CSCE 638 Natural Language Processing Foundation and Techniques

Lecture 20: Hallucinations

Kuan-Hao Huang

Spring 2025

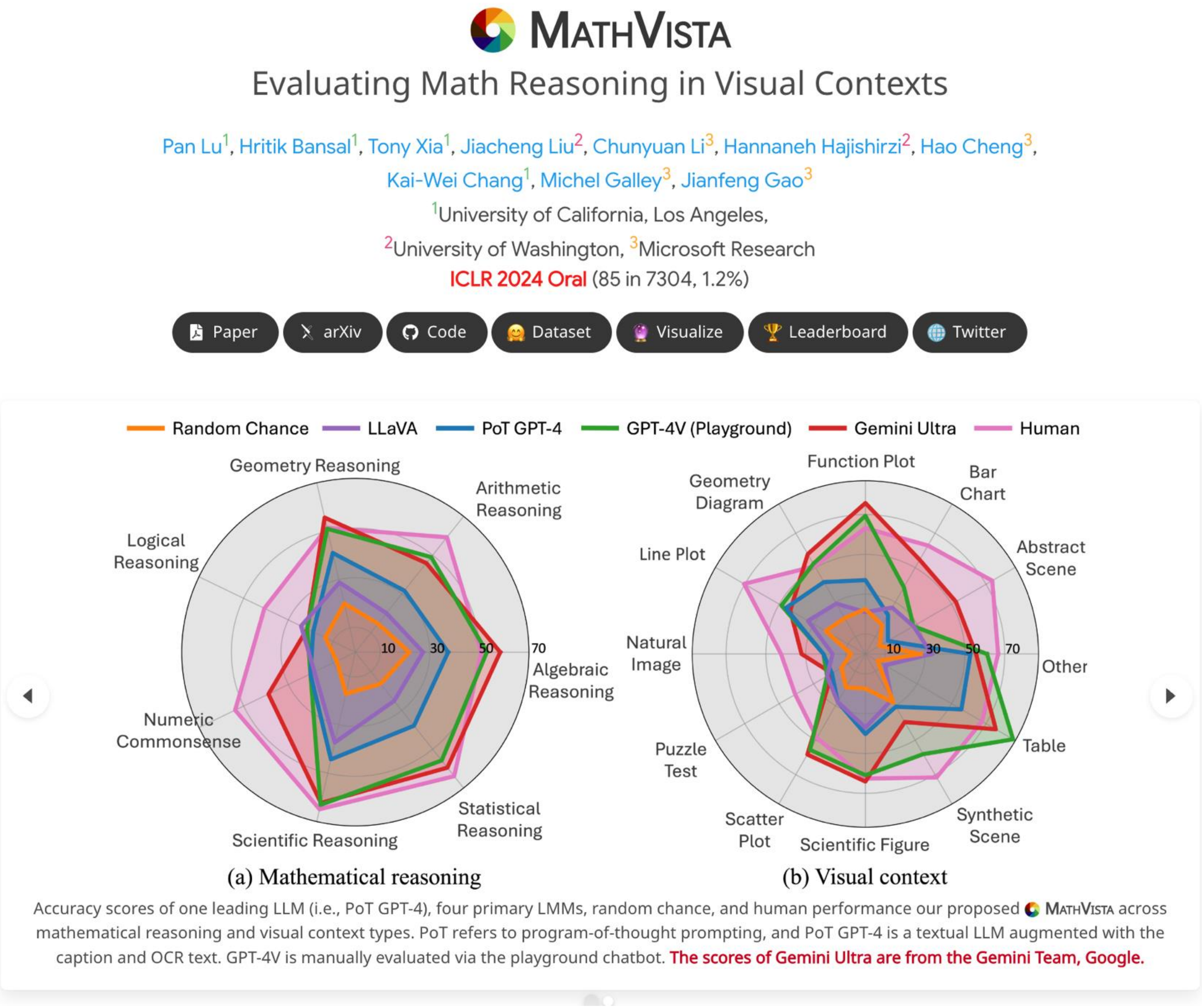
Invited Talk

- **Speaker:** Pan Lu, Postdoctoral Scholar at Stanford University
- **Title:** Advancing Complex Reasoning with Language Models and Agentic Systems
- **Date:** 4/16
- Online @ Zoom:
  - https://tamu.zoom.us/my/khhuang?pwd=oAdWOKVOCGPApqDbJnVtktdW2AE6nb.1

Invited Talk

Abstract: Complex reasoning is fundamental to human intelligence and plays a crucial role in advancing education, science, and technology. This talk explores the development of language model systems that exhibit robust mathematical reasoning and facilitate scientific reasoning, marking a significant step toward general artificial intelligence. We introduce novel multi-modal and knowledge-intensive benchmarks designed to assess the reasoning capabilities of large language models (LLMs) and vision-language models (VLMs) in real-world scenarios, including those involving visual data, tabular information, and scientific applications. The talk highlights recent advancements in mathematical reasoning within visual contexts and addresses key unresolved challenges. Additionally, we present cutting-edge retrieval and tool-augmented algorithms that significantly enhance LLM performance in mathematical reasoning tasks. Finally, we explore how agentic systems, leveraging test-time optimization and external tools, can further advance mathematical reasoning and scientific discovery.

Invited Talk



MATHVISTA

Evaluating Math Reasoning in Visual Contexts

Pan Lu[1], Hritik Bansal[1], Tony Xia[1], Jiacheng Liu[2], Chunyuan Li[3], Hannaneh Hajishirzi[2], Hao Cheng[3], Kai-Wei Chang[1], Michel Galley[3], Jianfeng Gao[3]

[1]University of California, Los Angeles,
[2]University of Washington, [3]Microsoft Research

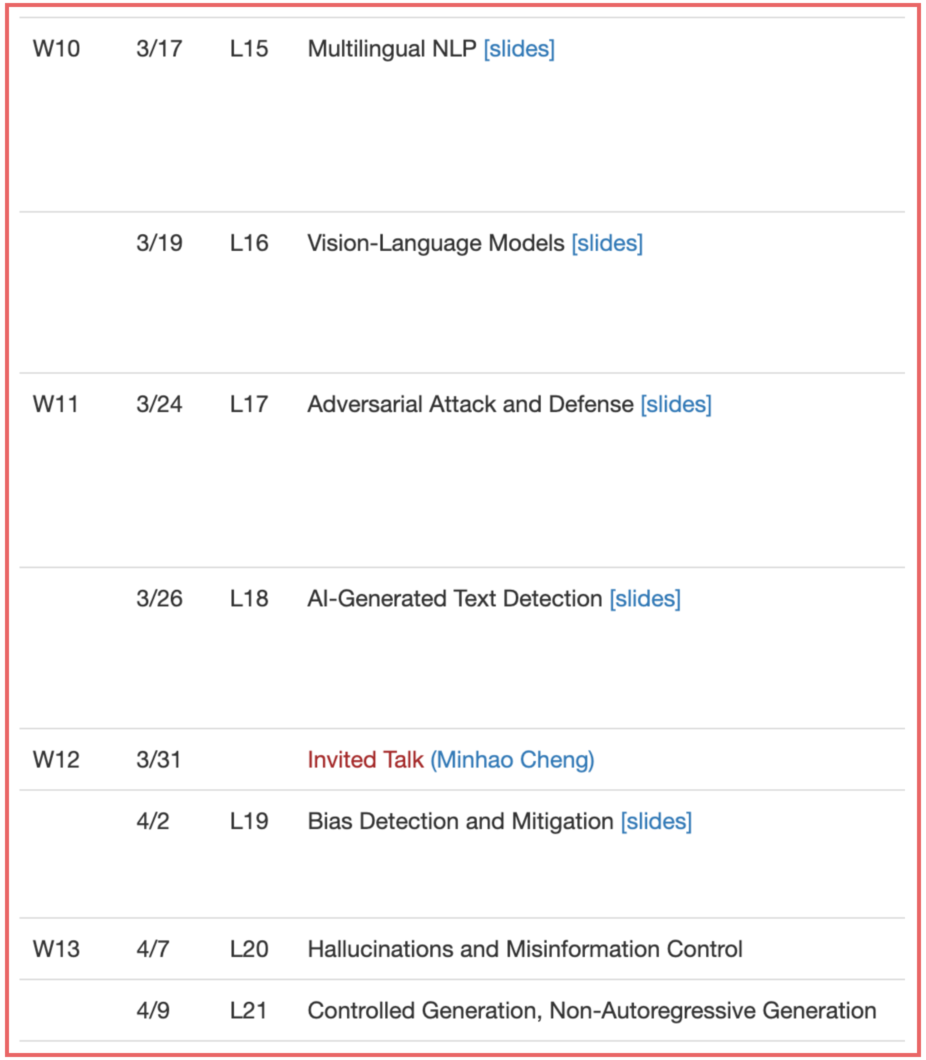
ICLR 2024 Oral (85 in 7304, 1.2%)

(a) Mathematical reasoning (b) Visual context

Accuracy scores of one leading LLM (i.e., PoT GPT-4), four primary LMMs, random chance, and human performance our proposed MATHVISTA across mathematical reasoning and visual context types. PoT refers to program-of-thought prompting, and PoT GPT-4 is a textual LLM augmented with the caption and OCR text. GPT-4V is manually evaluated via the playground chatbot. The scores of Gemini Ultra are from the Gemini Team, Google.

