

# CSCSE 689: Special Topics in Trustworthy NLP

## Lecture 23: Hallucinations and Misinformation Control (1)

Kuan-Hao Huang  
khhuang@tamu.edu



# Course Project – Final Presentation

- 10-minute presentation and 2-minute Q&A
  - The topic you choose
  - An introduction to the task
  - Evaluation metrics
  - The dataset, models, and approaches you have worked with
  - Results and findings

# Course Project – Final Presentation

- Lottery drawing in class
  - 11/20: Vicram Rajagopala, Sriram Balasubramanian, Rahul Baid
  - 11/22: Dylan Harden, Divij Bajaj, Arunim Chaitanya Samudra
  - 11/25: Chan-Wei Hu & Junru Liu, Saransh Agrawal, Jaehoon Lee

# Hallucinations and Misinformation Control



Assessment Learning Technologies

## Why does ChatGPT generate fake references?

Uncovering the ingredients that go into robotic porkie pies and some possible sweet dessert at the end.

TECHÉ Mathew Hillier on 20 February, 2023



## What is an AI Hallucination?

An AI hallucination is a false or misleading output generated by an artificial intelligence system. It can be a confident response by an AI that does not seem to be justified by its training data.

**Example:**  
When asked about Microsoft's annual revenue in 2021, a hallucinating chatbot falsely states that the figure is **\$10 billion** (the right answer should have been **\$168.1 billion**).

# What is Hallucination?

- The generation of content that is irrelevant, made-up, or inconsistent with the input data

# What is Hallucination?



Who was the first person to walk on the moon?



**Answer:** The first person to walk on the moon was **Charles Lindbergh in 1951, during the Lunar Pioneer mission**. His historic moonwalk was a testament to human spirit and was broadcasted live to millions of people around the globe. ❌



**Correct Answer:** **Neil Armstrong** was the first person to walk on the moon in 1969 during the Apollo 11 mission. ✅

(a) Factuality Hallucination



Please summarize the following news article:

**Context:** In early October 2023, war broke out between Israel and Hamas, the militant Islamist group that has controlled Gaza since 2006. Hamas fighters fired rockets ... civilians and taking dozens of hostages.



**Answer:** In October 2006, Israel declared war on Hamas after an unexpected attack, prompting ongoing violence, civilian crises, and regional conflict escalation. ❌

(b) Faithfulness Hallucination

# Factuality Hallucination

Sub-Type	User Input	Model Output	Explanation
Factual Inconsistency	Tell me about the first person to land on the Moon.	<b>Yuri Gagarin</b> was the first person to land on the Moon	The LLM's response is factual inconsistency as <b>Neil Armstrong was the first person to land on the Moon</b> , not Yuri Gagarin, who was the first human in space
Factual Fabrication	Tell me about the historical origins of unicorns	<b>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.</b>	The LLM's response is a fabricated claim since <b>there's no verified, real-world evidence supporting the existence of unicorns in Atlantis</b> , especially in connection with royalty

# Faithfulness Hallucination

Sub-Type	User Input	Model Output	Explanation
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. <b>Instead of performing the translation as instructed, the LLM chose to answer the question.</b>
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 <b>the paragraph clearly mentions that the Nile originates in the Great The Lakes region of central Africa.</b>
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, <b>the next step inaccurately divides 8 by 2 to yield a result of <math>x = 3</math></b> , which is inconsistent with the earlier reasoning.

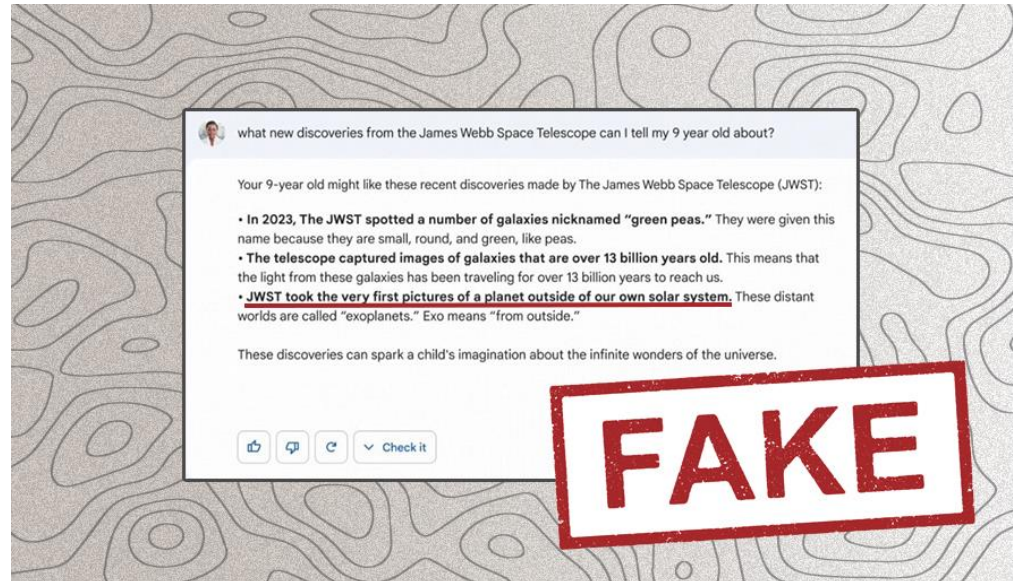


# Causes of Hallucinations in LLMs

- Training data issues
- Inference stage challenges
- Prompt Engineering

# Training Data Issues

- Verifying data's fairness, unbiasedness, and factual correctness is challenging
  - Models cannot distinguish between truth and fiction
- Internet-sourced datasets may include biased or incorrect information
  - propagate into the model's outputs



