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

- Presenter: Abdullah Mamun
- *Date: Dec 11, 2024*

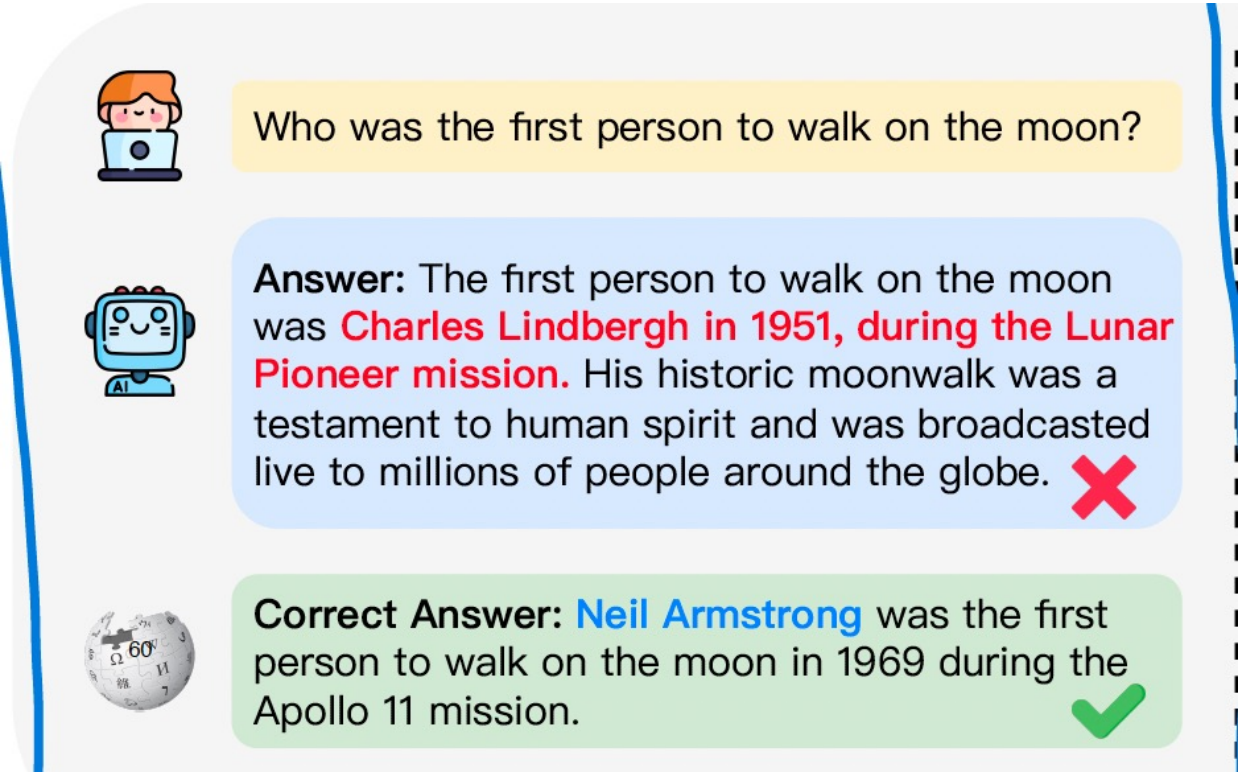
- Just accepted at ACM Transactions on Information Systems
- On arxiv since Nov 2023 and it has 640+ citations as of today

