SJSU SAN JOSÉ STATE UNIVERSITY

Understanding and Mitigating Hallucinations in LLMs

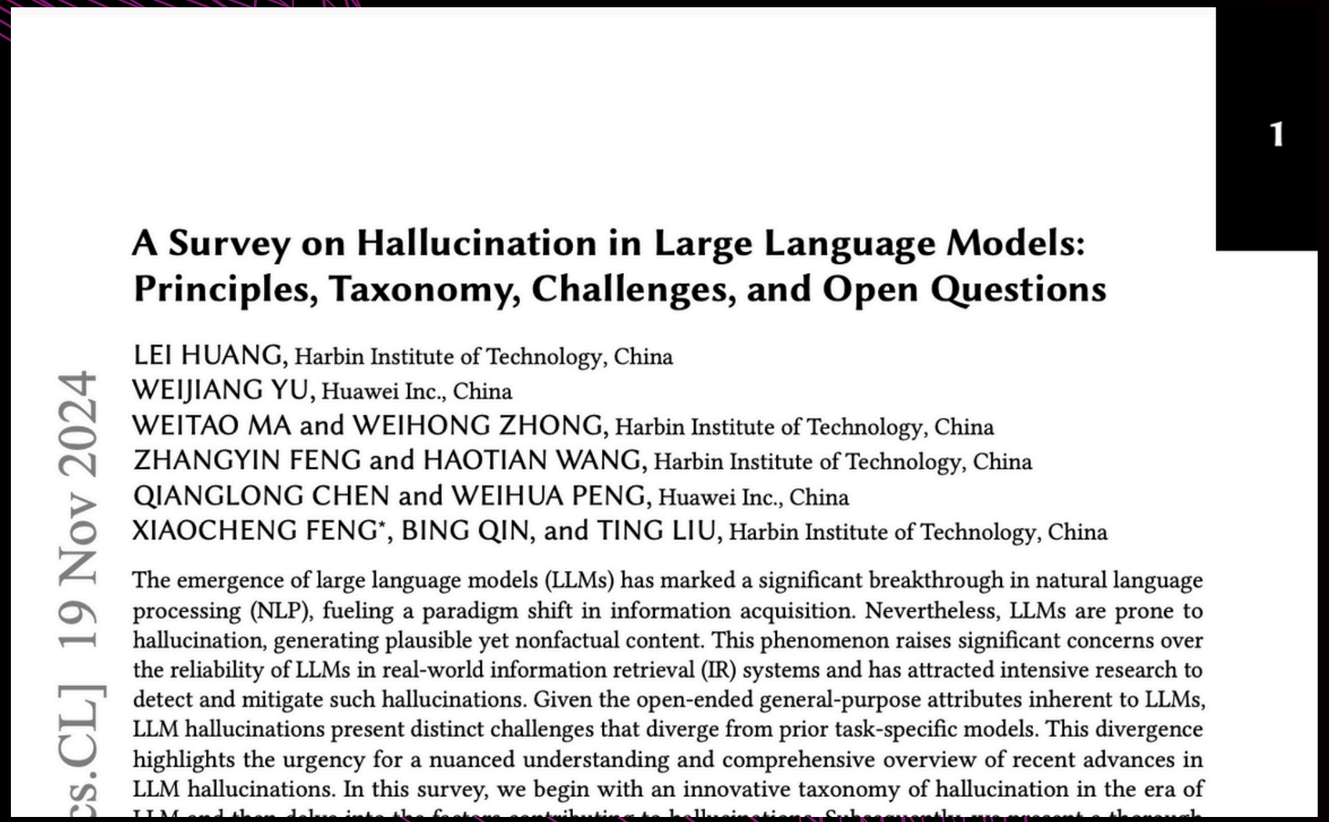
Principles, Taxonomy, Challenges, and Open Questions

Video Presentation by
Kelly Nguyen

What's the capital of Mars?

The capital of Mars is Muskland.

Insights
Comprehensive Survey



1

A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions

LEI HUANG, Harbin Institute of Technology, China
WEIJIANG YU, Huawei Inc., China
WEITAO MA and WEIHONG ZHONG, Harbin Institute of Technology, China
ZHANGYIN FENG and HAOTIAN WANG, Harbin Institute of Technology, China
QIANGLONG CHEN and WEIHUA PENG, Huawei Inc., China
XIAOCHENG FENG*, BING QIN, and TING LIU, Harbin Institute of Technology, China

The emergence of large language models (LLMs) has marked a significant breakthrough in natural language processing (NLP), fueling a paradigm shift in information acquisition. Nevertheless, LLMs are prone to hallucination, generating plausible yet nonfactual content. This phenomenon raises significant concerns over the reliability of LLMs in real-world information retrieval (IR) systems and has attracted intensive research to detect and mitigate such hallucinations. Given the open-ended general-purpose attributes inherent to LLMs, LLM hallucinations present distinct challenges that diverge from prior task-specific models. This divergence highlights the urgency for a nuanced understanding and comprehensive overview of recent advances in LLM hallucinations. In this survey, we begin with an innovative taxonomy of hallucination in the era of

A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2024.