# Inference Stage Challenges

- Defective decoding strategies
- Inherent randomness in the sampling methods

# Inputs and Prompt Engineering

- LLMs may generate hallucinated content when faced with unclear or imprecise input
- LLMs might generate an incorrect or unrelated answer if a prompt lacks adequate context or is ambiguously worded

# **FACTSCORE: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation**

**Sewon Min<sup>†1</sup> Kalpesh Krishna<sup>†2</sup> Xinxi Lyu<sup>1</sup> Mike Lewis<sup>4</sup> Wen-tau Yih<sup>4</sup>  
Pang Wei Koh<sup>1</sup> Mohit Iyyer<sup>2</sup> Luke Zettlemoyer<sup>1,4</sup> Hannaneh Hajishirzi<sup>1,3</sup>**

<sup>1</sup>University of Washington    <sup>2</sup>University of Massachusetts Amherst

<sup>3</sup>Allen Institute for AI    <sup>4</sup>Meta AI

{sewon,alrope,pangwei,lsz,hannaneh}@cs.washington.edu

{kalpesh,miyyer}@cs.umass.edu    {mikelewis,scottyih}@meta.com

# Types of Hallucinations



Who was the first person to walk on the moon?



**Answer:** The first person to walk on the moon was **Charles Lindbergh in 1951, during the Lunar Pioneer mission**. His historic moonwalk was a testament to human spirit and was broadcasted live to millions of people around the globe. ❌



**Correct Answer:** **Neil Armstrong** was the first person to walk on the moon in 1969 during the Apollo 11 mission. ✅

(a) Factuality Hallucination



Please summarize the following news article:

**Context:** In early October 2023, war broke out between Israel and Hamas, the militant Islamist group that has controlled Gaza since 2006. Hamas fighters fired rockets ... civilians and taking dozens of hostages.



**Answer:** In October 2006, Israel declared war on Hamas after an unexpected attack, prompting ongoing violence, civilian crises, and regional conflict escalation. ❌

(b) Faithfulness Hallucination

# FActScore: Factual Precision in Atomicity Score

Bridget Moynahan is an American actress, model and producer. She is best known for her roles in Grey's Anatomy, I, Robot and Blue Bloods. She studied acting at the American Academy of Dramatic Arts, and ...

## Atomic facts



- Bridget Moynahan is American.
- Bridget Moynahan is an actress.
- Bridget Moynahan is a model.
- Bridget Moynahan is a producer.
- She is best known for her roles in Grey's Anatomy.
- She is best known for her roles in I, Robot.
- She is best known for her roles in Blue Bloods.
- She studied acting.
- She studied at the American Academy of Dramatic Arts.
- ...

# FActScore: Factual Precision in Atomicity Score

Bridget Moynahan is an American actress, model and producer. She is best known for her roles in Grey's Anatomy, I, Robot and Blue Bloods. She studied acting at the American Academy of Dramatic Arts, and ...

## Atomic facts

- Bridget Moynahan is American. ✓
- Bridget Moynahan is an actress. ✓
- Bridget Moynahan is a model. ✓
- Bridget Moynahan is a producer. ✗
- She is best known for her roles in Grey's Anatomy. ✗
- She is best known for her roles in I, Robot. ✓
- She is best known for her roles in Blue Bloods. ✓
- She studied acting. ✓
- She studied at the American Academy of Dramatic Arts. ✗
- ...



**66.7%**





# FActScore: Factual Precision in Atomicity Score

Bridget Moynahan is an American filmmaker and writer. She is best known for her work on the soap opera General Hospital, which she co-created with husband Charles Kelly. Moynahan was raised in a middle-class family in Los Angeles, ...

## Atomic facts

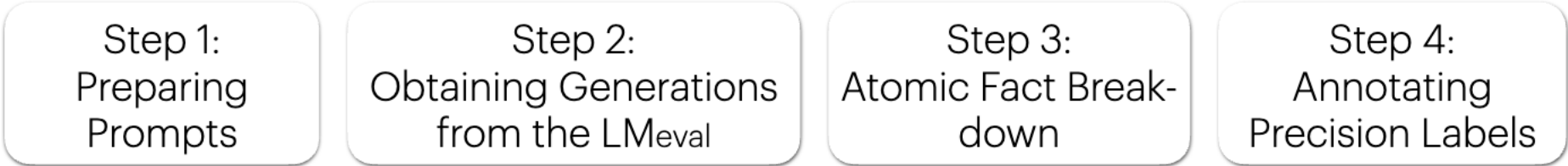
- Bridget Moynahan is American. ✓
- Bridget Moynahan is a filmmaker. ✗
- Bridget Moynahan is a writer. ✗
- She is best known for her work on General Hospital. ✗
- General Hospital is the soap opera. ✗
- She co-created General Hospital. ✗
- She co-created General Hospital with her husband. ✗
- Her husband is Charles Kelly. ✗
- Moynahan was raised in a middle-class family. ✗
- Moynahan was raised in Los Angeles. ✗
- ...