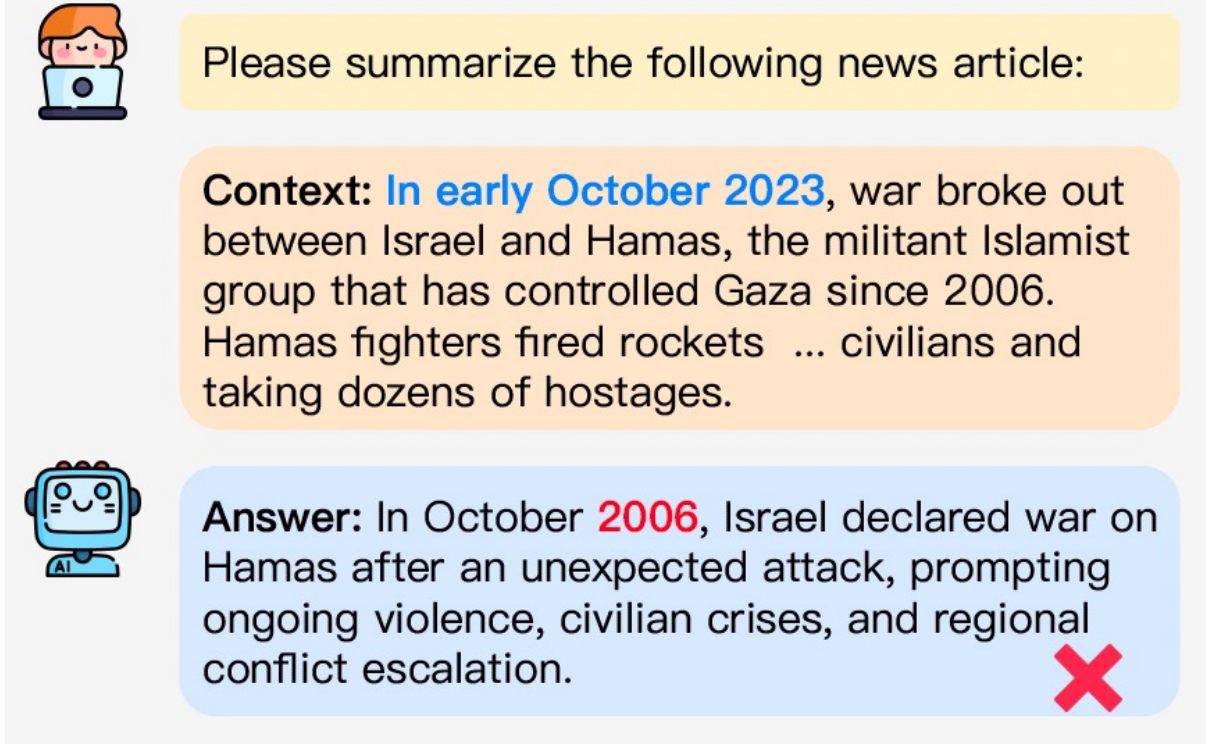
A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions

Lei Huang[1]*, **Weijiang Yu**[2]*, **Weitao Ma**[1], **Weihong Zhong**[1]
Zhangyin Feng[1], **Haotian Wang**[1], **Qianglong Chen**[2], **Weihua Peng**[2]
Xiaocheng Feng[1]†, **Bing Qin**[1], **Ting Liu**[1]
[1]Harbin Institute of Technology, Harbin, China
[2]Huawei Inc., Shenzhen, China
{lhuang,wtma,whzhong,zyfeng,xcfeng†,qinb,tliu}@ir.hit.edu.cn
{weijiangyu8,wanght1998,chenqianglong.ai,pengwh.hit}@gmail.com

What is Hallucination?

In simple terms: generated content that is nonsensical or unfaithful to the provided source content.





Goal of this paper

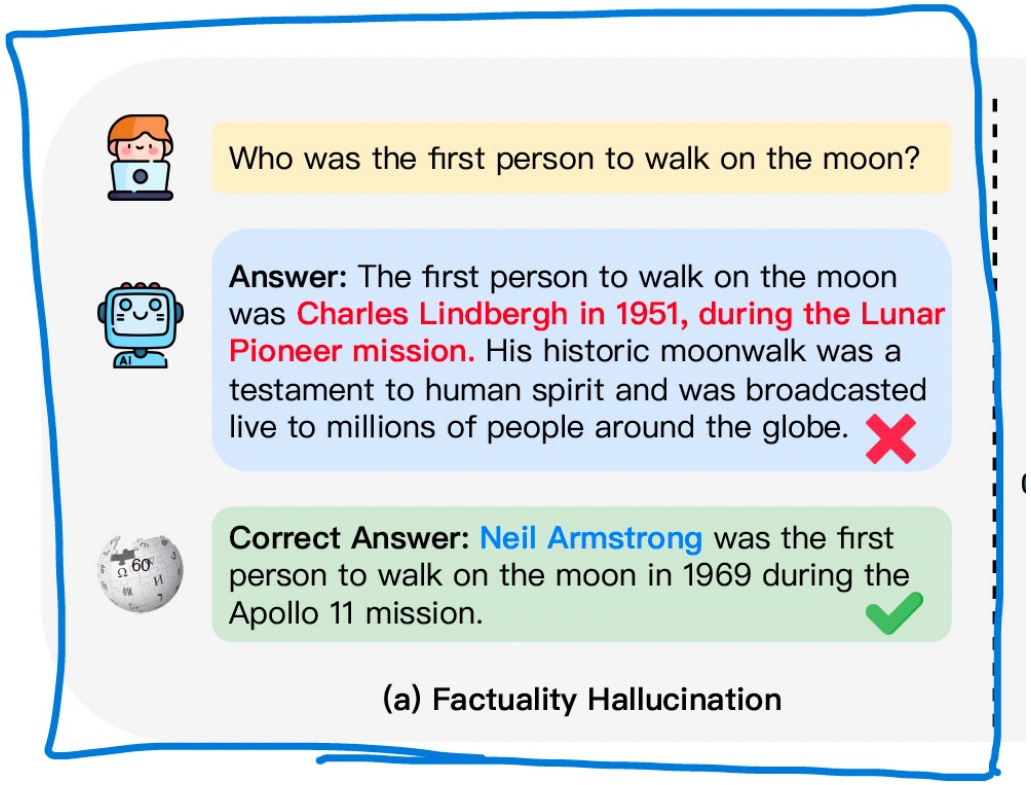
- Redefine the taxonomy of hallucinations
- Present the opportunities and challenges associated with hallucinations in LLMs.

