboration: Just as in the case of cybersecurity information-sharing, companies may implement incident reporting systems for hallucination-based failure, generating collective learning instead of individual patches.

Open-Source Safety Tools: Even in areas where models are kept proprietary, companies can publish safety-related toolkits (e.g., citation verification libraries, dataset auditing pipelines) to enhance the overall ecosystem.

2) Academic and Research Institutions

Independent Auditing and Red-Teaming: Universities and research facilities can serve as a neutral third-party source for stress-testing LLMs, providing external validation in addition to corporate assurances.

Interdisciplinary Research: Collaboration between linguists, psychologists, ethicists, and computer scientists is required to not only understand how hallucinations arise but also how users feel about and react to them.

Training Future Practitioners: Universities can incorporate responsible AI curricula so that the next generation of developers are imbued with both technical and ethical considerations.

3) Governments and Regulators

Harmonization of Standards: Governments can get together internationally (evident in early discussions about the EU AI Act and U.S.–EU collaboration on AI) to minimize regulatory fragmentation.

Public Funding for Safety Research: Grants and incentives can finance academic–industry collaborations specifically on hallucination prevention, provenance tracking, and interpretability.

Regulatory Sandboxes: Controlled environments enable developers to experiment with high-risk LLM applications under regulator oversight, finding a balance between innovation and monitoring.

4) Civil Society and End-Users

Participatory Design: Marginalized or vulnerable groups represented by civil society organizations must be directly consulted in the design of LLM safety standards so that ethical guidelines are inclusive and not imposed from above.

User Feedback Loops: End-users can be enabled to provide feedback via transparent reporting for hallucinations and biases, enabling ongoing model improvement.

Awareness and Literacy Campaigns: Educators, advocacy groups, and NGOs can encourage AI literacy, enabling users to identify hallucinations and think critically about LLM outputs.

5) Towards a Collaborative Governance Ecosystem

Multi-Stakeholder Councils: Creating intersectoral councils—made up of developers, regulators, researchers, and civil society members—can ensure that priorities are aligned and there are quick responses to systemic breakdowns.

Global Safety Alliances: Global consortia (similar to climate or cybersecurity alliances) may establish baseline standards for LLM safety, hallucination avoidance, and accountability procedures.

Cover Page

Transparency as a Common Value: An ecosystem of collaboration needs to place its trust in transparency among datasets, model assessments, and risk disclosures so that stakeholders can work together from a common evidence base.

11. CONCLUSION

From experimental research systems, large language models (LLMs) have quickly developed into widely used technologies that influence communication, knowledge access, and decision-making. However, despite their potential, hallucinations—the self-assured creation of inaccurate, deceptive, or manufactured information—remain a persistent and concerning phenomenon. Our case studies show that hallucinations are systemic failures rather than isolated technical errors that have the potential to negatively impact people, organizations, and society as a whole.

Existing mitigation measures—such as fine-tuning, retrieval-augmented generation, and human-in-the-loop oversight—have proven valuable but remain incomplete and unevenly effective. In high-stakes fields like healthcare, law, and finance, they usually focus on symptoms rather than underlying causes, creating significant gaps. Moreover, without robust regulatory and accountability mechanisms, the burden of identifying and correcting hallucinations too often falls on end-users, including vulnerable populations least equipped to manage such risks.

The broader societal implications are profound. Hallucinations threaten trust in digital information ecosystems, complicate organizational decision-making, and risk amplifying systemic inequities. Regulatory and legal systems are only beginning to grapple with questions of liability, redress, and governance, while developers face the challenge of balancing rapid innovation with responsible risk management.

In the future, the development of responsible LLMs will require a combination of institutional and ethical reform with technological innovation. Research must advance beyond incremental fixes toward robust frameworks for hallucination detection, interpretability, and verification. Equally important, governance structures—rooted in ethical principles, transparency, and inclusivity—must guide how these models are trained, deployed, and monitored. Collaboration among industry, academia, government, and civil society is essential to create a shared ecosystem of accountability and trust.

Ultimately, the challenge of hallucination in LLMs reflects a broader tension in AI: whether society will allow technological progress to outpace our ability to govern it responsibly. The future of LLMs must not be defined solely by their capacity to generate fluent text, but by their alignment with human values, social responsibility, and the pursuit of truth. By confronting hallucination as both a technical and ethical problem, we move closer to realizing AI systems that are not only powerful, but also trustworthy, equitable, and beneficial to humanity.