**10.0%**



# FActScore: Factual Precision in Atomicity Score

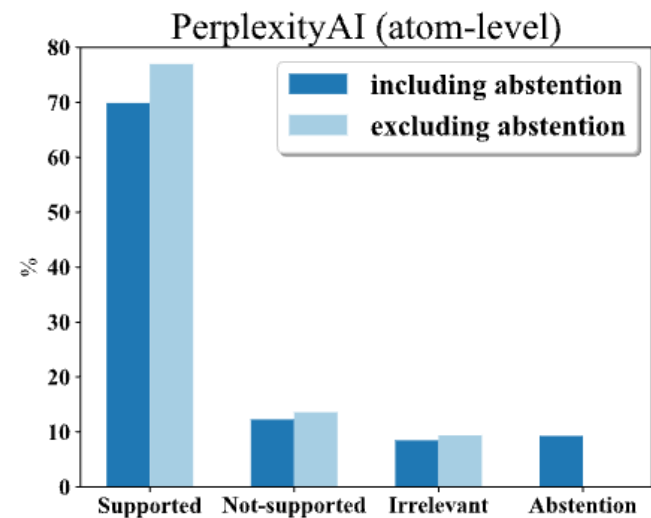
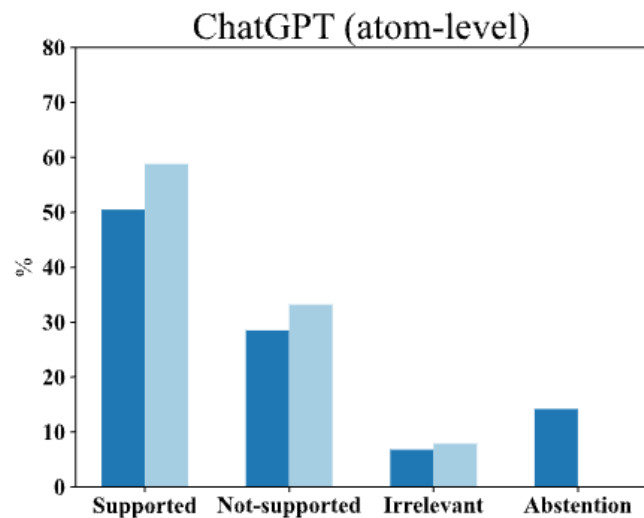
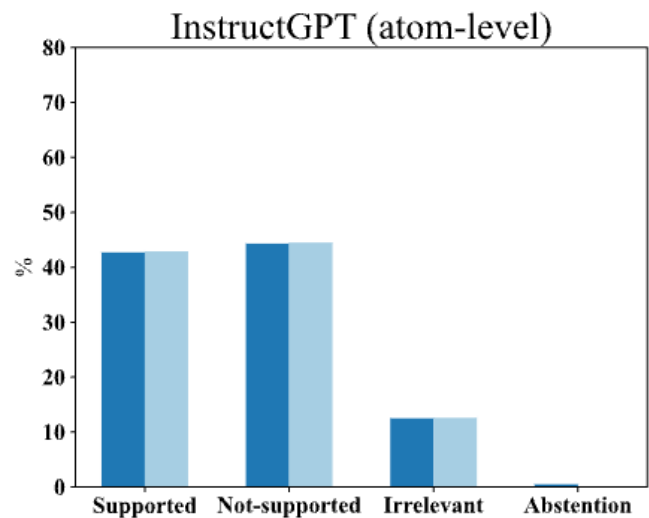


Bridget Moynahan is an American.  
Bridget Moynahan is an actress.  
Bridget Moynahan is a model.  
Bridget Moynahan is a producer.  
...  
She studied acting at the American Academy of Dramatic Arts.  
She began her career in the late 1990s.