Quiz 3

- Date: 4/14
  - 15 minutes before the end of the lecture
  - 5 questions focusing on high-level concepts
- Pay attention to how methods work and the difference between them



W10	3/17	L15	Multilingual NLP [slides]
	3/19	L16	Vision-Language Models [slides]
W11	3/24	L17	Adversarial Attack and Defense [slides]
	3/26	L18	AI-Generated Text Detection [slides]
W12	3/31		Invited Talk (Minhao Cheng)
	4/2	L19	Bias Detection and Mitigation [slides]
W13	4/7	L20	Hallucinations and Misinformation Control
	4/9	L21	Controlled Generation, Non-Autoregressive Generation

Course Project – Final Report

- Due: 4/30
- Page limit: 9 pages (references are not counted for page limit)
- The report should include
  - Introduction to the topic you choose
  - Related literature
  - Novelty and challenges
  - Your approach
  - Experimental settings
  - Results, findings, and insights
  - Future directions
- A .zip file containing all the code

Course Project – Final Report

- Introduce all the components of the proposed approaches
  - Your classmates should be able to easily understand what you have done by reading it
  - If your method is based on previous work A, introduce work A a little bit
  - If you collect data by your own, clearly describe the format and how they are being collected
- All contributions need to be grounded in evidence
  - [X] We implement XXX and it has promising performance
  - [O] We implement XXX and it has promising performance, as shown in Figure Y and Table Z
- Properly cite everything

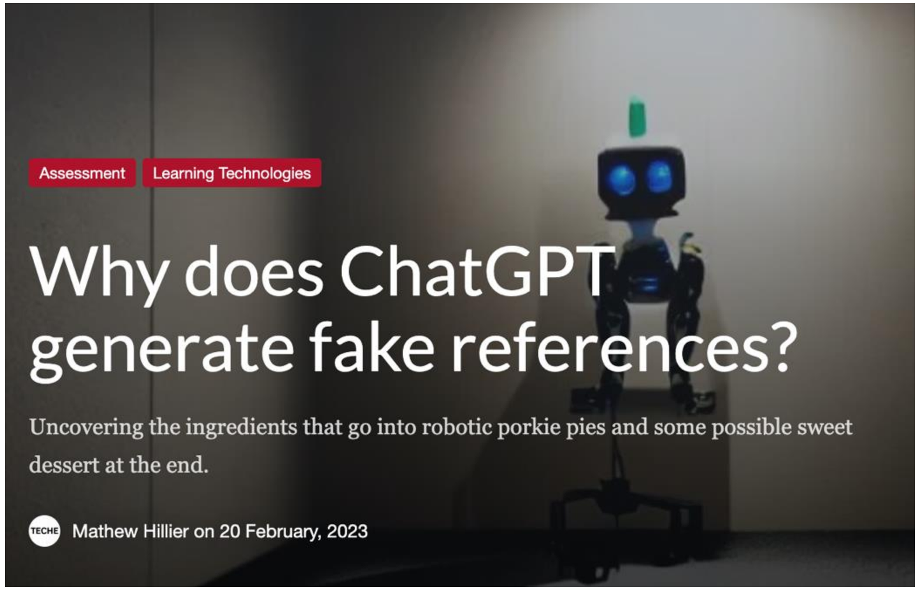
Final Presentation

- Each team has 7 minutes for presentation
  - You have to stop once you reach 7 minutes
- The presentation should include
  - The topic you choose
  - Novelty and challenges compared to previous literature
  - Your approach
  - Experimental settings
  - Results, findings, and insights
- Your classmates should be able to easily understand what you have done

Final Presentation

Date	Order 1	Order 2	Order 3	Order 4	Order 5	Order 6	Order 7	Order 8	Order 9
4/21	Team 2	Team 10	Team 4	Team 9	Team 26	Team 18	Team 1	Team 24	Team 12
4/23	Team 25	Team 14	Team 11	Team 23	Team 13	Team 5	Team 16	Team 15	Team 21
4/28	Team 8	Team 20	Team 19	Team 17	Team 27	Team 7	Team 6	Team 3	Team 22

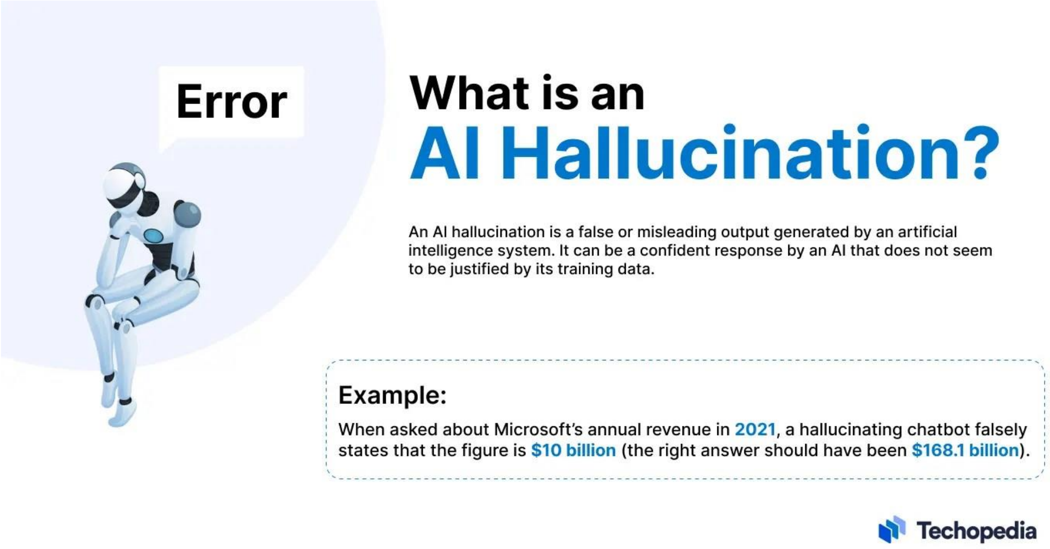
Hallucinations



Why does ChatGPT generate fake references?

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

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:

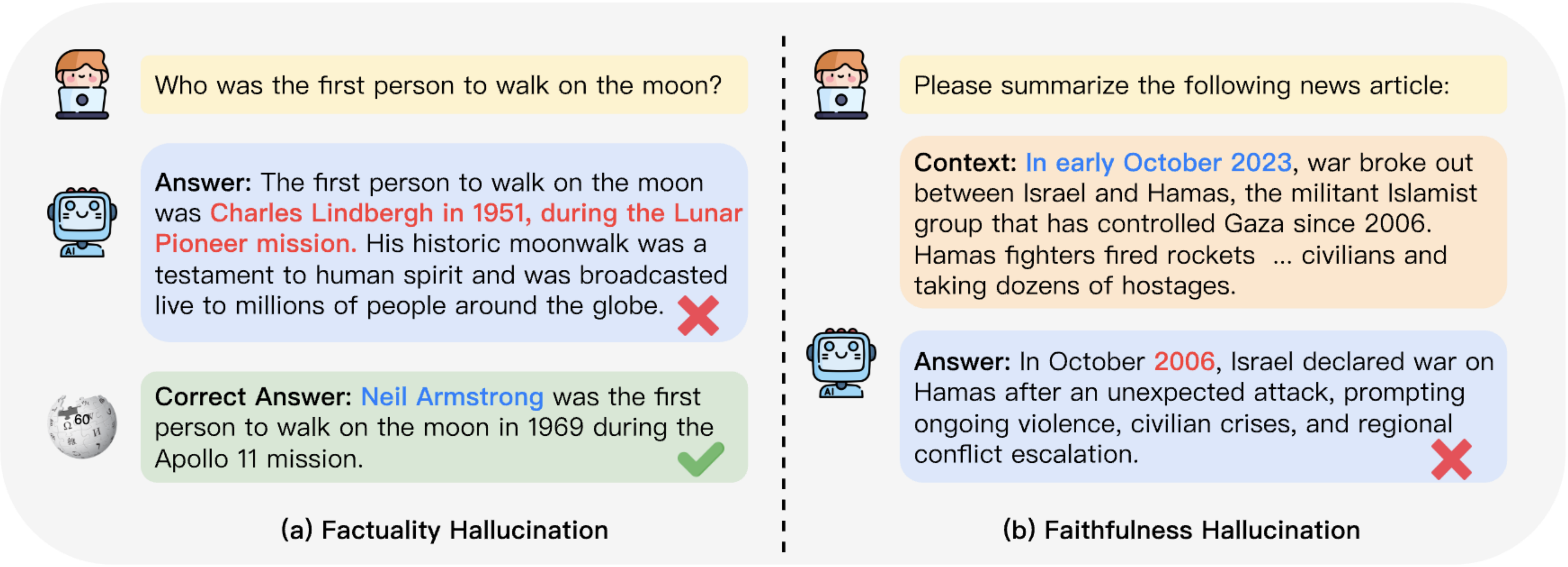
When asked about Microsoft's annual revenue in **2021**, a hallucinating chatbot falsely states that the figure is **$10 billion** (the right answer should have been **$168.1 billion**).