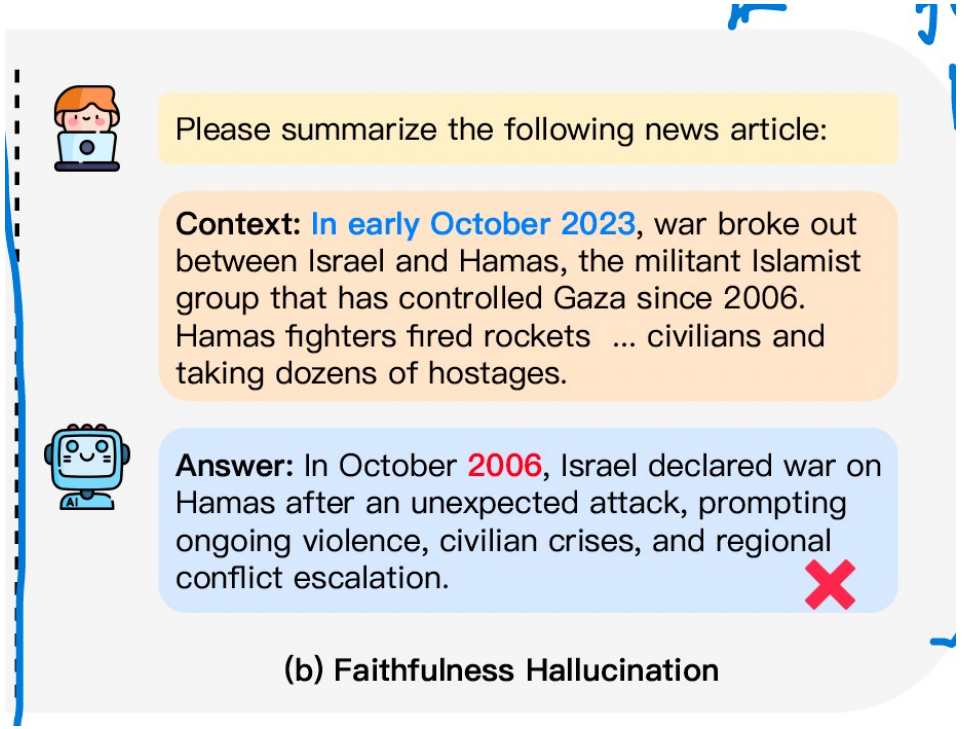
Previous surveys: Ji et al., 2023a, Liu et al., 2023h, Wang et al., 2023c. Each of them covers a specific type or application.

This survey focuses on the trustworthiness of LLMs

Finally, understand the causes, methods to detect, and method to mitigate hallucinations