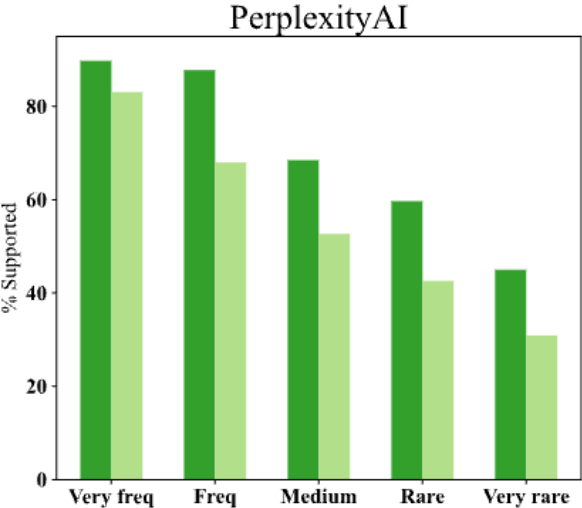
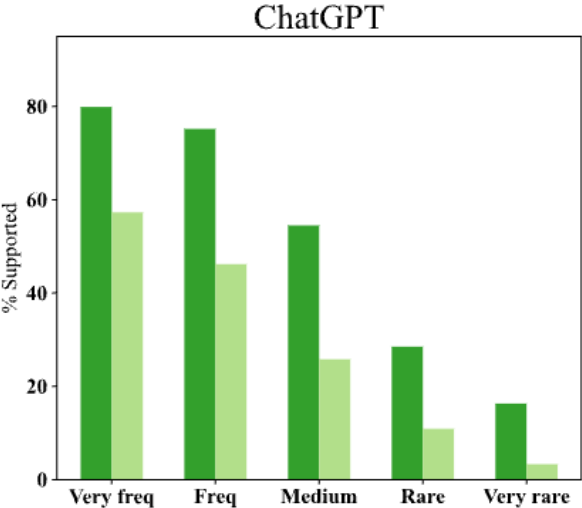
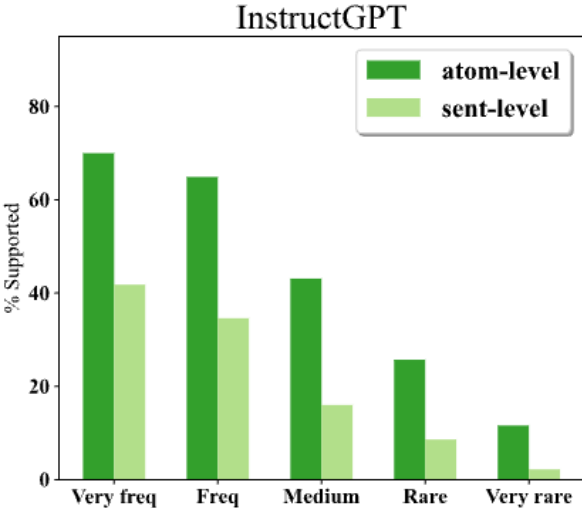
- Irrelevant
- Supported
- Not-supported

# Results



All models make substantial precision errors

# Results



More errors as the rarity of entities increases

# How to Mitigate Hallucinations?

- Some general suggestions
  - **Simplify complex tasks:** Break down intricate actions into simpler steps.
  - **Harness affordances:** Utilize built-in functions within your metaprompt.
  - **Use few-shot learning:** Include examples when you can.
  - **Iterative refinement:** Don't hesitate to tweak the model's output.

# How Language Model Hallucinations Can Snowball

**Muru Zhang**<sup>♡</sup> **Ofir Press**<sup>♡</sup> **William Merrill**<sup>♠</sup> **Alisa Liu**<sup>♡</sup> **Noah A. Smith**<sup>♡♣</sup>

<sup>♡</sup>Paul G. Allen School of Computer Science and Engineering, University of Washington

<sup>♠</sup>New York University

<sup>♣</sup>Allen Institute for Artificial Intelligence

nanami17@cs.washington.edu

# Types of Hallucinations



Who was the first person to walk on the moon?



**Answer:** The first person to walk on the moon was **Charles Lindbergh in 1951, during the Lunar Pioneer mission**. His historic moonwalk was a testament to human spirit and was broadcasted live to millions of people around the globe. ❌



**Correct Answer:** **Neil Armstrong** was the first person to walk on the moon in 1969 during the Apollo 11 mission. ✅

(a) Factuality Hallucination



Please summarize the following news article:




**Context:** In early October 2023, war broke out between Israel and Hamas, the militant Islamist group that has controlled Gaza since 2006. Hamas fighters fired rockets ... civilians and taking dozens of hostages.


**Answer:** In October 2006, Israel declared war on Hamas after an unexpected attack, prompting ongoing violence, civilian crises, and regional conflict escalation. ❌

(b) Faithfulness Hallucination


# Hallucination Snowballing


 Is 9677 a prime number?

---


 No, 9677 is not a prime number.  
It can be factored into 13 and 745, as  $9677 = 13 \times 745$ .

} incorrect assertion  
} snowballed hallucination




 Is 9677 divisible by 13?

---




 No

in a separate session, GPT-4 recognizes its claim as incorrect!





# Verification

Dataset	Original Question	Verification Question
 <b>Primality Testing</b>	<p><b>User:</b> Is 10733 a prime number?</p> <p><b>GPT-4:</b> No... It can be <u>factored into <math>3 \times 3577</math></u>.</p>	<p><b>User:</b> Is 10733 divisible by 3? Answer with either Yes or No.</p> <p><b>GPT-4:</b> <u>No</u></p>
 <b>Senator Search</b>	<p><b>User:</b> Was there ever a US senator that represented the state of New Hampshire and whose alma mater was the University of Pennsylvania?</p> <p><b>GPT-4:</b> Yes... His name was <u>John P. Hale</u></p>	<p><b>User:</b> Was John P. Hale's alma mater University of Pennsylvania?</p> <p><b>GPT-4:</b> <u>No</u>... [it] was Bowdoin</p>
 <b>Graph Connectivity</b>	<p><b>User:</b> Current flight information (the following flights are one-way only, and all the flights available are included below): There is a flight from city F to city K There is a flight from city H to city A [... 10 other rules cut for space ...] Question: Is there a series of flights that goes from city B to city E?</p> <p><b>GPT-4:</b> Yes... the route is as follows: ... <u>City K to City G</u>...</p>	<p><b>User:</b> [...flight information given in the context...] Based on the above flight information, is City K to City G a valid flight?</p> <p><b>GPT-4:</b> <u>No</u>, based on the above flight information, there is no direct flight from City K to City G.</p>