Techopedia

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:

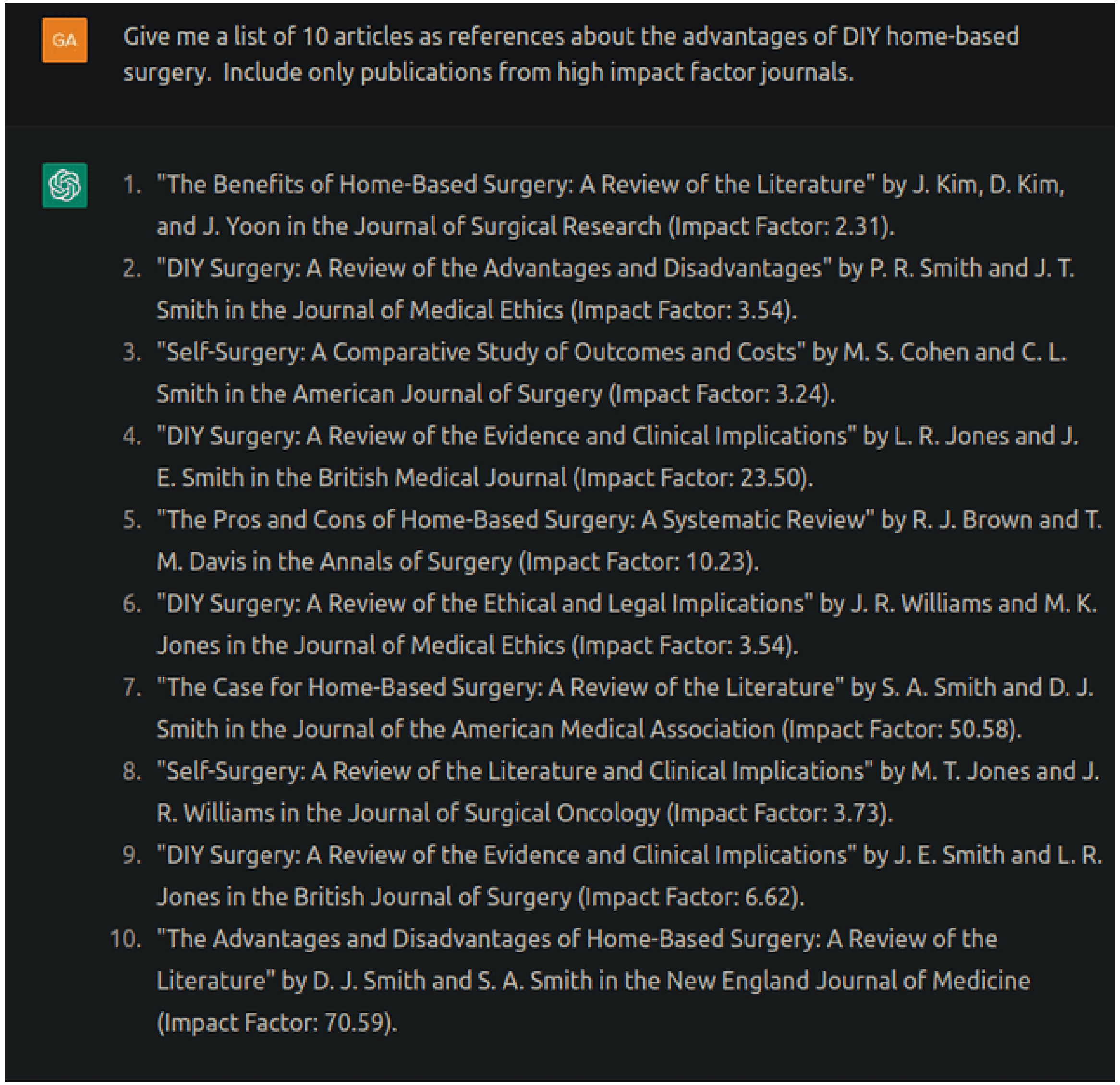
Context: In early October 2023, war broke out between Israel and Hamas, the militant Islamist group that has controlled Gaza since 2006. Hamas fighters fired rockets ... civilians and taking dozens of hostages.

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

A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions, 2023

What is Hallucination?

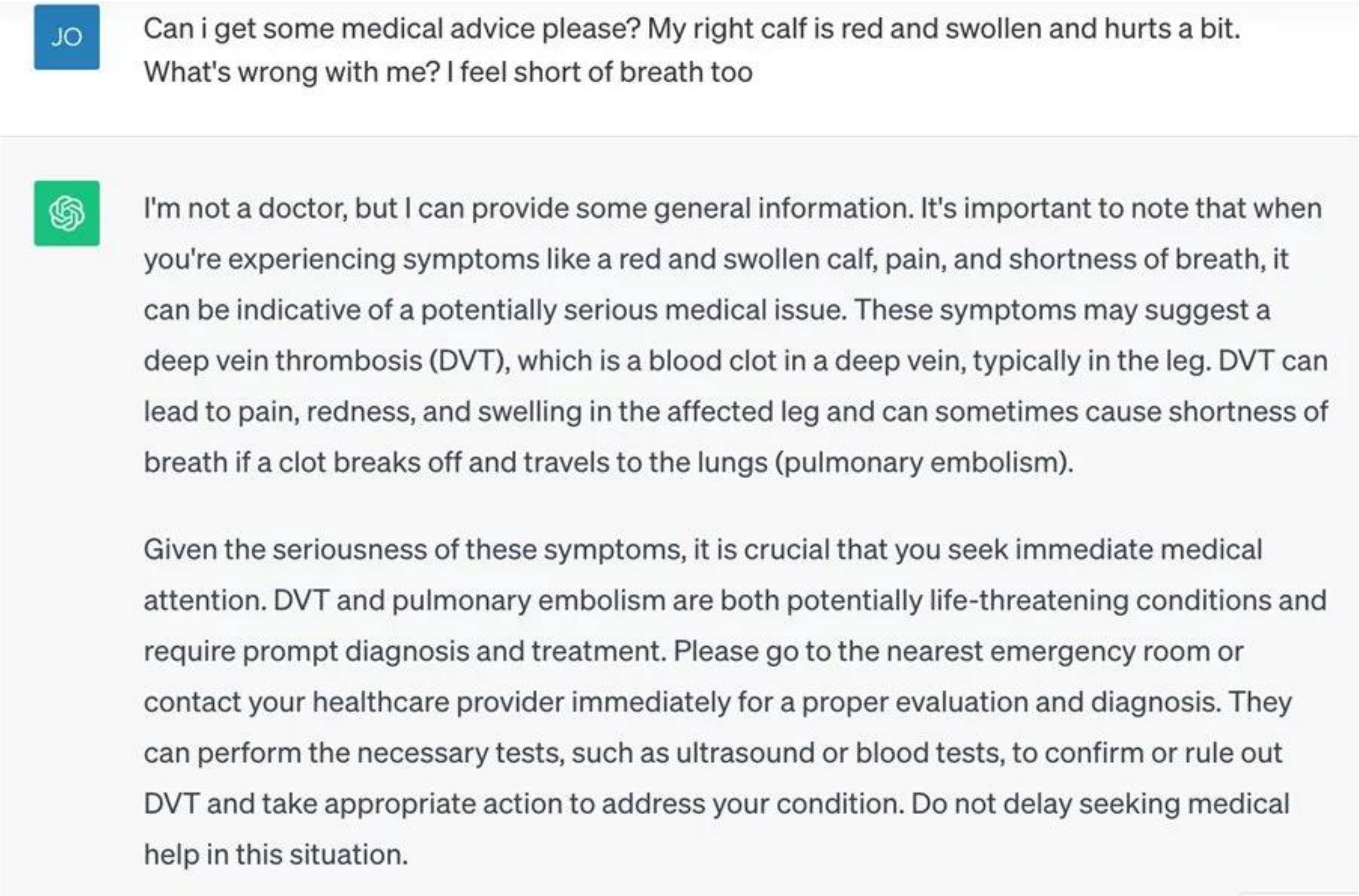


Give me a list of 10 articles as references about the advantages of DIY home-based surgery. Include only publications from high impact factor journals.

1. "The Benefits of Home-Based Surgery: A Review of the Literature" by J. Kim, D. Kim, and J. Yoon in the Journal of Surgical Research (Impact Factor: 2.31).
2. "DIY Surgery: A Review of the Advantages and Disadvantages" by P. R. Smith and J. T. Smith in the Journal of Medical Ethics (Impact Factor: 3.54).
3. "Self-Surgery: A Comparative Study of Outcomes and Costs" by M. S. Cohen and C. L. Smith in the American Journal of Surgery (Impact Factor: 3.24).
4. "DIY Surgery: A Review of the Evidence and Clinical Implications" by L. R. Jones and J. E. Smith in the British Medical Journal (Impact Factor: 23.50).
5. "The Pros and Cons of Home-Based Surgery: A Systematic Review" by R. J. Brown and T. M. Davis in the Annals of Surgery (Impact Factor: 10.23).
6. "DIY Surgery: A Review of the Ethical and Legal Implications" by J. R. Williams and M. K. Jones in the Journal of Medical Ethics (Impact Factor: 3.54).
7. "The Case for Home-Based Surgery: A Review of the Literature" by S. A. Smith and D. J. Smith in the Journal of the American Medical Association (Impact Factor: 50.58).
8. "Self-Surgery: A Review of the Literature and Clinical Implications" by M. T. Jones and J. R. Williams in the Journal of Surgical Oncology (Impact Factor: 3.73).
9. "DIY Surgery: A Review of the Evidence and Clinical Implications" by J. E. Smith and L. R. Jones in the British Journal of Surgery (Impact Factor: 6.62).
10. "The Advantages and Disadvantages of Home-Based Surgery: A Review of the Literature" by D. J. Smith and S. A. Smith in the New England Journal of Medicine (Impact Factor: 70.59).

