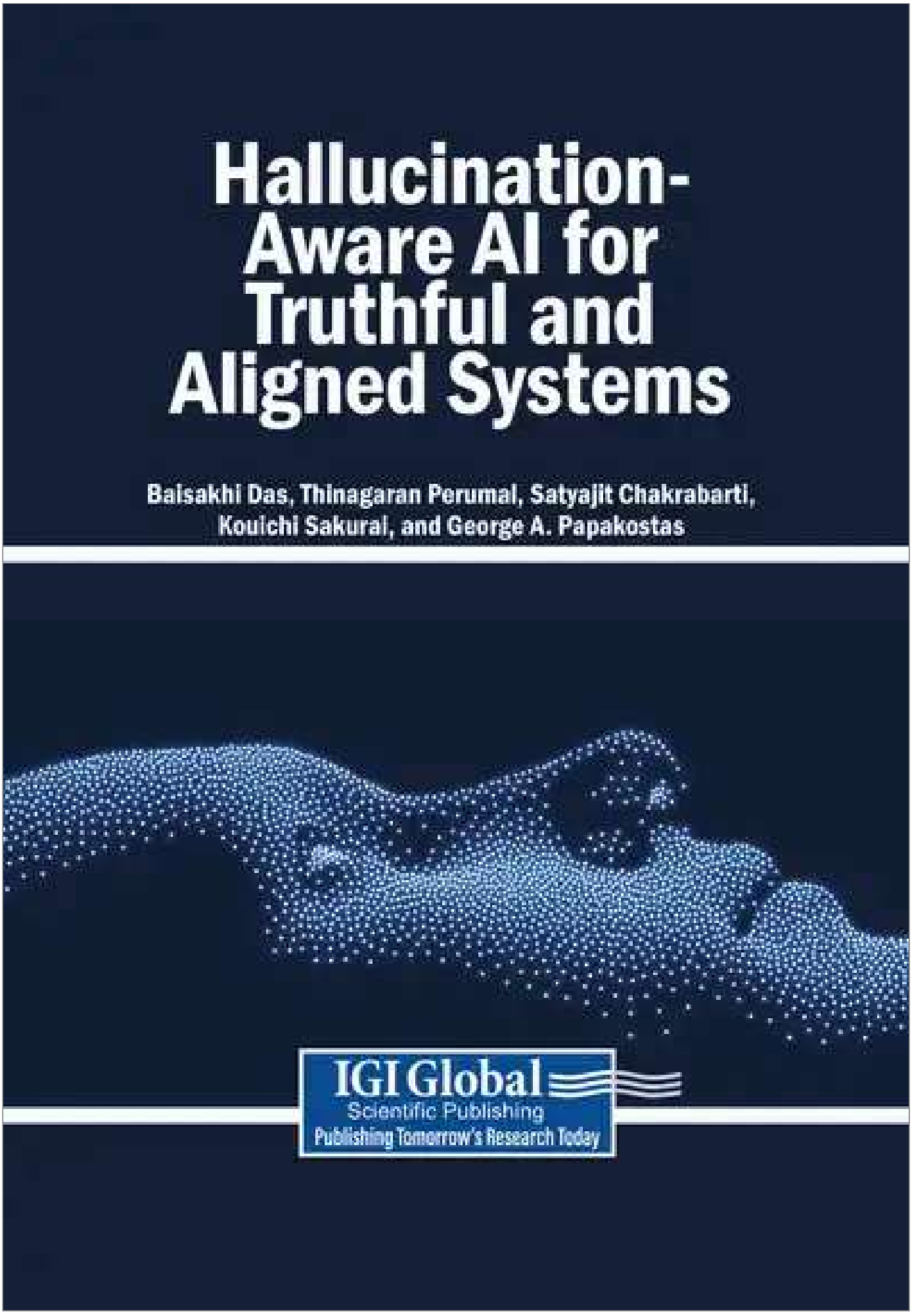

Hallucination in AI Systems: Understanding and Framework ⊗

Swarnalee Ray (Institute of Engineering and Management, India), Apurba Paul (Institute of Engineering and Management, India), Dipankar Das (Jadavpur University, India), and Jiya Raj (Institute of Engineering and Management, India)