# Results

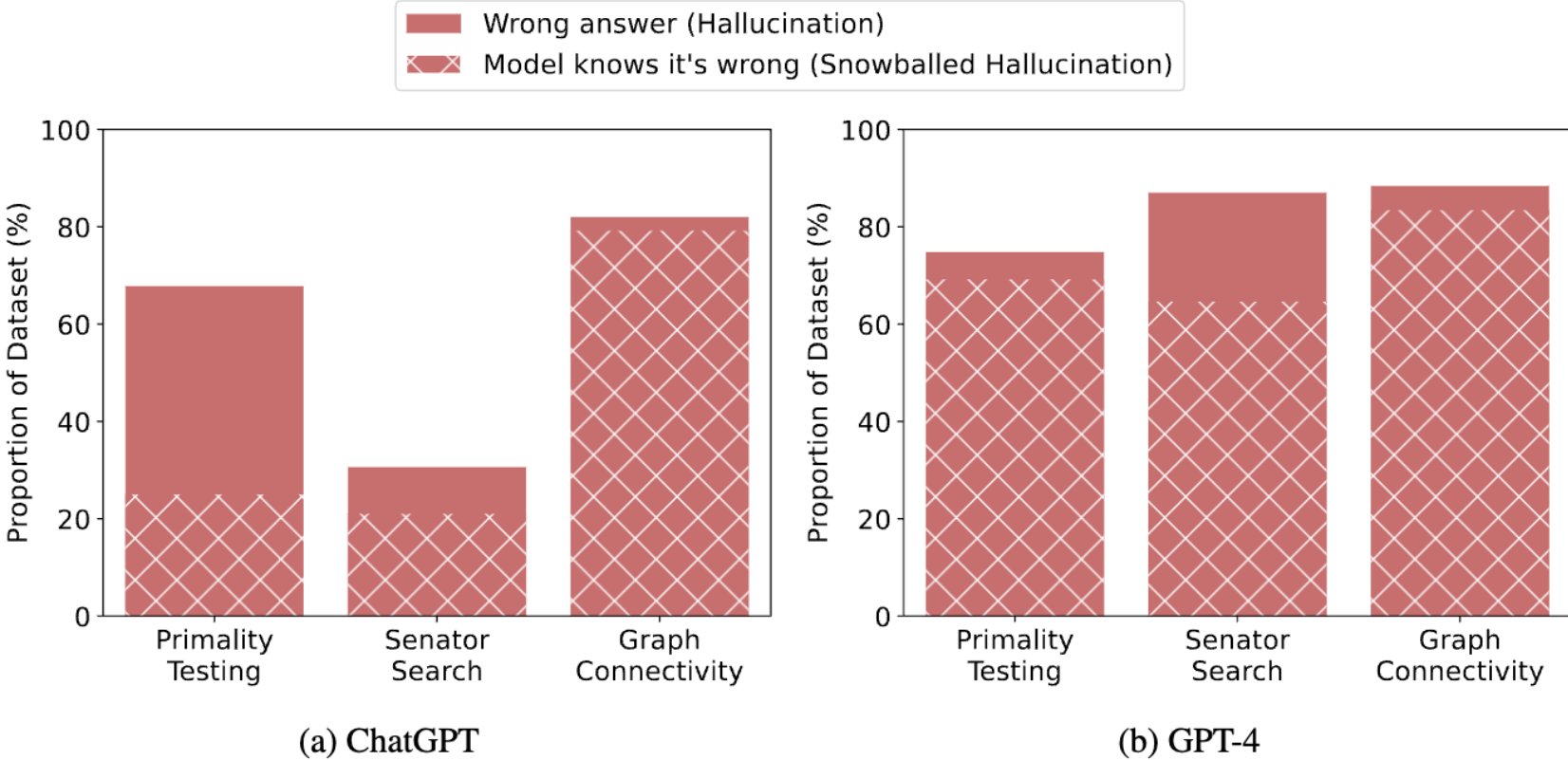


Figure 2: Percentage of hallucination and percentage of snowballed hallucination (both calculated with respect to the entire dataset) for ChatGPT and GPT-4. The precise numbers for this plot are available in [Table 6](#) and [Table 7](#) in the Appendix.