Factuality vs Faithfullness Hallucinations



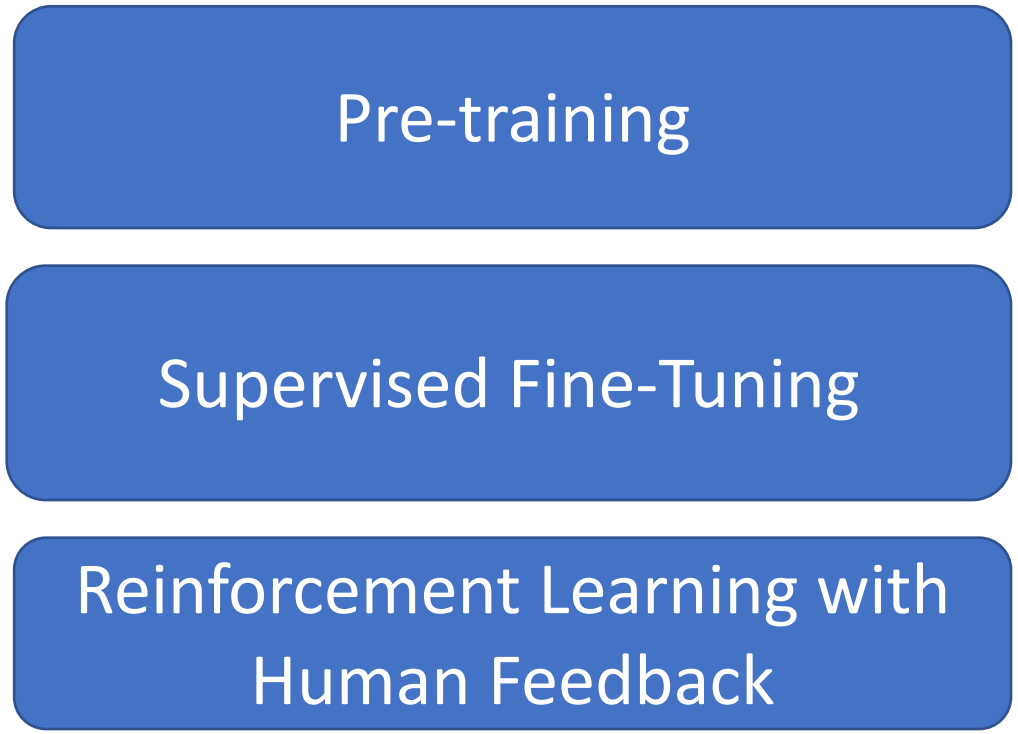


Answer inconsistent with known facts

Answer inconsistent with user-provided context

Training stages of LLMs

- Why? To figure out the origins of the hallcinations



Pre-training

Pre-training Approach

- Models predict the next word in a sequence (autoregressive learning).

Self-Supervised Learning

- Trains on extensive text datasets without labeled examples.

Capabilities Acquired

- Language syntax and grammar understanding.
- World knowledge and reasoning (?) skills.

Foundation for Fine-Tuning:

- Enables efficient adaptation to specific downstream tasks.

```
1  Translate English to French:
2  cheese =>
```

Supervised Fine-Tuning

- **Limitations of Pre-trained LLMs**
  - Function primarily as "completion machines."
  - Misaligned objectives: next-word prediction vs. user-specific responses.
- **Introduction of Supervised Fine-Tuning (SFT)**
  - Involves additional training on annotated instruction-response pairs.
  - Enhances model capabilities and user controllability.
- **Benefits of SFT**
  - Improves generalization to unseen tasks.
  - Proven effectiveness in achieving high performance.

Reinforcement Learning with Human Feedback

What is RLHF?

- Aligns models with human preferences using reinforcement learning.

How it Works

- Uses a preference model trained on human-labeled data (e.g., prompt-response pairs).
- Optimizes outputs to maximize rewards from the preference model.

Techniques Employed

- Reinforcement learning algorithm: Proximal Policy Optimization (PPO).

Key Benefits

- Produces high-quality, aligned, and safe responses.