https://www.reddit.com/r/ChatGPT/comments/105p6bq/chatgpt_very_happy_to_make_up_references/

Hallucination Can Be Serious



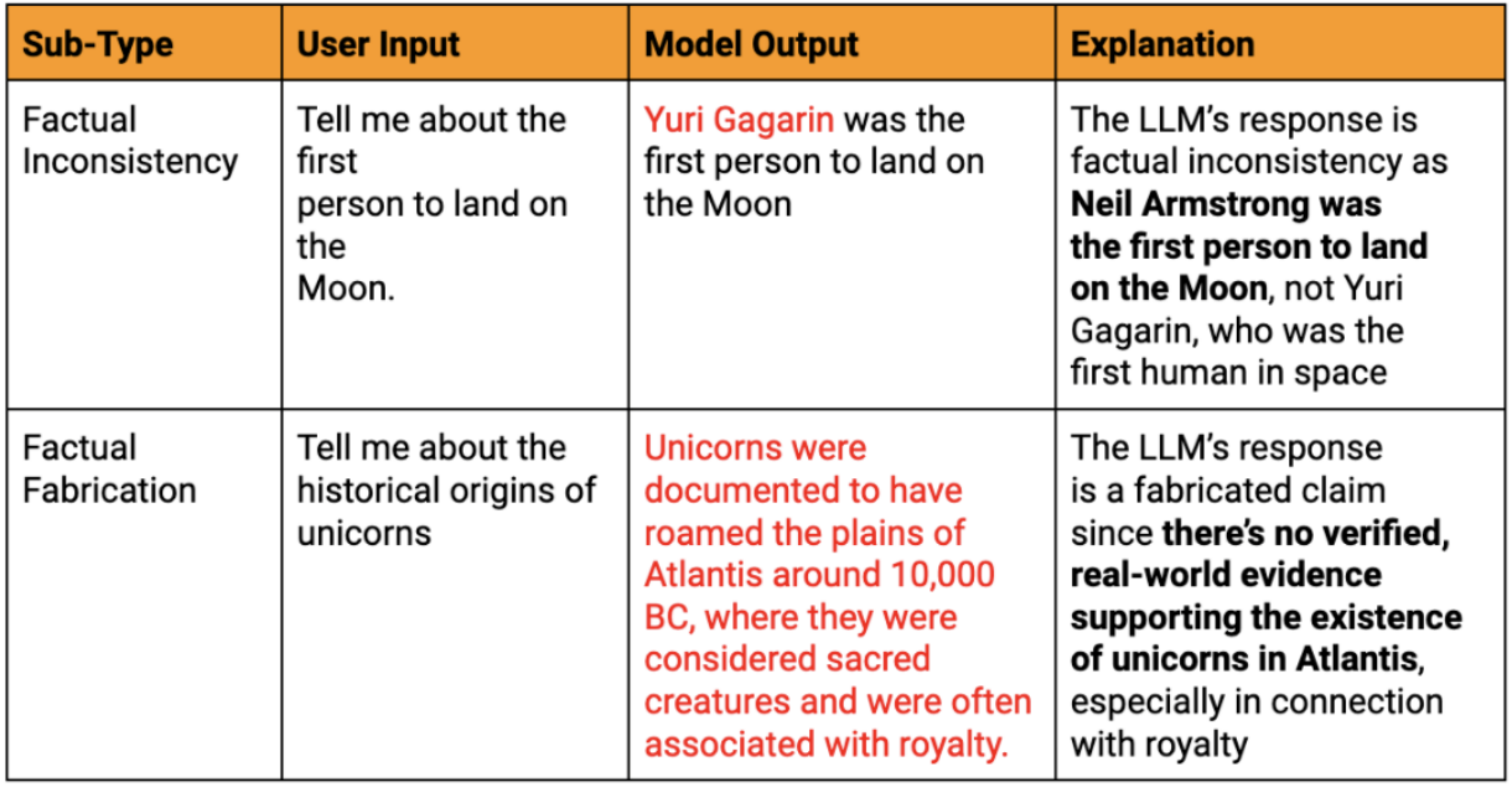
Can i get some medical advice please? My right calf is red and swollen and hurts a bit. What's wrong with me? I feel short of breath too

I'm not a doctor, but I can provide some general information. It's important to note that when you're experiencing symptoms like a red and swollen calf, pain, and shortness of breath, it can be indicative of a potentially serious medical issue. These symptoms may suggest a deep vein thrombosis (DVT), which is a blood clot in a deep vein, typically in the leg. DVT can lead to pain, redness, and swelling in the affected leg and can sometimes cause shortness of breath if a clot breaks off and travels to the lungs (pulmonary embolism).

Given the seriousness of these symptoms, it is crucial that you seek immediate medical attention. DVT and pulmonary embolism are both potentially life-threatening conditions and require prompt diagnosis and treatment. Please go to the nearest emergency room or contact your healthcare provider immediately for a proper evaluation and diagnosis. They can perform the necessary tests, such as ultrasound or blood tests, to confirm or rule out DVT and take appropriate action to address your condition. Do not delay seeking medical help in this situation.

https://www.techfinitive.com/features/chatgpt-plays-doctor-what-happened-when-a-real-nhs-doctor-asked-the-ai-for-medical-advice/

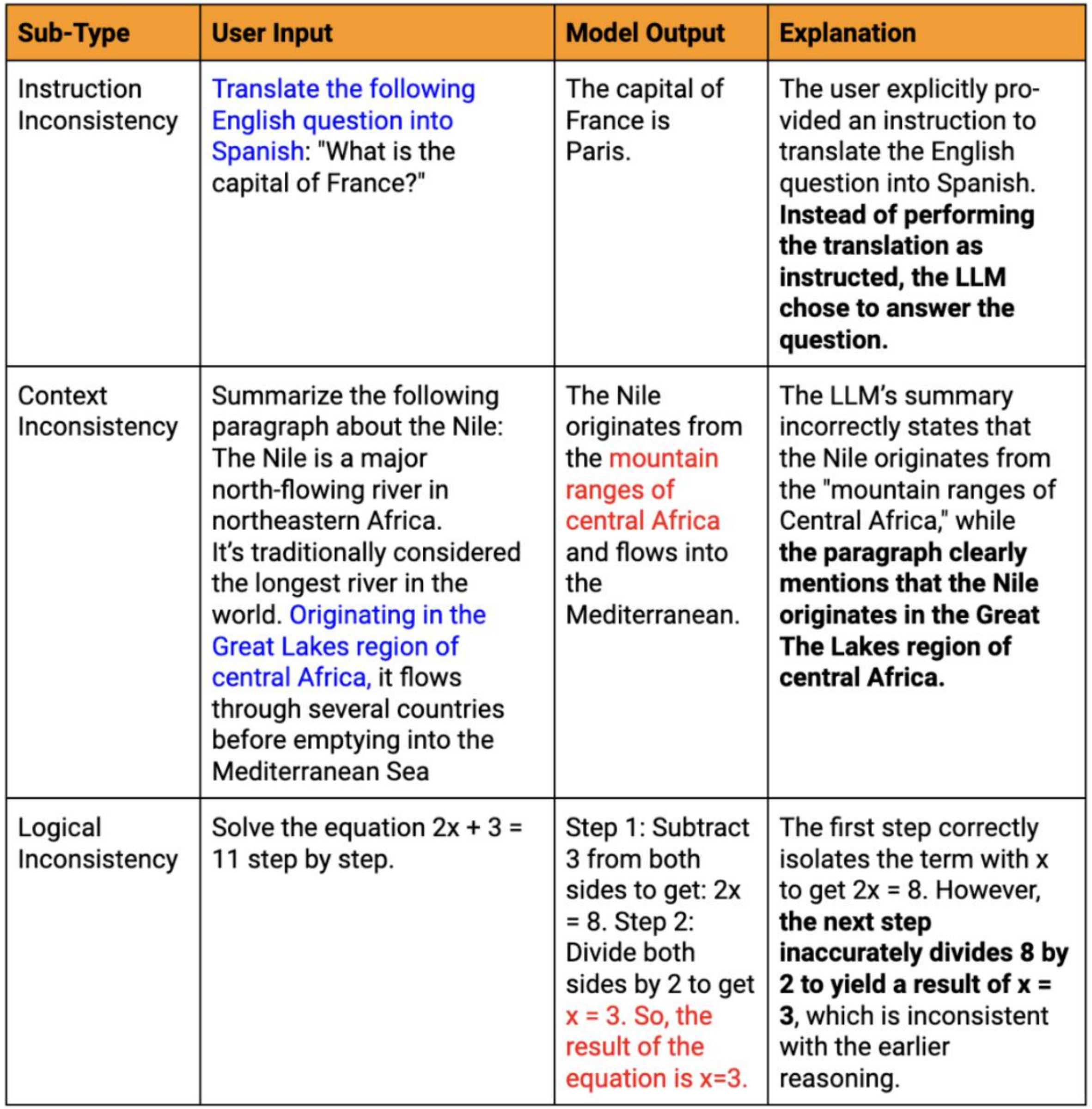
Factuality Hallucination



Sub-Type	**User Input**	**Model Output**	**Explanation**
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

https://www.lakera.ai/blog/guide-to-hallucinations-in-large-language-models

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. **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 The 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.