# Chain-of-Thought Helps!

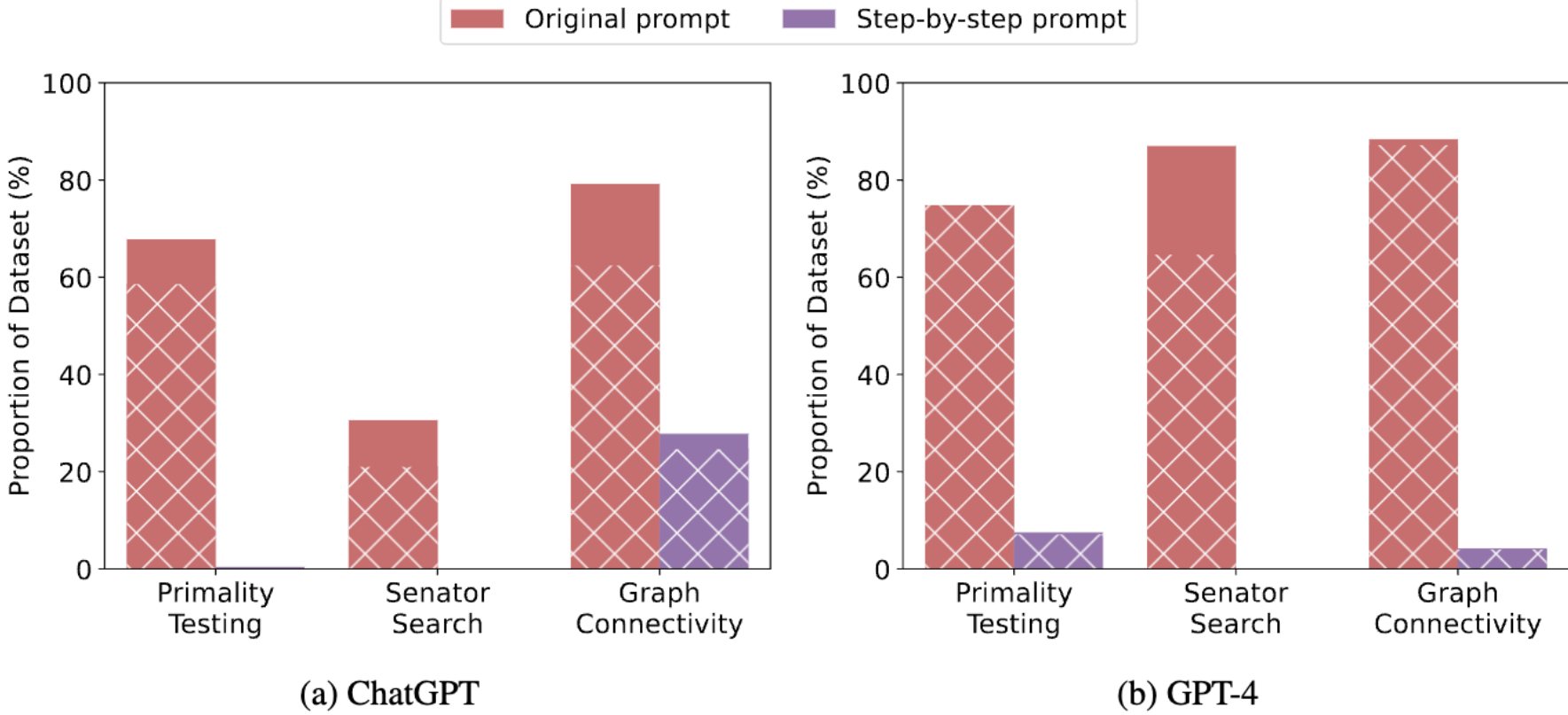


Figure 3: Error rate and snowballed hallucination rate (hatch pattern) for ChatGPT and GPT-4, when using the original prompt versus “Let’s think step by step”. See Appendix Table 8 and Table 9 for the exact numbers.

# **SELF CHECK GPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models**

**Potsawee Manakul, Adian Liusie, Mark J. F. Gales**

ALTA Institute, Department of Engineering, University of Cambridge

pm574@cam.ac.uk, a1826@cam.ac.uk, mjfg@eng.cam.ac.uk

# Types of Hallucinations



Who was the first person to walk on the moon?



**Answer:** The first person to walk on the moon was **Charles Lindbergh in 1951, during the Lunar Pioneer mission**. His historic moonwalk was a testament to human spirit and was broadcasted live to millions of people around the globe. ❌



**Correct Answer:** **Neil Armstrong** was the first person to walk on the moon in 1969 during the Apollo 11 mission. ✅

(a) Factuality Hallucination



Please summarize the following news article:

**Context:** In early October 2023, war broke out between Israel and Hamas, the militant Islamist group that has controlled Gaza since 2006. Hamas fighters fired rockets ... civilians and taking dozens of hostages.