Back to Taxonomy: Factuality Hallucination

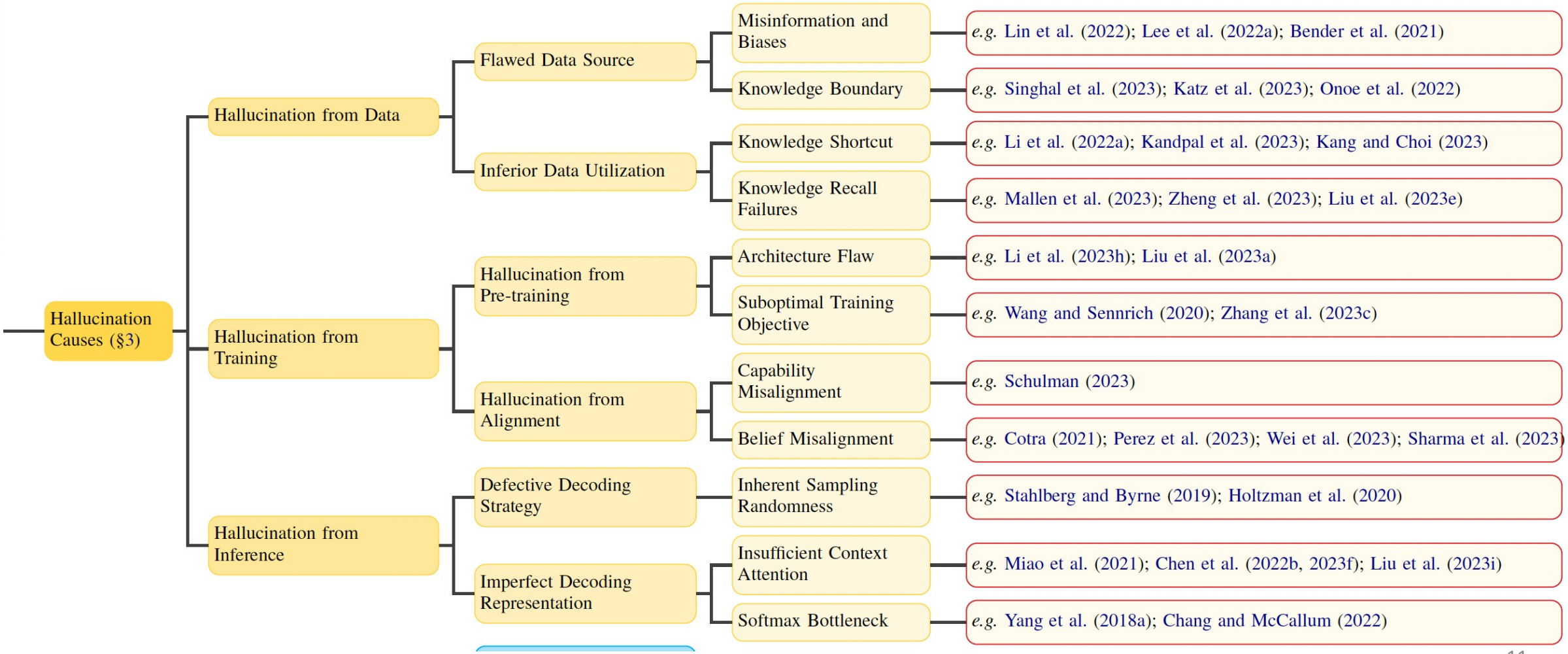
Hallucination Type	Sub-Type	User Input	Model Output	Explanation
Factuality Hallucination	Factual Inconsistency	Tell me about the first person to land on the Moon.	Yuri Gagarin was the first person to land on the Moon.	The LLM's response is factual inconsistency as **Neil Armstrong was the first person to land on the Moon**, not Yuri Gagarin, who was the first human in space.
	Factual Fabrication	Tell me about the historical origins of unicorns.	Unicorns were documented to have roamed the plains of Atlantis around 10,000 BC, where they were considered sacred creatures and were often associated with royalty.	The LLM's response is a fabricated claim since **there's no verified, real-world evidence supporting the existence of unicorns in Atlantis**, especially in connection with royalty.

Back to Taxonomy: Faithfulness Hallucination

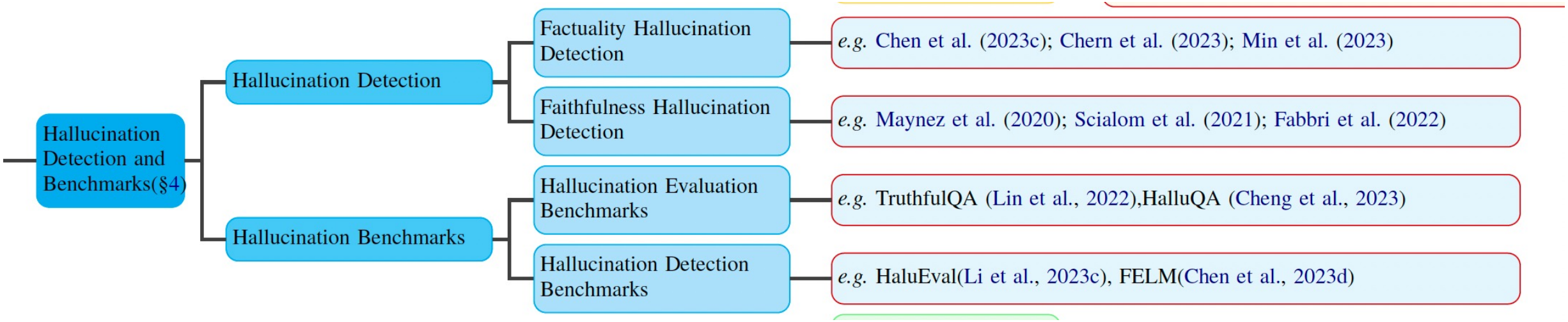
Faithfulness Hallucination

Instruction Inconsistency	Translate the following English question into Spanish: "What is the capital of France?"	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	Summarize the following paragraph about the Nile: 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.	The Nile originates from the mountain ranges of central Africa and flows into the Mediterranean.	The LLM's summary incorrectly states that the Nile originates from the "mountain ranges of central Africa," while **the paragraph clearly mentions that the Nile originates in the Great Lakes region of central Africa.**
Logical Inconsistency	Solve the equation 2x + 3 = 11 step by step.	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.