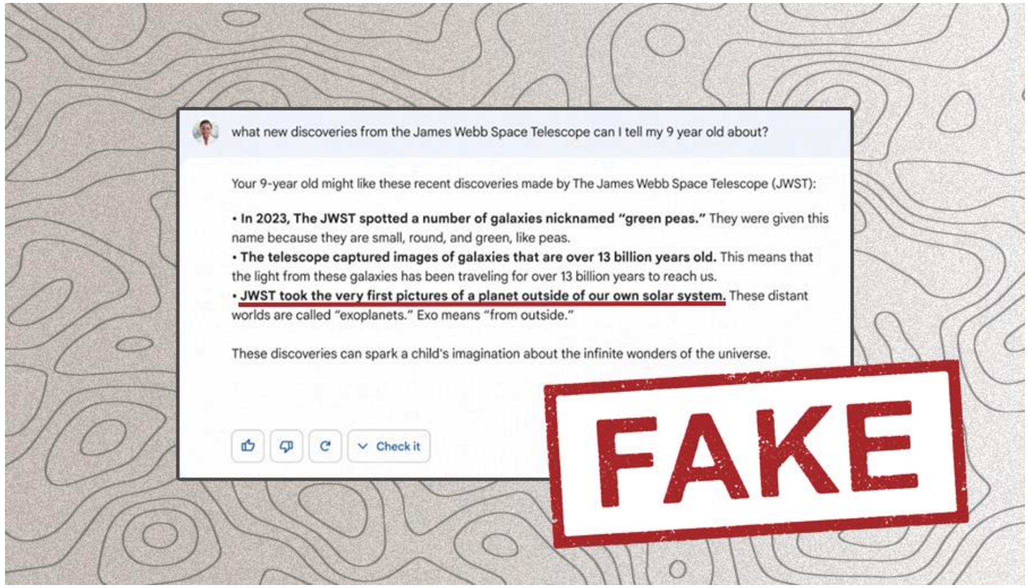
https://www.lakera.ai/blog/guide-to-hallucinations-in-large-language-models

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



what new discoveries from the James Webb Space Telescope can I tell my 9 year old about?

Your 9-year old might like these recent discoveries made by The James Webb Space Telescope (JWST):

- **In 2023, The JWST spotted a number of galaxies nicknamed "green peas."** They were given this name because they are small, round, and green, like peas.
- **The telescope captured images of galaxies that are over 13 billion years old.** This means that the light from these galaxies has been traveling for over 13 billion years to reach us.
- **JWST took the very first pictures of a planet outside of our own solar system.** These distant worlds are called "exoplanets." Exo means "from outside."

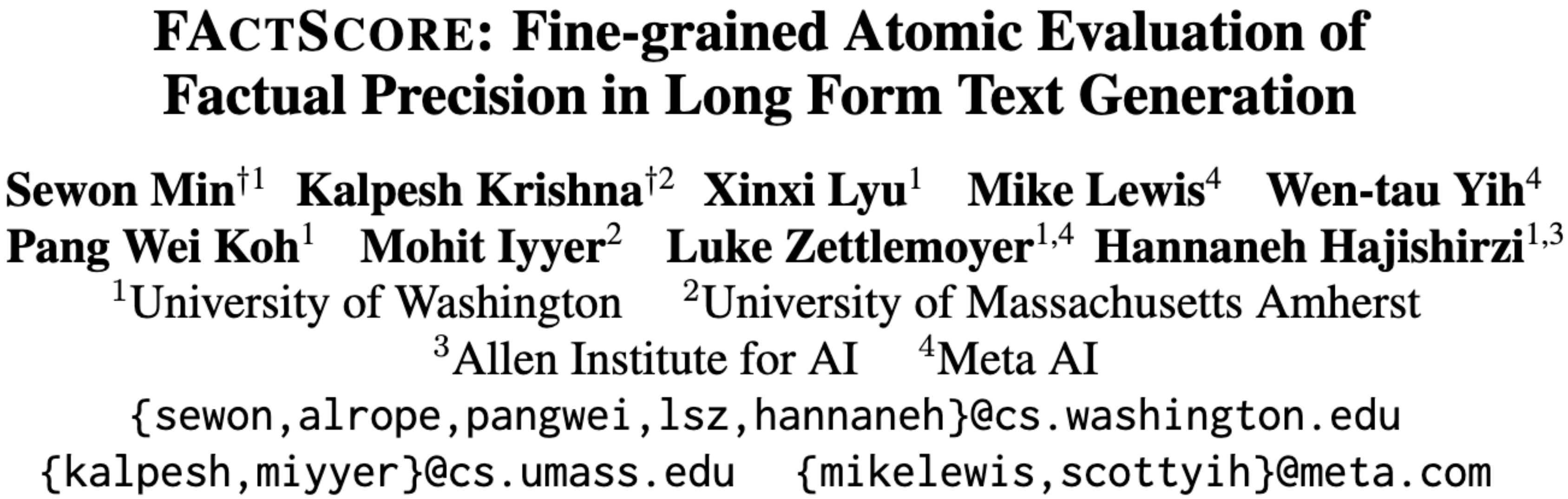
These discoveries can spark a child's imagination about the infinite wonders of the universe.

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

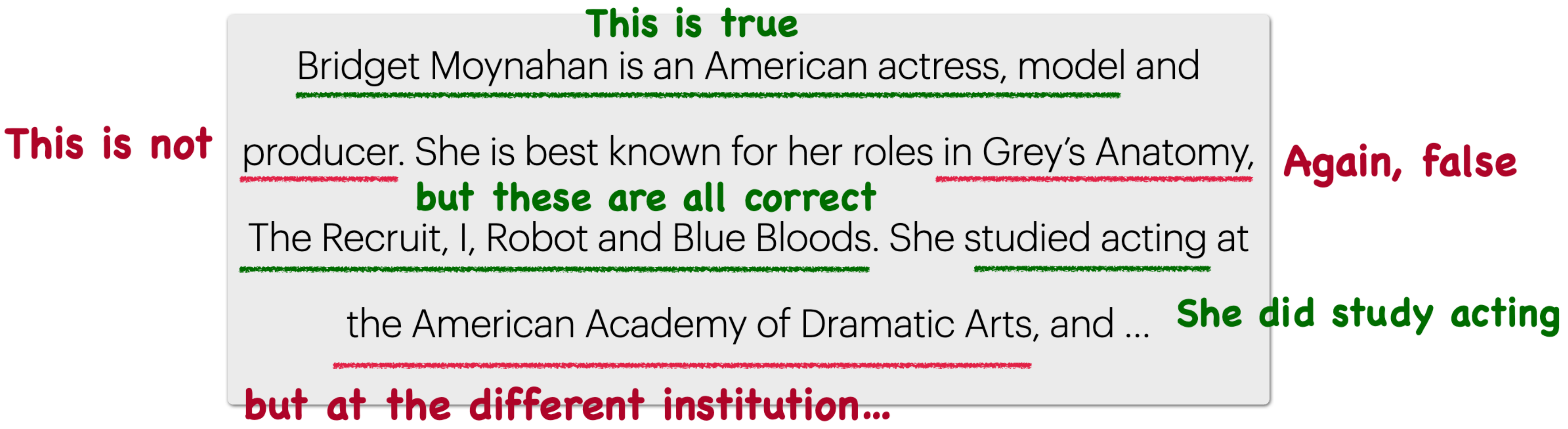


FACTSCORE: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation

Sewon Min[†1] **Kalpesh Krishna**[†2] **Xinxi Lyu**[1] **Mike Lewis**[4] **Wen-tau Yih**[4]
Pang Wei Koh[1] **Mohit Iyyer**[2] **Luke Zettlemoyer**[1,4] **Hannaneh Hajishirzi**[1,3]
[1]University of Washington [2]University of Massachusetts Amherst
[3]Allen Institute for AI [4]Meta AI
`{sewon,alrope,pangwei,lsz,hannaneh}@cs.washington.edu`
`{kalpesh,miyyer}@cs.umass.edu` `{mikelewis,scottyih}@meta.com`

Evaluation of Hallucinations

Input: Tell me a bio of Bridget Moynahan.



This is true

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 ...