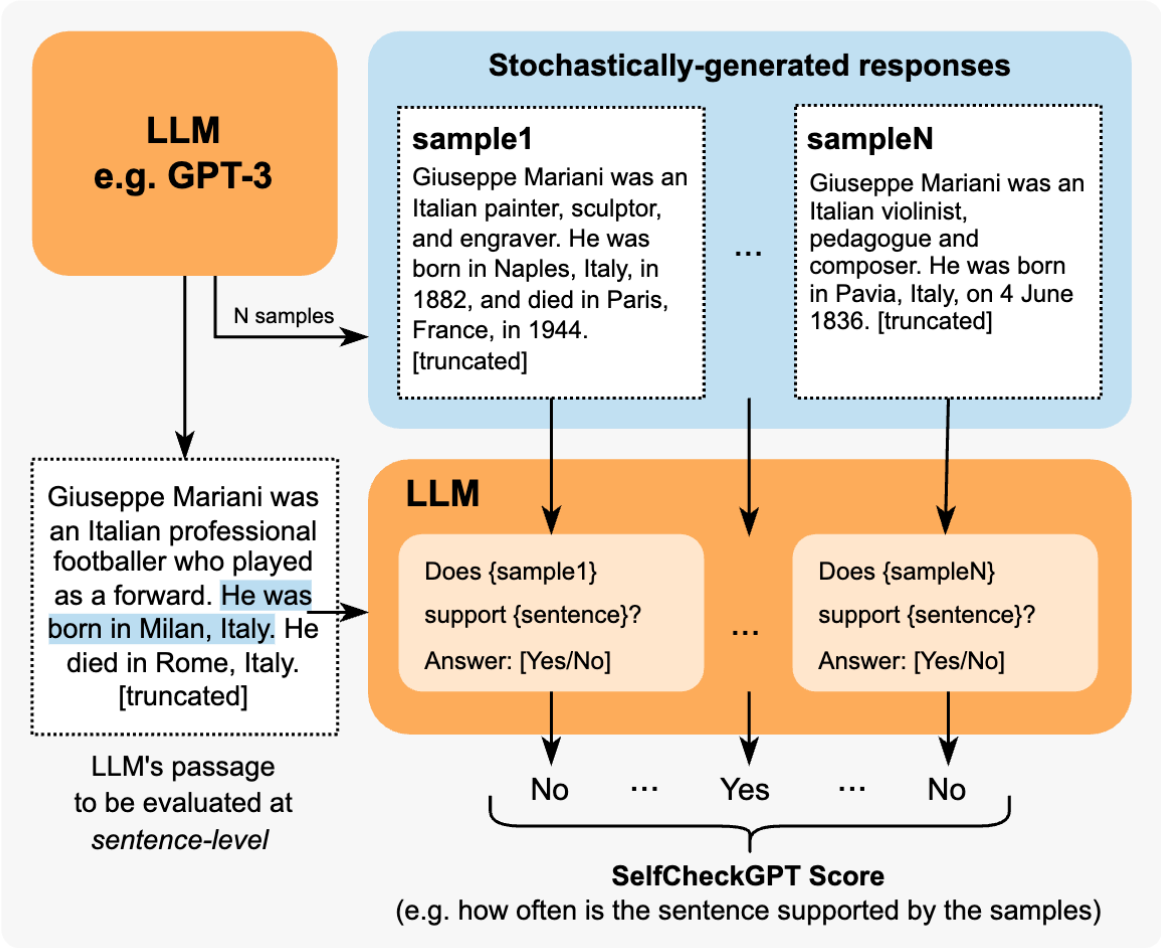
**Answer:** In October 2006, Israel declared war on Hamas after an unexpected attack, prompting ongoing violence, civilian crises, and regional conflict escalation. ❌

(b) Faithfulness Hallucination

# LLMs Can Verify

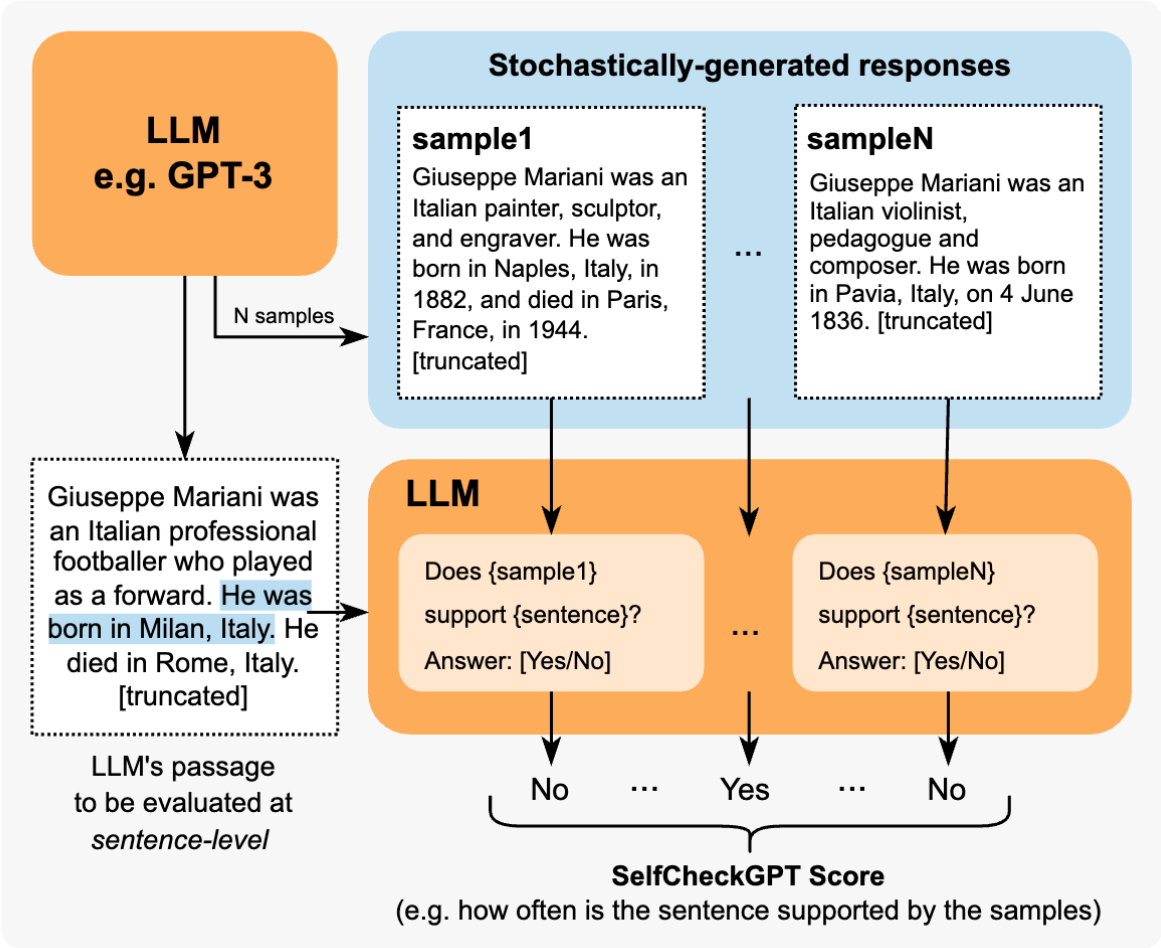
- If an LLM has knowledge of a given concept, sampled responses are likely to be similar and contain consistent facts