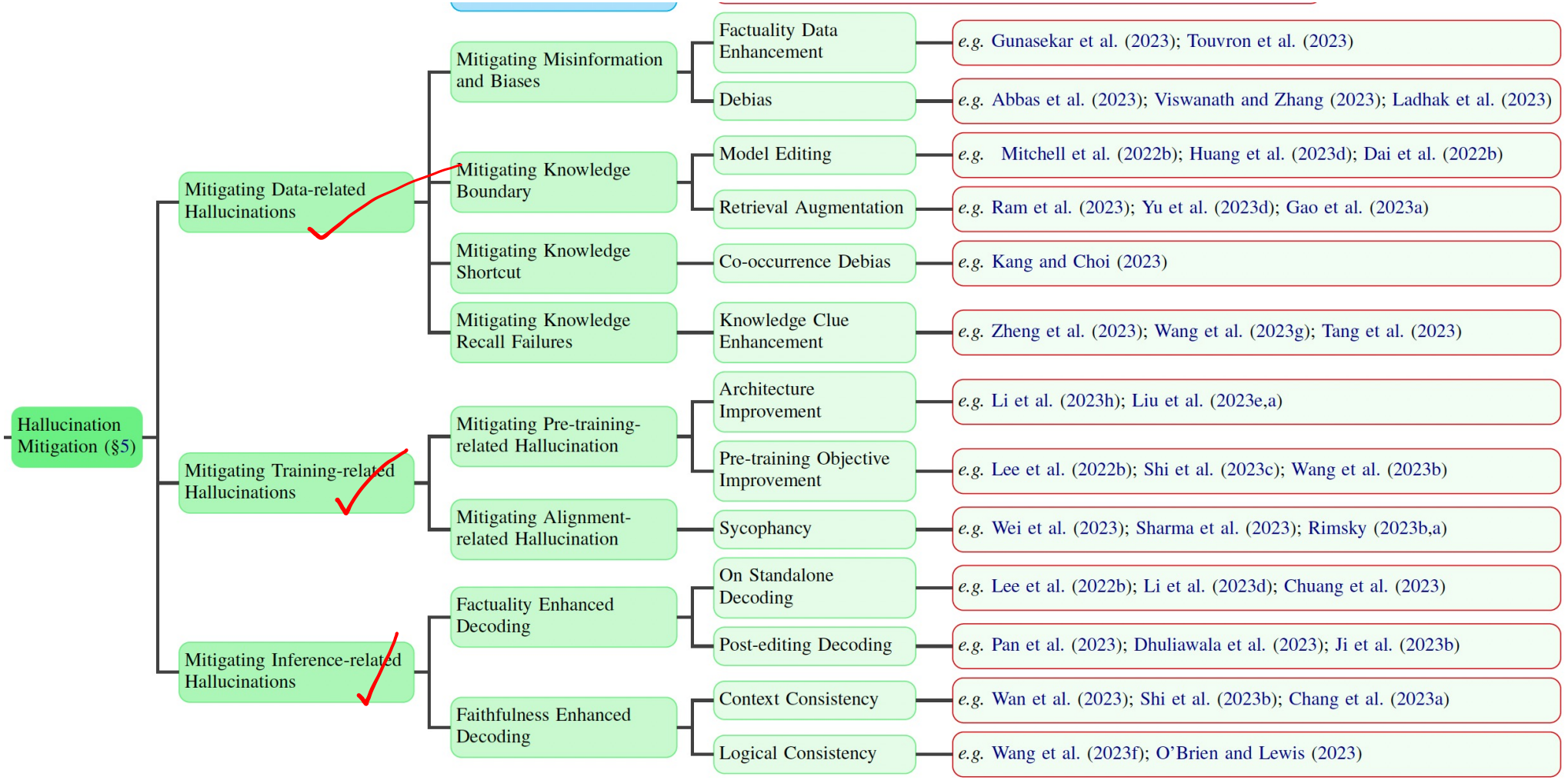
Causes


Hallucination Causes (§3)
Hallucination from Data
Flawed Data Source
Misinformation and Biases
e.g. Lin et al. (2022); Lee et al. (2022a); Bender et al. (2021)
Knowledge Boundary
e.g. Singhal et al. (2023); Katz et al. (2023); Onoe et al. (2022)
Inferior Data Utilization
Knowledge Shortcut
e.g. Li et al. (2022a); Kandpal et al. (2023); Kang and Choi (2023)
Knowledge Recall Failures
e.g. Mallen et al. (2023); Zheng et al. (2023); Liu et al. (2023e)
Hallucination from Training
Hallucination from Pre-training
Architecture Flaw
e.g. Li et al. (2023h); Liu et al. (2023a)
Suboptimal Training Objective
e.g. Wang and Sennrich (2020); Zhang et al. (2023c)
Hallucination from Alignment
Capability Misalignment
e.g. Schulman (2023)
Belief Misalignment
e.g. Cotra (2021); Perez et al. (2023); Wei et al. (2023); Sharma et al. (2023)
Hallucination from Inference
Defective Decoding Strategy
Inherent Sampling Randomness
e.g. Stahlberg and Byrne (2019); Holtzman et al. (2020)
Imperfect Decoding Representation
Insufficient Context Attention
e.g. Miao et al. (2021); Chen et al. (2022b, 2023f); Liu et al. (2023i)
Softmax Bottleneck
e.g. Yang et al. (2018a); Chang and McCallum (2022)