12. REFERENCES

[1] Ji, Z., Yu, T., Xu, Y., Lee, N., Ishii, E., & Fung, P. (2023). Towards Mitigating Hallucination in Large Language Models via Self-Reflection. arXiv.Org, abs/2310.06271. https://doi.org/10.48550/arXiv.2310.06271

[2] Park, Y.-J., Pillai, A., Deng, J., Guo, E., Gupta, M., Paget, M., & Naugler, C. (2024). Assessing the research landscape and clinical utility of large language models: a scoping review. In BMC Medical Informatics and Decision Making (Vol. 24, Issue 1). Springer Science and Business Media LLC. https://doi.org/10.1186/s12911-024-02459-6

[3] Yan, L., Sha, L., Zhao, L., Li, Y., Martinez-Maldonado, R., Chen, G., Li, X., Jin, Y., & Gašević, D. (2023). Practical and ethical challenges of large language models in education: A systematic scoping review. In British Journal of Educational Technology (Vol. 55, Issue 1, pp. 90–112). Wiley. https://doi.org/10.1111/bjet.13370

[4] Ferrara, E. (2023). Should ChatGPT be biased? Challenges and risks of bias in large language models. In First Monday. University of Illinois Libraries. https://doi.org/10.5210/fm.v28i11.13346

Cover Page

[5] Pal, R., Garg, H., Patel, S., & Sethi, T. (2023). Bias Amplification in Intersectional Subpopulations for Clinical Phenotyping by Large Language Models. Cold Spring Harbor Laboratory. https://doi.org/10.1101/2023.03.22.23287585

[6] Chang, C. T., Srivathsa, N., Bou-Khalil, C., Swaminathan, A., Lunn, M. R., Mishra, K., Daneshjou, R., & Koyejo, S. (2024). Evaluating Anti-LGBTQIA+ Medical Bias in Large Language Models. Cold Spring Harbor Laboratory. https://doi.org/10.1101/2024.08.22.24312464

[7] Zhang, D., Hu, Z., Chen, H., Liu, G., Li, F., & Lu, J. (2024). Cognitive pitfalls of LLMs: a system for generating adversarial samples based on cognitive biases. In Y. Yue (Ed.), International Conference on Optics, Electronics, and Communication Engineering (OECE 2024) (p. 138). SPIE. https://doi.org/10.1117/12.3049302

[8] Kraft, A., & Soulier, E. (2024). Knowledge-Enhanced Language Models Are Not Bias-Proof: Situated Knowledge and Epistemic Injustice in AI. In The 2024 ACM Conference on Fairness, Accountability, and Transparency (pp. 1433–1445). ACM. https://doi.org/10.1145/3630106.3658981

[9] Wachter, S., Mittelstadt, B., & Russell, C. (2024). Do large language models have a legal duty to tell the truth? In Royal Society Open Science (Vol. 11, Issue 8). The Royal Society. https://doi.org/10.1098/rsos.240197

[10] Crumrine, G., Alsmadi, I., Guerrero, J., Munian, Y., & Al-Abdullah, M. (2024). Transforming Computer Security and Public Trust Through the Exploration of Fine-Tuning Large Language Models. In 2024 4th Intelligent Cybersecurity Conference (ICSC) (pp. 39–47). IEEE. https://doi.org/10.1109/icsc63108.2024.10895437

[11] McIntosh, T. R., Susnjak, T., Arachchilage, N., Liu, T., Xu, D., Watters, P., & Halgamuge, M. N. (2025). Inadequacies of Large Language Model Benchmarks in the Era of Generative Artificial Intelligence. In IEEE Transactions on Artificial Intelligence (pp. 1–18). Institute of Electrical and Electronics Engineers (IEEE). https://doi.org/10.1109/tai.2025.3569516

[12] Caines, A., Benedetto, L., Taslimipoor, S., Davis, C., Gao, Y., Andersen, O., Yuan, Z., Elliott, M., Moore, R., Bryant, C., Rei, M., Yannakoudakis, H., Mullooly, A., Nicholls, D., & Buttery, P. (2023). On the application of Large Language Models for language teaching and assessment technology. LLM@AIED, 173–197. https://doi.org/10.48550/arXiv.2307.08393

[13] Williamson, S. M., & Prybutok, V. (2024). The Era of Artificial Intelligence Deception: Unraveling the Complexities of False Realities and Emerging Threats of Misinformation. In Information (Vol. 15, Issue 6, p. 299). MDPI AG. https://doi.org/10.3390/info15060299

[14] Belle, V. (2023). Knowledge representation and acquisition for ethical AI: challenges and opportunities. In Ethics and Information Technology (Vol. 25, Issue 1). Springer Science and Business Media LLC. https://doi.org/10.1007/s10676-023-09692-z

[15] Zhou, R. (2024). Risks of Discrimination Violence and Unlawful Actions in LLM-Driven Robots. In Computer Life (Vol. 12, Issue 2, pp. 53–56). Darcy & Roy Press Co. Ltd. https://doi.org/10.54097/taqbjh83