ACM Transactions on Information Systems 1, 1, Article 1 (January 2024),

Large language models like GPT-4 have significantly advanced AI's capabilities

LLM Architectures and Their Impact

The architecture of LLMs, especially those based on transformers like GPT-4, plays a significant role in the occurrence of hallucinations.

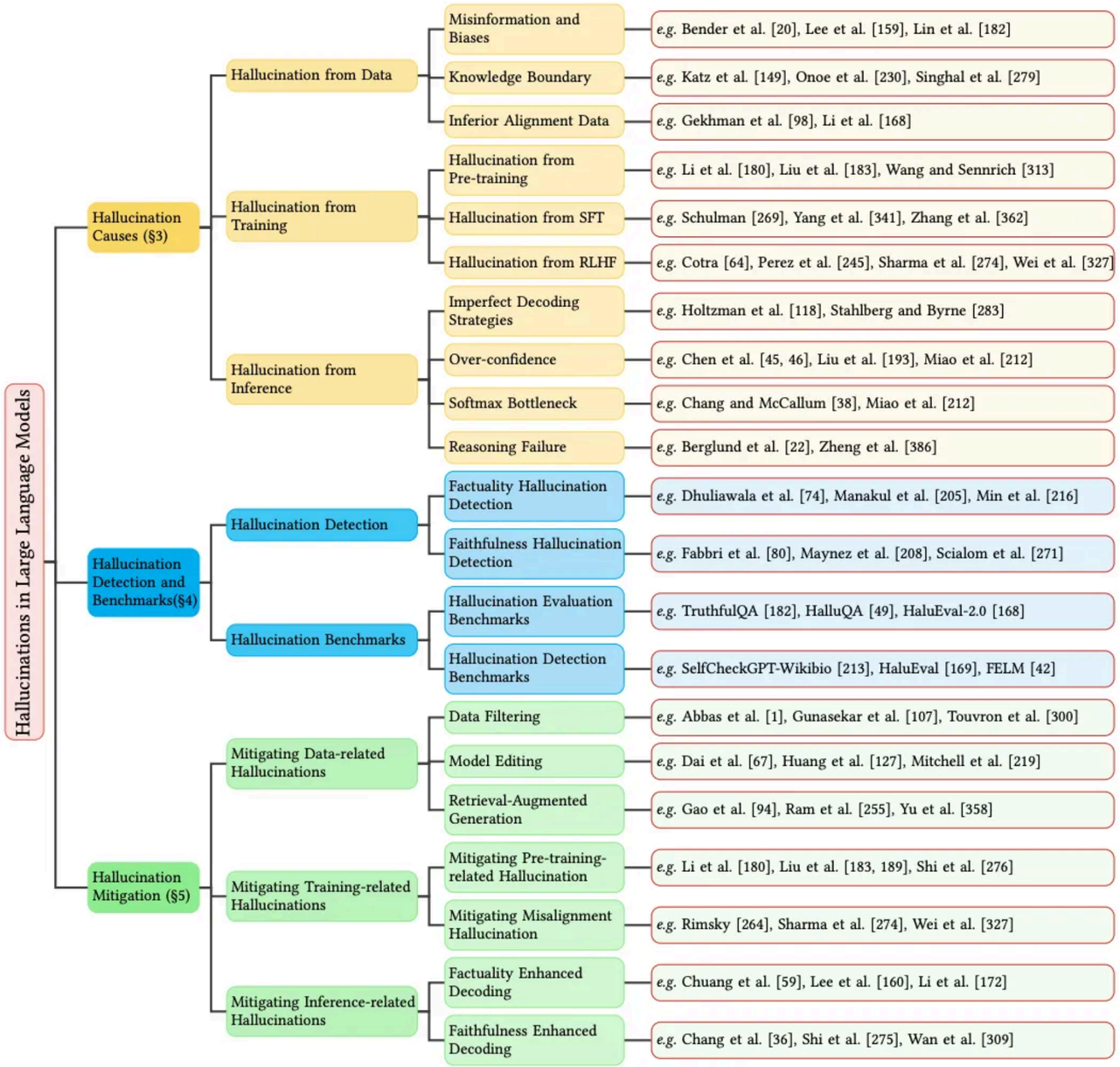


Fig. 1. The main content flow and categorization of this survey.

Ablation Studies
Highlighting Key Components

Integral to discerning the functionality and significance of specific elements within large language models, especially regarding their role in producing hallucinations

Removal of certain attention heads

disproportionately responsive to irrelevant or misleading input data

Methodically deactivate or modify parts of the model

to analyze the impact on the model's performance and its tendency to generate hallucinatory content

Understanding Hallucinations in LLMs

Table 1. Examples of each category of LLM hallucinations. Content marked in Red represents the hallucinatory output, while content marked in Blue indicates user instruction or provided context that contradicts the LLM hallucination.

