Chapter 12
Reimagining Reality:
Innovations in AI Hallucination Management and Trustworthy Generation

Atanu Nag
https://orcid.org/0000-0003-0561-1946
IFTM University, India

Debika Chaudhuri
IFTM University, India

ABSTRACT

AI hallucination—the production of fluent but incorrect or fabricated outputs—has become a major challenge for large language models and generative AI. This chapter reviews the evolution of hallucination management, from early recognition to modern mitigation and creative reinterpretation. It defines hallucination in AI contexts and identifies key causes such as data noise, model drift, prompt ambiguity, and domain mismatch. The chapter then examines evaluation frameworks, taxonomies, and metrics, including factual consistency measures, FActScore, and domain-specific benchmarks. Core mitigation strategies are discussed, notably retrieval-augmented generation, multi-stage verification, human-in-the-loop approaches, and prompt optimization. Emerging trends include agentic AI systems with multi-agent validation, sector-specific solutions in healthcare and finance, and viewing hallucination as a driver of human–machine co-creation. The chapter concludes that effective hallucination management is a technical and ethical necessity for trustworthy AI.

DOI: 10.4018/979-8-3373-7534-2.ch012

INTRODUCTION

The phenomenon of hallucination in artificial intelligence (AI) systems—particularly in large language models (LLMs)—has emerged as one of the defining challenges in the pursuit of trustworthy, ethically aligned, and socially responsible AI (Förster & Skop, 2025; Lin et al., 2024, Li et al., 2024). Hallucination refers to cases in which AI systems produce outputs that are coherent, grammatically fluent, and often highly persuasive, yet lack factual grounding or contextual relevance. These outputs may include fabricated information, inaccurate reasoning, or invented citations, all of which can compromise reliability when LLMs are deployed in sensitive domains such as education, governance, healthcare, and law (Dahl et al., 2024; Anh-Hoang et al., 2025). As generative AI technologies proliferate, understanding, mitigating, and reimagining hallucination has become central to the broader discourse on trustworthy AI. Early examinations of hallucination focused on defining the phenomenon and documenting its behavior across diverse tasks. Researchers observed that LLMs frequently produce statements that are 'plausible yet ungrounded,' reflecting the probabilistic nature of neural text generation rather than true comprehension (Kryściński et al., 2020; Farquhar et al., 2024; Lin et al., 2024). Computer science literature defined hallucination as a generative error that leads to misinformation or logical inconsistency, emphasizing its role in distorting the semantic mapping between inputs and outputs (Anh-Hoang et al., 2025; Kazlaris et al., 2025; van Deemter, 2024). Meanwhile, scholars in law, clinical decision support, and computational education highlighted hallucination as a critical threat to user trust, noting that fabricated content in high-stakes scenarios could lead to serious harm (Dahl et al., 2024). This recognition marked the first stage in the evolution of hallucination research: definition and conceptualization. During this phase, hallucination was often framed as an enigmatic behavior embedded within the 'black-box' functioning of deep learning models, raising fundamental questions about interpretability, epistemic validity, and safety. Following conceptual clarity, the research community moved toward building systematic taxonomies and evaluative frameworks (Huang et al., 2024). A turning point came with the development of quantitative hallucination metrics designed to measure factual consistency. The FActScore metric, for instance, introduced a structured, evidence-based evaluation pipeline for assessing the factual grounding of LLM outputs (Min et al., 2023). In parallel, the summarization community developed fine-grained typologies of factual errors, distinguishing between intrinsic hallucinations (not derived from source text) and extrinsic hallucinations (unsupported by external knowledge) (Maynez, 2020). In healthcare, benchmarks such as med-HALT and clinical-HALT systematically captured the risks posed by hallucinations in biomedical reasoning and clinical summarization (Pal et al., 2023; Asgari et al., 2025). These tools demonstrated that