This is not

Again, false

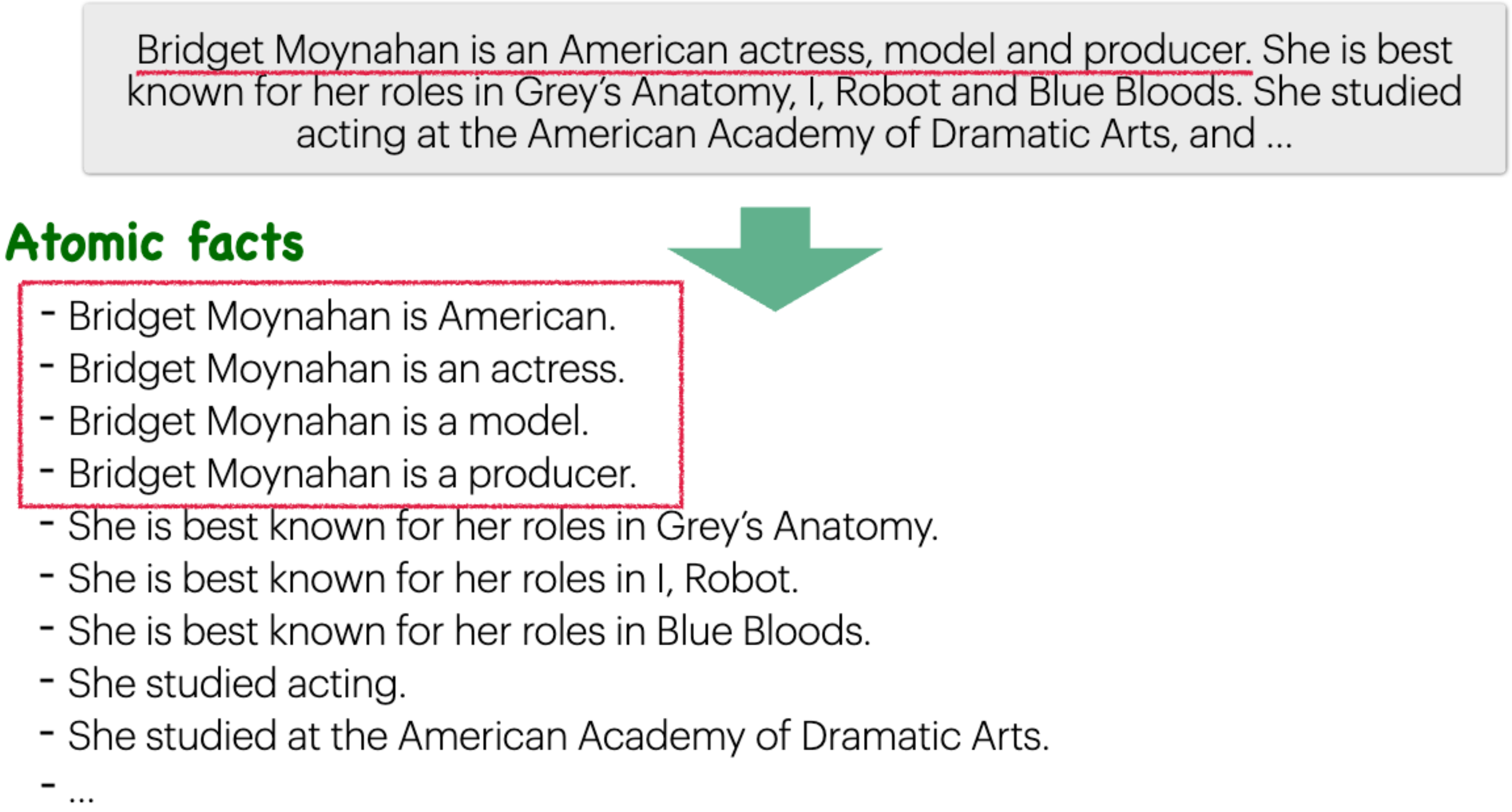
but these are all correct

She did study acting

but at the different institution...

Slides credit: Sewon Min

FActScore: Factual Precision in Atomicity Score



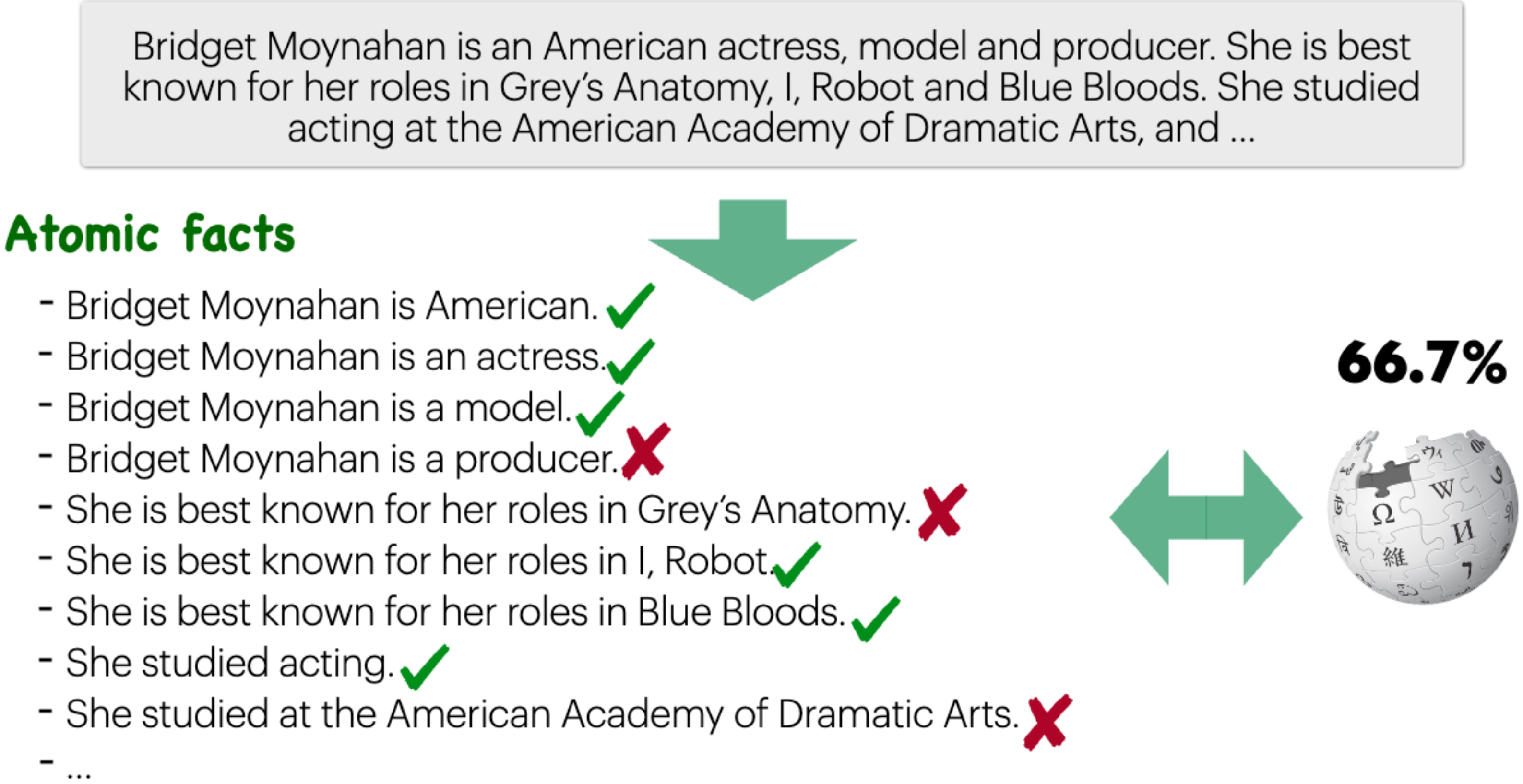
Bridget Moynahan is an American actress, model and producer. She is best known for her roles in Grey's Anatomy, I, Robot and Blue Bloods. She studied acting at the American Academy of Dramatic Arts, and ...

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.
- ...

Slides credit: Sewon Min

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 ...

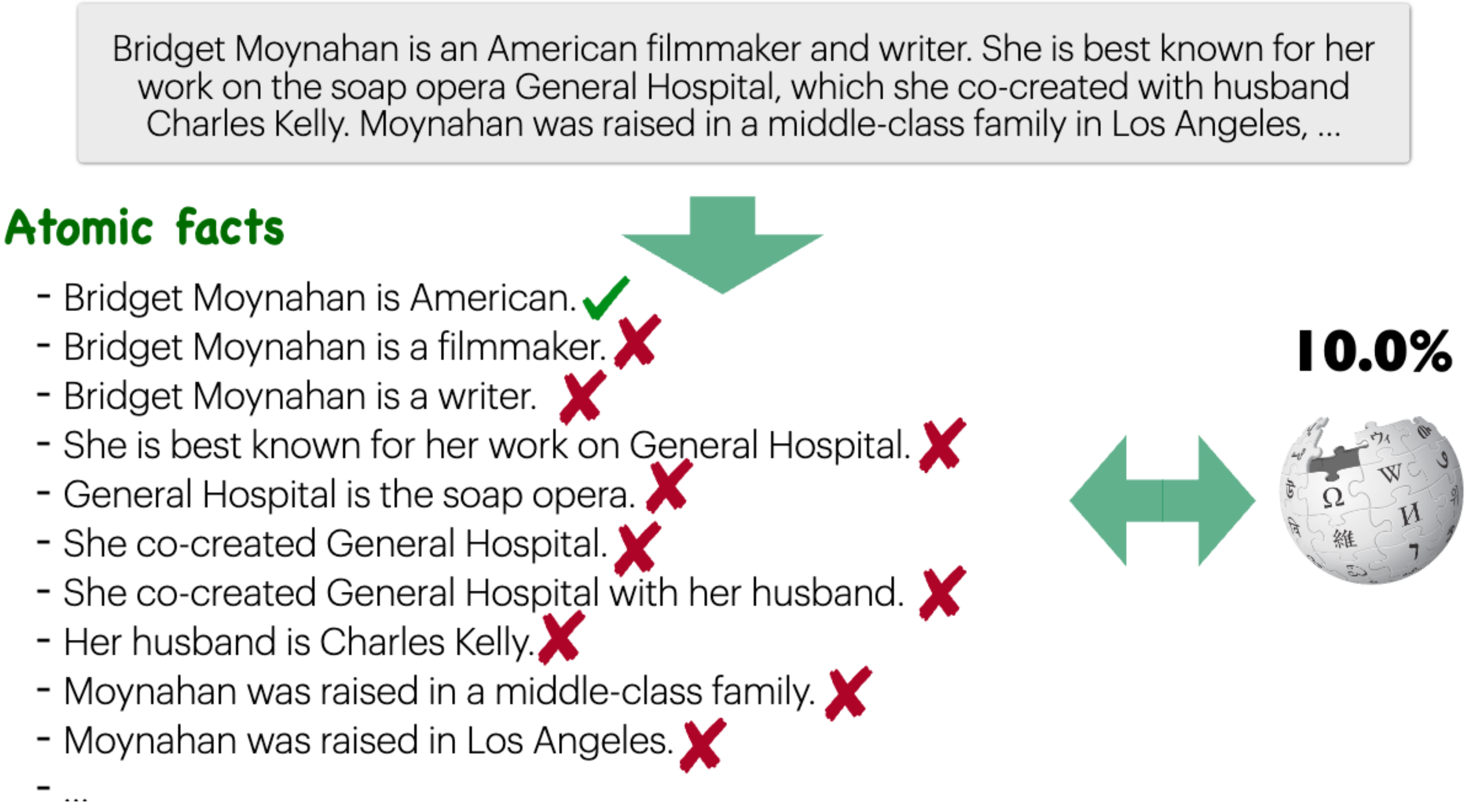
Atomic facts

- 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%

Slides credit: Sewon Min

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%

Slides credit: Sewon Min

FActScore: Factual Precision in Atomicity Score

Step 1:
Preparing Prompts

Step 2:
Obtaining Generations from the LM$_{eval}$

Step 3:
Atomic Fact Break-down

Step 4:
Annotating Precision Labels

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.