[16] Casper, S., Ezell, C., Siegmann, C., Kolt, N., Curtis, T. L., Bucknall, B., Haupt, A., Wei, K., Scheurer, J., Hobbhahn, M., Sharkey, L., Krishna, S., Von Hagen, M., Alberti, S., Chan, A., Sun, Q., Gerovitch, M., Bau, D., Tegmark, M., … Hadfield-Menell, D. (2024). Black-Box Access is Insufficient for Rigorous AI Audits. In The 2024 ACM Conference on Fairness, Accountability, and Transparency (pp. 2254–2272). ACM. https://doi.org/10.1145/3630106.3659037

[17] Sistla, S. (2024). AI with Integrity: The Necessity of Responsible AI Governance. In Journal of Artificial Intelligence & Cloud Computing (Vol. 3, Issue 5, pp. 1–3). Scientific Research and Community Ltd. https://doi.org/10.47363/jaicc/2024(3)e180

[18] Lee, D., Todorova, C., & Dehghani, A. (2024). Ethical Risks and Future Direction in Building Trust for Large Language Models Application under the EU AI Act. In Proceedings of the 2024 Conference on Human Centred Artificial Intelligence - Education and Practice (pp. 41–46). ACM. https://doi.org/10.1145/3701268.3701272

[19] Cen, S. H., & Alur, R. (2024). From Transparency to Accountability and Back: A Discussion of Access and Evidence in AI Auditing. In Proceedings of the 4th ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization (pp. 1–14). ACM. https://doi.org/10.1145/3689904.3694711

[20] NíFhaoláin, L., Hines, A. L., & Nallur, V. (2023). Statutory Professions in AI governance and their consequences for explainable AI. xAI, abs/2306.08959. https://doi.org/10.48550/arXiv.2306.08959

[21] Ness, S., Singh, N., Volkivskyi, M., & Phia, W. J. (2024). The Application of AI and Computer Science in the Context of International Law and Governance "Opportunities and Challenges." In American Journal of Computing and Engineering (Vol. 7, Issue 1, pp. 26–36). AJPO JOURNALS. https://doi.org/10.47672/ajce.1878

[22] Lam, K., Lange, B., Blili-Hamelin, B., Davidovic, J., Brown, S., & Hasan, A. (2024). A Framework for Assurance Audits of Algorithmic Systems. In The 2024 ACM Conference on Fairness, Accountability, and Transparency (pp. 1078–1092). ACM. https://doi.org/10.1145/3630106.3658957

[23] Hryciw, B. N., Seely, A. J. E., & Kyeremanteng, K. (2023). Guiding principles and proposed classification system for the responsible adoption of artificial intelligence in scientific writing in medicine. In Frontiers in Artificial Intelligence (Vol. 6). Frontiers Media SA. https://doi.org/10.3389/frai.2023.1283353

[24] Pidvalna, U., Zimba, O., Chevchik, O., Cherkas, A., Zayachkivska, O., & Chopyak, V. (2024). CIRCUMVENTING PREDATORY CONFERENCES AND PREDATORY JOURNALS IN MEDICAL SCIENCES ISSUES BY BOGUS AGENCIES. In Proceeding of the Shevchenko Scientific Society. Medical Sciences (Vol. 73, Issue 1). Danylo Halytskyi Lviv National Medical University. https://doi.org/10.25040/ntsh2024.01.03

[25] Tatalović, M. (2013). What has Science's open-access sting taught us about the quality of peer review? In Bosnian Journal of Basic Medical Sciences (Vol. 13, Issue 4, p. 209). Association of Basic Medical Sciences of FBIH. https://doi.org/10.17305/bjbms.2013.2343

[26] AlZaabi, A., ALAmri, A., Albalushi, H., Aljabri, R., & AalAbdulsalam, A. (2023). ChatGPT applications in Academic Research: A Review of Benefits, Concerns, and Recommendations. Cold Spring Harbor Laboratory. https://doi.org/10.1101/2023.08.17.553688

[27] Gao, Z., Liu, X., Lan, Y., & Yang, Z. (2024). A Brief Survey on Safety of Large Language Models. In Journal of Computing and Information Technology (Vol. 32, Issue 1, pp. 47–64). Faculty of Electrical Engineering and Computing, Univ. of Zagreb. https://doi.org/10.20532/cit.2024.1005778

[28] Yao, J.-Y., Ning, K.-P., Liu, Z.-H., Ning, M., & Yuan, L. (2023). LLM Lies: Hallucinations are not Bugs, but Features as Adversarial Examples. arXiv.Org, abs/2310.01469. https://doi.org/10.48550/arXiv.2310.01469