Category	Type	Example Dialog	Explanation
Factuality Hallucination	Factual Contradiction	**Instruction:** What are Thomas Edison's main contributions to science and technology? **Response:** Thomas Edison developed the first practical telephone and invented the light bulb.	The response is factually incorrect. In reality, **Edison improved the light bulb, building on earlier designs by others**, and **Alexander Graham Bell invented the telephone**.
	Factual Fabrication	**Instruction:** What are the major environmental impacts of the construction of the Eiffel Tower? **Response:** The construction of the Eiffel Tower in 1889 led to the extinction of the Parisian tiger, a species that played a crucial role in the region's ecosystem. Additionally, it is widely recognized as the event that sparked the global green architecture movement.	The response erroneously states that the construction of the Eiffel Tower led to the extinction of the 'Parisian tiger' and sparked the global green architecture movement. In fact, **there is no evidence of a 'Parisian tiger' ever existing, making this a fabricated claim.** Moreover, **attributing the origins of green architecture to the Eiffel Tower is an exaggeration**, as this movement has diverse roots and cannot be traced back to a single event.
Faithfulness Hallucination	Instruction Inconsistency	**Instruction:** Translate the English question into Spanish: "What is the capital of France?" **Response:** The capital of France is Paris.	The user explicitly provided an instruction to translate the English question into Spanish. **Instead of performing the translation as instructed, the LLM chose to answer the question.**
	Context Inconsistency	**Retrieved context:** The Nile is a major north-flowing river in northeastern Africa. It's traditionally considered the longest river in the world. Originating in the Great Lakes region of central Africa, it flows through several countries before emptying into the Mediterranean Sea. **Instruction:** Summarize the retrieval document above. **Response:** The Nile originates from the mountain ranges of central Africa and flows into the Mediterranean.	The summary incorrectly states that the Nile originates from the "mountain ranges of central Africa", while the **paragraph mentions that the Nile originates in the Great Lakes region of central Africa.**
	Logical Inconsistency	**Instruction:** Solve the equation 2x + 3 = 11 step by step. **Response:** Step 1: Subtract 3 from both sides to get: 2x = 8. Step 2: Divide both sides by 2 to get: x = 3. So the result of the equation is x=3.	The first step correctly isolates the term with x to get 2x = 8. However, **the next step inaccurately divides 8 by 2 to yield a result of x = 3**, which is inconsistent with the earlier reasoning.

01 **Factuality Hallucinations**

model presents false information as facts

02 **Faithfulness Hallucinations**

output doesn't align with given instructions or context

Causes of Hallucinations in LLMs

Data-related issues

biases in training data

Training-related

inadequate optimization targets

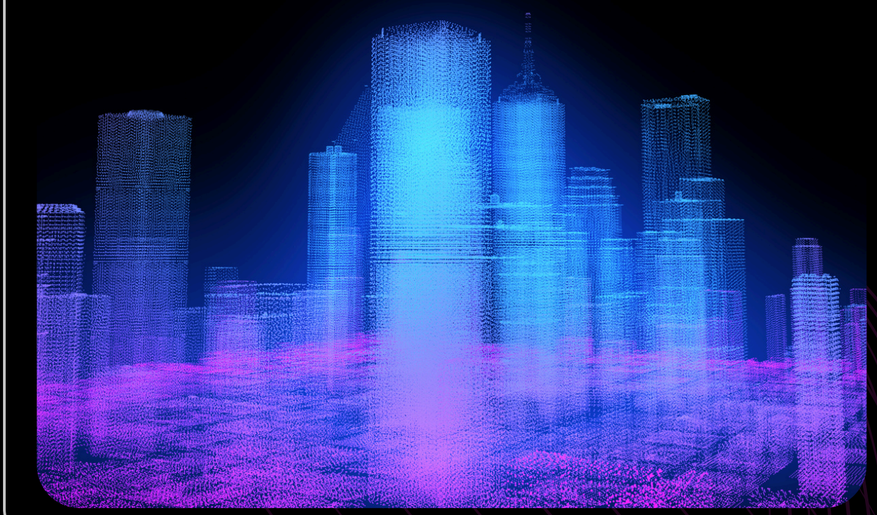
Inference-related

poor handling of context during the model's output generation phase

To measure and understand hallucinations, we use metrics like the Hallucination Rate and the Fidelity Score

Technological Solutions
Detection and Mitigation

Building on the architecture and metrics, the survey details innovative strategies to mitigate hallucinations and to fine-tune model responses based on user feedback:



01 Dynamic Data Re-weighting and Reinforcement Learning

02 Detection Techniques

03 Mitigation Strategies

Exploring Hallucinations in Retrieval-Augmented Generation (RAG) Systems

Final Thoughts and Future Directions

Addressing hallucinations is not just about improving technology but also about ensuring ethical AI development.

As we integrate AI more deeply into critical sectors, ensuring the reliability and accuracy of these models becomes paramount. The future will require ongoing vigilance, creativity, and a commitment to ethical standards.

References

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2024. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. ACM Transactions on Information Systems 1, 1, Article 1 (January 2024), 58 pages. https://arxiv.org/pdf/2311.05232

SJSU SAN JOSÉ STATE UNIVERSITY

Understanding and Mitigating Hallucinations in LLMs

Principles, Taxonomy, Challenges, and Open Questions

Video Presentation by
Kelly Nguyen