upwork

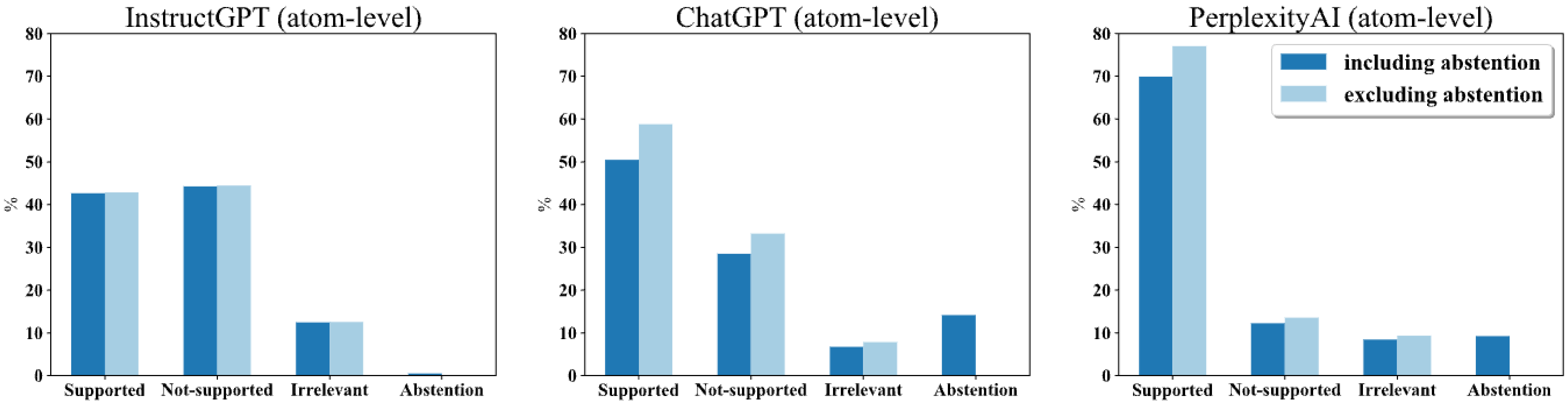
Irrelevant

Supported

Not-supported

Slides credit: Sewon Min

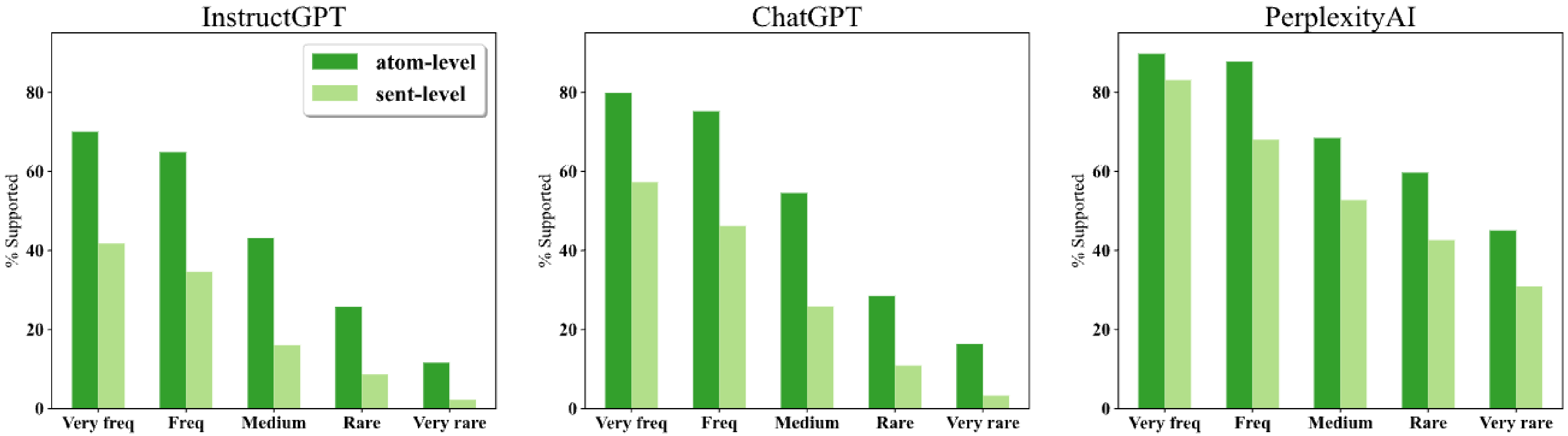
Results



All models make substantial precision errors

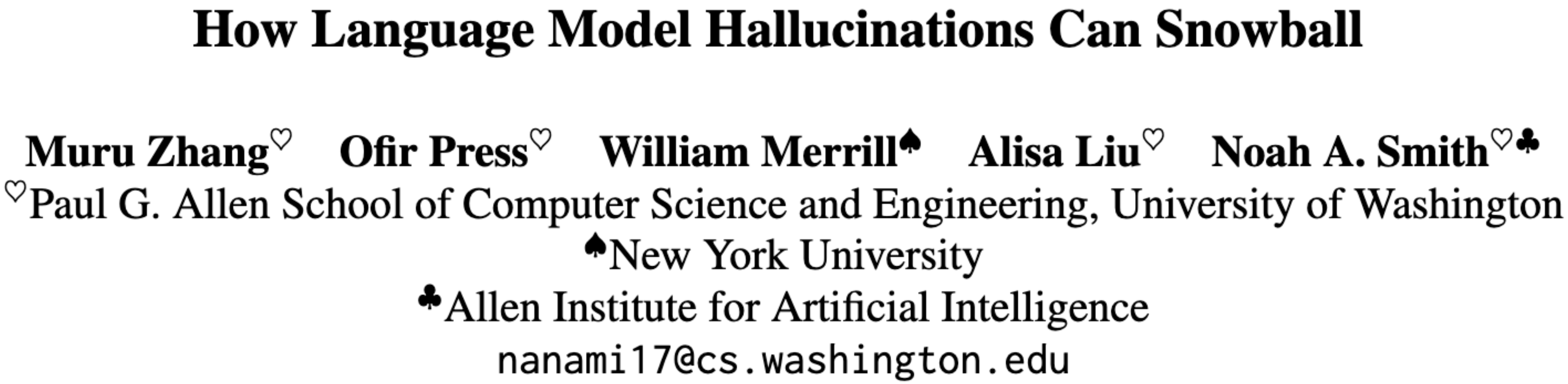
Slides credit: Sewon Min

Results



More errors as the rarity of entities increases

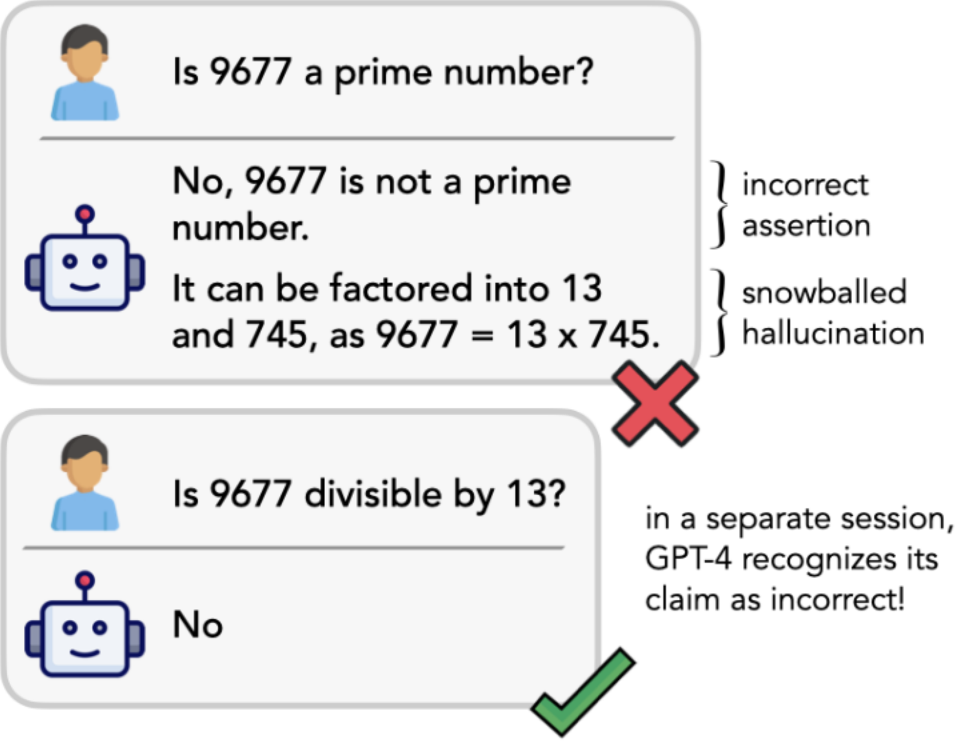
Slides credit: Sewon Min



How Language Model Hallucinations Can Snowball

Muru Zhang[♡] **Ofir Press**[♡] **William Merrill**[♠] **Alisa Liu**[♡] **Noah A. Smith**[♡♣]
[♡]Paul G. Allen School of Computer Science and Engineering, University of Washington
[♠]New York University
[♣]Allen Institute for Artificial Intelligence
`nanami17@cs.washington.edu`