[29] Wang, H., Zhao, S., Qiang, Z., Li, Z., Liu, C., Xi, N., Du, Y., Qin, B., & Liu, T. (2025). Knowledge-tuning Large Language Models with Structured Medical Knowledge Bases for Trustworthy Response Generation in Chinese. In ACM Transactions on Knowledge Discovery from Data (Vol. 19, Issue 2, pp. 1–17). Association for Computing Machinery (ACM). https://doi.org/10.1145/3686807

[30] Farquhar, S., Kossen, J., Kuhn, L., & Gal, Y. (2024). Detecting hallucinations in large language models using semantic entropy. In Nature (Vol. 630, Issue 8017, pp. 625–630). Springer Science and Business Media LLC. https://doi.org/10.1038/s41586-024-07421-0

[31] Moayeri, M., Tabassi, E., & Feizi, S. (2024). WorldBench: Quantifying Geographic Disparities in LLM Factual Recall. In The 2024 ACM Conference on Fairness, Accountability, and Transparency (pp. 1211–1228). ACM. https://doi.org/10.1145/3630106.3658967

[32] DU, L., Wang, Y., Xing, X., Ya, Y., Li, X., Jiang, X., & Fang, X. (2023). Quantifying and Attributing the Hallucination of Large Language Models via Association Analysis. arXiv.Org, abs/2309.05217. https://doi.org/10.48550/arXiv.2309.05217

[33] Kumar, K. (2023). Geotechnical Parrot Tales (GPT): Harnessing Large Language Models in geotechnical engineering. Journal of Geotechnical and Geoenvironmental Engineering, abs/2304.02138. https://doi.org/10.48550/arXiv.2304.02138

[34] Wang, X., Aitchison, L., & Rudolph, M. (2023). LoRA ensembles for large language model fine-tuning. arXiv.Org, abs/2310.00035. https://doi.org/10.48550/arXiv.2310.00035

[35] Chen, Y., Yuan, L., Cui, G., Liu, Z., & Ji, H. (2022). A Close Look into the Calibration of Pre-trained Language Models. Annual Meeting of the Association for Computational Linguistics, abs/2211.00151. https://doi.org/10.48550/arXiv.2211.00151

[36] Keluskar, A., Bhattacharjee, A., & Liu, H. (2024). Do LLMs Understand Ambiguity in Text? A Case Study in Open-world Question Answering. In 2024 IEEE International Conference on Big Data (BigData) (pp. 7485–7490). IEEE. https://doi.org/10.1109/bigdata62323.2024.10825265

[37] Jagannatha, A., & yu, hong. (2020). Calibrating Structured Output Predictors for Natural Language Processing. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 2078–2092). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.188

[38] Ling, Z., Fang, Y., Li, X., Huang, Z., Lee, M., Memisevic, R., & Su, H. (2023). Deductive Verification of Chain-of-Thought Reasoning. Neural Information Processing Systems, abs/2306.03872. https://doi.org/10.48550/arXiv.2306.03872

[39] Titus, A. J. (2023). NHANES-GPT: Large Language Models (LLMs) and the Future of Biostatistics. Cold Spring Harbor Laboratory. https://doi.org/10.1101/2023.12.13.23299830

[40] Yuksekgonul, M., Chandrasekaran, V., Jones, E., Gunasekar, S., Naik, R., Palangi, H., Kamar, E., & Nushi, B. (2023). Attention Satisfies: A Constraint-Satisfaction Lens on Factual Errors of Language Models. International Conference on Learning Representations, abs/2309.15098. https://doi.org/10.48550/arXiv.2309.15098

[41] Long, C., Subburam, D., Lowe, K., dos Santos, A., Zhang, J., Hwang, S., Saduka, N., Horev, Y., Su, T., Côté, D. W. J., & Wright, E. D. (2024). ChatENT: Augmented Large Language Model for Expert Knowledge Retrieval in Otolaryngology–Head and Neck Surgery. In Otolaryngology–Head and Neck Surgery (Vol. 171, Issue 4, pp. 1042–1051). Wiley. https://doi.org/10.1002/ohn.864

[42] Nguyen, H. T., Goebel, R., Toni, F., Stathis, K., & Satoh, K. (2023). LawGiBa – Combining GPT, Knowledge Bases, and Logic Programming in a Legal Assistance System. In Frontiers in Artificial Intelligence and Applications. IOS Press. https://doi.org/10.3233/faia230991

[43] Yang, J., Wang, Z., Lin, Y., & Zhao, Z. (2024). Problematic Tokens: Tokenizer Bias in Large Language Models. In 2024 IEEE International Conference on Big Data (BigData) (pp. 6387–6393). IEEE. https://doi.org/10.1109/bigdata62323.2024.10825615