Detection


Hallucination Detection and Benchmarks(§4)
Hallucination Detection
Hallucination Benchmarks
Factuality Hallucination Detection
Faithfulness Hallucination Detection
Hallucination Evaluation Benchmarks
Hallucination Detection Benchmarks
e.g. Chen et al. (2023c); Chern et al. (2023); Min et al. (2023)
e.g. Maynez et al. (2020); Scialom et al. (2021); Fabbri et al. (2022)
e.g. TruthfulQA (Lin et al., 2022),HalluQA (Cheng et al., 2023)
e.g. HaluEval(Li et al., 2023c), FELM(Chen et al., 2023d)

Mitigation


Hallucination Mitigation (§5)
Mitigating Data-related Hallucinations
Mitigating Training-related Hallucinations
Mitigating Inference-related Hallucinations
Mitigating Misinformation and Biases
Mitigating Knowledge Boundary
Mitigating Knowledge Shortcut
Mitigating Knowledge Recall Failures
Mitigating Pre-training-related Hallucination
Mitigating Alignment-related Hallucination
Factuality Enhanced Decoding
Faithfulness Enhanced Decoding
Factuality Data Enhancement
Debias
Model Editing
Retrieval Augmentation
Co-occurrence Debias
Knowledge Clue Enhancement
Architecture Improvement
Pre-training Objective Improvement
Sycophancy
On Standalone Decoding
Post-editing Decoding
Context Consistency
Logical Consistency
e.g. Gunasekar et al. (2023); Touvron et al. (2023)
e.g. Abbas et al. (2023); Viswanath and Zhang (2023); Ladhak et al. (2023)
e.g. Mitchell et al. (2022b); Huang et al. (2023d); Dai et al. (2022b)
e.g. Ram et al. (2023); Yu et al. (2023d); Gao et al. (2023a)
e.g. Kang and Choi (2023)
e.g. Zheng et al. (2023); Wang et al. (2023g); Tang et al. (2023)
e.g. Li et al. (2023h); Liu et al. (2023e,a)
e.g. Lee et al. (2022b); Shi et al. (2023c); Wang et al. (2023b)
e.g. Wei et al. (2023); Sharma et al. (2023); Rimsky (2023b,a)
e.g. Lee et al. (2022b); Li et al. (2023d); Chuang et al. (2023)
e.g. Pan et al. (2023); Dhuliawala et al. (2023); Ji et al. (2023b)
e.g. Wan et al. (2023); Shi et al. (2023b); Chang et al. (2023a)
e.g. Wang et al. (2023f); O'Brien and Lewis (2023)

Some examples: Misinformation and Biases

Type	Premise	Input	Model Ouput	Explanation
Imitative Falsehood	Within the LLM's training data, a frequently referenced but incorrect fact appears multiple times: *Thomas Edison invented the light bulb.*	Who is credited with the invention of the light bulb?	Thomas Edison is credited with the invention of the light bulb.	**While Edison improved the light bulb's design, he was not its sole inventor.** The model's answer reflects a misconception prevalent in its training data, exemplifying an imitative falsehood.
Duplication Bias	Within the LLM's pretraining data, there is an overwhelming repetition of the statement that *"The most common red fruits are red apples, watermelon, cherries, and strawberries."*	Please list some red fruits, excluding apples.	Red fruits are red apples, watermelon, cherries, and strawberries.	**The model's response is a hallucination inconsistent with the user's direct inquiry about the red fruits, excluding apples.** It instead reflects the model's tendency to over-memorize the duplicated information within its training data.
Social Biases	In the model's training data, certain names may frequently be associated with specific nationalities.	Provide a summary of the following news: Dr. Kim recently published a breakthrough paper on quantum computing. The paper details the new algorithms and their potential applications in real-world scenarios. It has been hailed as revolutionary by experts in the field.	Dr. Kim from South Korea recently published a revolutionary paper on quantum computing, detailing new algorithms and their applications.	**The model erroneously added "South Korea" based on the name "Kim"** revealing its bias in linking certain names to specific nationalities, which can cause hallucinations in the summary.