# SelfCheckGPT with BERTScore



$$S_{\text{BERT}}(i) = 1 - \frac{1}{N} \sum_{n=1}^N \max_k (\mathcal{B}(r_i, s_k^n))$$

# SelfCheckGPT with Prompt



-----

Context: {}  
Sentence: {}  
Is the sentence supported by the context above?  
Answer Yes or No:

-----

$$S_{\text{Prompt}}(i) = \frac{1}{N} \sum_{n=1}^N x_i^n$$



# Results

Method	Sentence-level (AUC-PR)			Passage-level (Corr.)	
	NonFact	NonFact*	Factual	Pearson	Spearman
Random	72.96	29.72	27.04	-	-
GPT-3 (text-davinci-003)'s probabilities ( <i>LLM, grey-box</i> )					
Avg( $-\log p$ )	83.21	38.89	53.97	57.04	53.93
Avg( $\mathcal{H}$ ) <sup>†</sup>	80.73	37.09	52.07	55.52	50.87
Max( $-\log p$ )	87.51	35.88	50.46	57.83	55.69
Max( $\mathcal{H}$ ) <sup>†</sup>	85.75	32.43	50.27	52.48	49.55
LLaMA-30B's probabilities ( <i>Proxy LLM, black-box</i> )					
Avg( $-\log p$ )	75.43	30.32	41.29	21.72	20.20
Avg( $\mathcal{H}$ )	80.80	39.01	42.97	33.80	39.49
Max( $-\log p$ )	74.01	27.14	31.08	-22.83	-22.71
Max( $\mathcal{H}$ )	80.92	37.32	37.90	35.57	38.94
<b>SelfCheckGPT</b> ( <i>black-box</i> )					
w/ BERTScore	81.96	45.96	44.23	58.18	55.90
w/ QA	84.26	40.06	48.14	61.07	59.29
w/ Unigram (max)	85.63	41.04	58.47	64.71	64.91
w/ NLI	92.50	45.17	66.08	74.14	73.78
w/ Prompt	<b>93.42</b>	<b>53.19</b>	<b>67.09</b>	<b>78.32</b>	<b>78.30</b>

---

# Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks

---

**Patrick Lewis<sup>†‡</sup>, Ethan Perez<sup>\*</sup>,**

**Aleksandra Piktus<sup>†</sup>, Fabio Petroni<sup>†</sup>, Vladimir Karpukhin<sup>†</sup>, Naman Goyal<sup>†</sup>, Heinrich Küttler<sup>†</sup>,**

**Mike Lewis<sup>†</sup>, Wen-tau Yih<sup>†</sup>, Tim Rocktäschel<sup>†‡</sup>, Sebastian Riedel<sup>†‡</sup>, Douwe Kiela<sup>†</sup>**

<sup>†</sup>Facebook AI Research; <sup>‡</sup>University College London; <sup>\*</sup>New York University;  
plewis@fb.com

# Types of Hallucinations



Who was the first person to walk on the moon?



**Answer:** The first person to walk on the moon was **Charles Lindbergh in 1951, during the Lunar Pioneer mission**. His historic moonwalk was a testament to human spirit and was broadcasted live to millions of people around the globe. ❌



**Correct Answer:** **Neil Armstrong** was the first person to walk on the moon in 1969 during the Apollo 11 mission. ✅

(a) Factuality Hallucination



Please summarize the following news article:



**Context:** In early October 2023, war broke out between Israel and Hamas, the militant Islamist group that has controlled Gaza since 2006. Hamas fighters fired rockets ... civilians and taking dozens of hostages.

**Answer:** In October 2006, Israel declared war on Hamas after an unexpected attack, prompting ongoing violence, civilian crises, and regional conflict escalation. ❌

(b) Faithfulness Hallucination

# Retrieval-Augmented Generation