Source Title: Hallucination-Aware AI for Truthful and Aligned Systems (/gateway/book/389724)

Pages: 32

ISBN13: 9798337373201ISBN13 Softcover: 9798337373218EISBN13: 9798337373225

DOI: 10.4018/979-8-3373-7320-1.ch001

Abstract

Hallucination in large language models (LLMs) refers to the ability to generate coherent but factually incorrect text that is not source-verified and sufficiently logically justified. This problem is caused by the probabilistic nature of sequence modeling and the limitations of the Transformer architecture. If not addressed in high-risk areas such as medicine, law, and autonomous vehicles, hallucinations can negatively affect the reliability, safety, and trustworthiness of users. The current chapter offers a comprehensive review of the theoretical bases, typologies, causes, assessment strategies, and solutions to the problem of hallucination. One of the key contributions of this chapter is the development of a diagnostic framework that associates sensitivity to the prompt, the intrinsic properties of the model, and hallucinations.

Full Text Preview

Introduction

LLM Hallucinations should not be considered mere computer bugs since LLMs have been developed to predict the most probable sequence of words. Thus, it is normal for them to drift away from reality occasionally. This means that LLMs can produce coherent and confident responses that are not based in reality. When applied in critical industries like healthcare, law, and finance, hallucinations in LLMs pose serious risks of medical mistakes, legal repercussions, and financial damage. To ensure the safe use of LLMs, it is therefore necessary to shift focus from zero-error performance to effective hallucination detection and multi-layered mitigation techniques like Retrieval-Augmented Generation (RAG), the incorporation of structured knowledge, and ongoing human monitoring (Lakera, 2025).

The implications of unmitigated hallucinations are particularly evident in critical domains where the integrity of information is critical.

- Healthcare: The creation of deceptive medical information, such as erroneous medication dosages or outdated medical treatment protocols, poses a direct threat to patient safety. Moreover, the use of fabricated citations in AI-based systematic reviews may also impinge upon the integrity of medical research (Kim et al., 2025).
- Law: There are known cases of lawyers being penalized for using AI-based legal document submissions that included deceptive citations of medical cases and non-existent legal precedents, thus impinging upon the integrity of the judicial system (Deloitte, n.d.).
- Autonomous Agents: In the case of AI systems that interact with the physical world, inaccuracies in information may have direct implications for safety. “Tool Calling Hallucinations,” where an agent invents parameters for an API call, may also lead to incorrect action sequences with potentially disastrous consequences.

This chapter presents a structured analysis of hallucinations in AI systems by categorising their forms, examining their theoretical and practical causes, and reviewing existing evaluation and mitigation approaches. In addition, the second portion presents a theoretical model that accounts for why such errors happen to them. It stresses the importance of understanding the design theory inherent in the framework itself, along with the limits placed upon any specific instance of reality perception (such as hallucination) by that framework if one wants to measure the phenomenon properly.

Foundations Of Hallucination In Ai Systems

This section explores the theory of how hallucinations function as inevitable outcomes of AI technologies and training approaches - moving beyond definitions of hallucinations. These theoretical foundations are important because they help us understand why hallucinations are a management issue, rather than simply an issue of eliminating them.

Formal Definition

“Hallucination” in AI systems can be defined formally as the creation of new content that is not only factually false but also irrelevant/unsupported by any of the source materials or logically inconsistent, despite having a fluent and coherent presentation. Hallucinations include everything from simple factual errors (e.g., incorrect dates or facts) to producing new, non-existent (and often complex) entities/data. There are several key distinctions to make here; hallucinations may be confused with a related (but different) phenomenon known as “Bias” (the presence of a persistent representational skew, which may not involve factual errors); with “noise” (the presence of unpredictable, stochastic randomness); and with “Model Uncertainty” (the level of certainty that exists when predicting future events based on historical examples), but all of these phenomena can occur together as contributing factors to the creation of “hallucinated” content (Palo Alto Networks, n.d.).