Some examples: Knowledge Boundary

Type	Input	Model Ouput	Explanation
Domain Knowledge Deficiency	What causes phenylketonuria?	Phenylketonuria is caused by a mutation in the hemoglobin gene, which results in an unusual interaction with hemoglobin, affecting its ability to transport oxygen in the bloodstream.	Phenylketonuria (PKU) is actually caused by **mutations in the gene that helps make an enzyme called phenylalanine hydroxylase**. Due to the lack of specialized medical knowledge in LLM's training data, it may sometimes provide incorrect information on medical topics.
Outdated Knowledge	Which city hosted the most recent Winter Olympics?	The most recent city to host the Winter Olympics was Pyeongchang, South Korea, in 2018.	The most recent city to host the Winter Olympics was **Beijing, in 2022**. This is due to the fact that LLM stores outdated knowledge, and the answer to this question exhibits a time-shift phenomenon.

Table 3: Example of Knowledge Boundary.

Some Examples: Knowledge Shortcut and Knowledge Recall Failures

Type	Sub-Type	User Input	Model Output	Explanation
Knowledge Shortcut	Co-occurrence Bias	What is the capital of Canada?	Toronto is the capital of Canada.	The model leans heavily on the frequent co-occurrence of the terms *Toronto* and *Canada* in its training data, without truly capturing the factual knowledge about the capital of Canada
Knowledge Recall Failures	Long-tail Knowledge	Please generate a biography for George James Rankin.	George James Rankin is a dedicated educator known for his contributions to the field of education and his passion for fostering learning.	**George James Rankin is actually a politician**, but the LLM may have difficulty effectively utilizing knowledge about this long-tail entity despite being trained on comprehensive Wikipedia data during pre-training.
	Complex Reasoning	If Mount Everest were to descend by 500 meters, which mountain would become the world's highest peak?	If Mount Everest were to descend by 500 meters, it would still remain the world's highest peak.	The height of Mount Everest is 8844.43 meters, while K2's height is 8611 meters. If Mount Everest were to descend by 500 meters, K2 would become the world's highest peak. Facing complex multi-step reasoning questions like this, LLM may struggle to recall all the relevant knowledge associated with it.

Factuality Hallucination Detection Example:

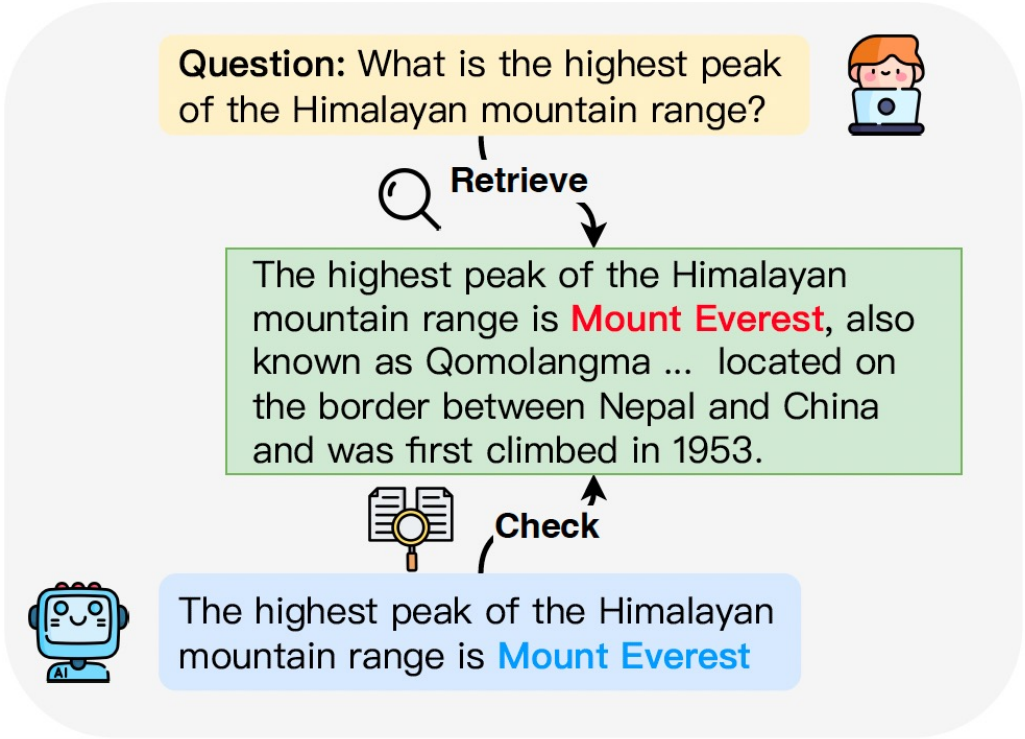


Figure 3: An example of detecting factuality hallucination by retrieving external facts.

https://abdullah-mamun.com
a.mamun@asu.edu