Continue Reading (/gateway/chapter/full-text-html/410312)

References

Follow Reference

Anh-Hoang D. Tran V. Nguyen L.-M. (2025). Survey and analysis of hallucinations in large language models: Attribution to prompting strategies or model behavior.Frontiers in Artificial Intelligence, 8, 1622292. 10.3389/frai.2025.162229241098969

Baker Donelson. (n.d.). *The Perils of Legal Hallucinations and the Need for AI Training for Your In-House Legal Team!* Retrieved from https://www.bakerdonelson.com/the-perils-of-legal-hallucinations-and-the-need-for-ai-training-for-your-in-house-legal-team (https://www.bakerdonelson.com/the-perils-of-legal-hallucinations-and-the-need-for-ai-training-for-your-in-house-legal-team)

Follow Reference

Bang Y. Ji Z. Schelten A. Hartshorn A. Fowler T. Zhang C. Cancedda N. Fung P. (2025). HalluLens: LLM Hallucination Benchmark. ACL Anthology.

Deloitte. (n.d.). *AI doesn't lie, it hallucinates and M&A due diligence must address that*. Retrieved from https://www.deloitte.com/ch/en/services/consulting/perspectives/ai-hallucinations-new-risk-m-a.html (https://www.deloitte.com/ch/en/services/consulting/perspectives/ai-hallucinations-new-risk-m-a.html)

Follow Reference

Ho D. E. (2025). Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools. Stanford University.

IAPP. (n.d.). *Hallucinations in LLMs: Technical challenges, systemic risks and AI governance implications*. Retrieved from https://iapp.org/news/a/hallucinations-in-llms-technical-challenges-systemic-risks-and-ai-governance-implications (https://iapp.org/news/a/hallucinations-in-llms-technical-challenges-systemic-risks-and-ai-governance-implications)

Follow Reference

Kazlaris I. Antoniou E. Diamantaras K. Bratsas C. (2025). From Illusion to Insight: A Taxonomic Survey of Hallucination Mitigation Techniques in LLMs.AI, 6(10), 260. 10.3390/ai6100260

Follow Reference

Kim Y. Jeong H. Chen S. Li S. S. Park C. Lu M. Alhamoud K. Mun J. Grau C. Jung M. Gameiro R. Fan L. Park E. Lin T. Yoon J. Yoon W. Sap M. Tsvetkov Y. Liang P. P. Breazeal C. (2025). Medical Hallucination in Foundation Models and Their Impact on Healthcare.medRxiv. 10.1101/2025.02.28.25323115

Lakera. (2025). *LLM Hallucinations in 2025: How to Understand and Tackle AI's....* Retrieved from https://www.lakera.ai/blog/guide-to-hallucinations-in-large-language-models (https://www.lakera.ai/blog/guide-to-hallucinations-in-large-language-models)

Min, S., et al. (2023). FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation. *NeurIPS*.

Palo Alto Networks. (n.d.). *What Are AI Hallucinations? [+ Protection Tips]*. Retrieved from https://www.paloaltonetworks.com/cyberpedia/what-are-ai-hallucinations (https://www.paloaltonetworks.com/cyberpedia/what-are-ai-hallucinations)

Peng, B., Narayanan, S., & Papadimitriou, C. (2024). On Limitations of the Transformer Architecture. *arXiv preprint arXiv:2402.08164*.

Pinecone. (n.d.). *Understanding Hallucinations in AI: A Comprehensive Guide*. Retrieved from https://www.pinecone.io/learn/ai-hallucinations/ (https://www.pinecone.io/learn/ai-hallucinations/)

Ravichander, A., Ghela, S., Wadden, D., & Choi, Y. (2025). THE HALOGEN BENCHMARK: FANTASTIC LLM HALLUCINATIONS AND WHERE TO FIND THEM. *OpenReview*.

Xenoss. (n.d.). *How to prevent AI hallucinations in production*. Retrieved from https://xenoss.io/blog/how-to-avoid-ai-hallucinations-in-production (https://xenoss.io/blog/how-to-avoid-ai-hallucinations-in-production)

Request Access

You do not own this content. Please login to recommend this title to your institution's librarian or purchase it from the IGI Global Scientific Publishing bookstore (/chapter/hallucination-in-ai-systems/410312)

Username or email:
apurba.saitech2021@gmail.com

Password:
••••••

Log In >

Forgot individual login password? (/gateway/login/reset-password/)

Create individual account (/gateway/login/create-account/)

(http://www.world-forgotten-children.org)

(https://publicationethics.org/category/publisher/igi-global)