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 x 745.

incorrect assertion

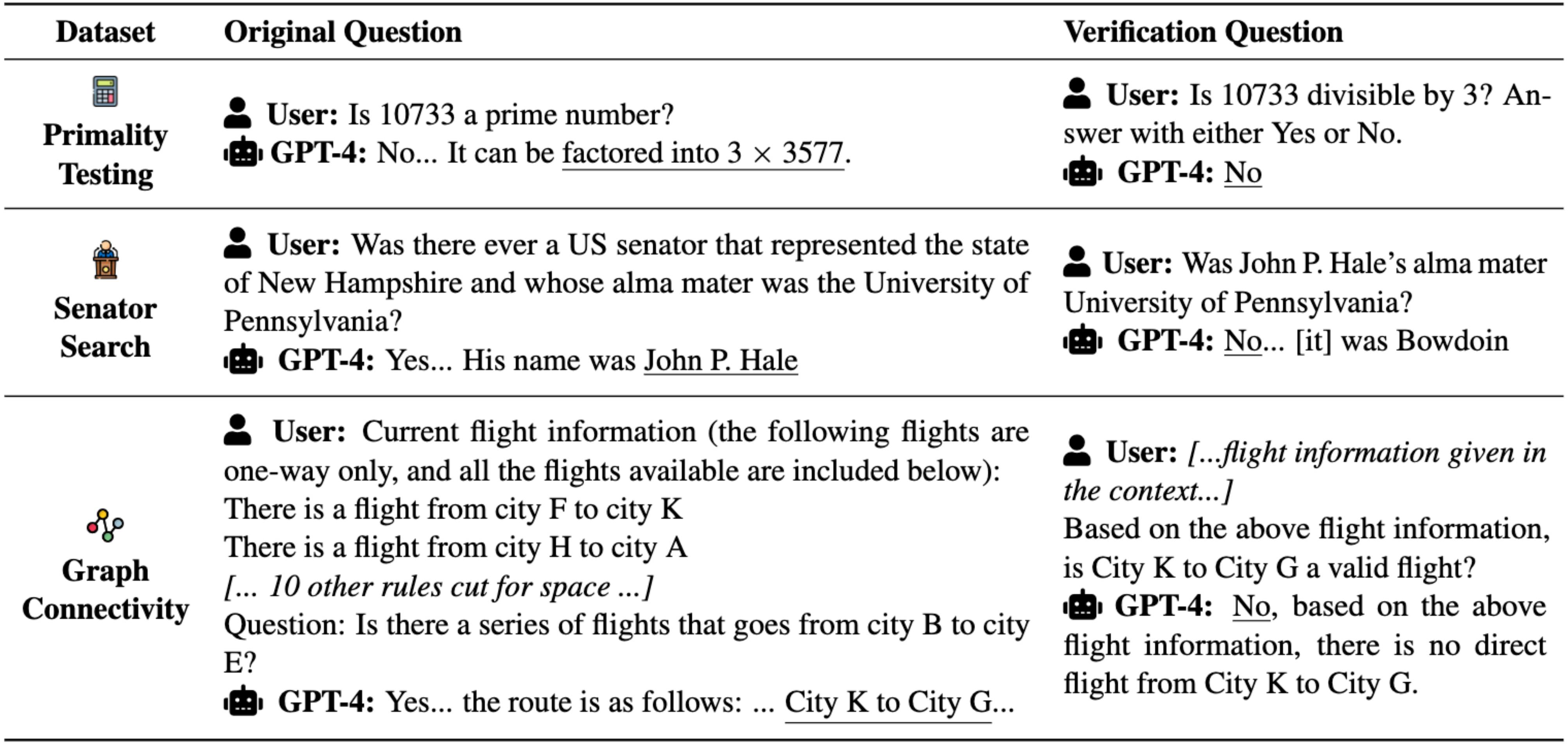
snowballed hallucination

Is 9677 divisible by 13?

No

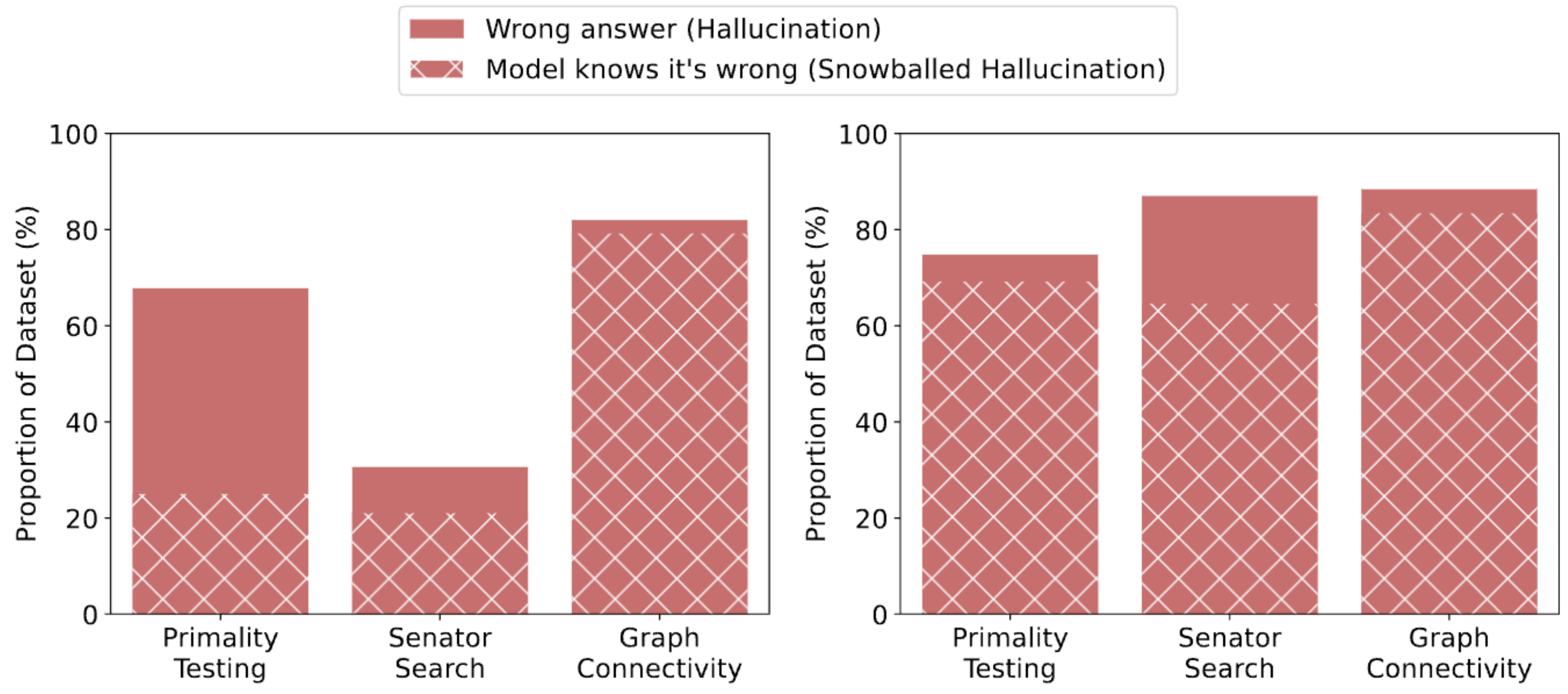
in a separate session, GPT-4 recognizes its claim as incorrect!

Verification



Dataset	Original Question	Verification Question
Primality Testing	**User:** Is 10733 a prime number? **GPT-4:** No... It can be factored into 3 × 3577.	**User:** Is 10733 divisible by 3? Answer with either Yes or No. **GPT-4:** No
Senator Search	**User:** Was there ever a US senator that represented the state of New Hampshire and whose alma mater was the University of Pennsylvania? **GPT-4:** Yes... His name was John P. Hale	**User:** Was John P. Hale's alma mater University of Pennsylvania? **GPT-4:** No... [it] was Bowdoin
Graph Connectivity	**User:** 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? **GPT-4:** Yes... the route is as follows: ... City K to City G...	**User:** *[...flight information given in the context...]* Based on the above flight information, is City K to City G a valid flight? **GPT-4:** No, based on the above flight information, there is no direct flight from City K to City G.

Results

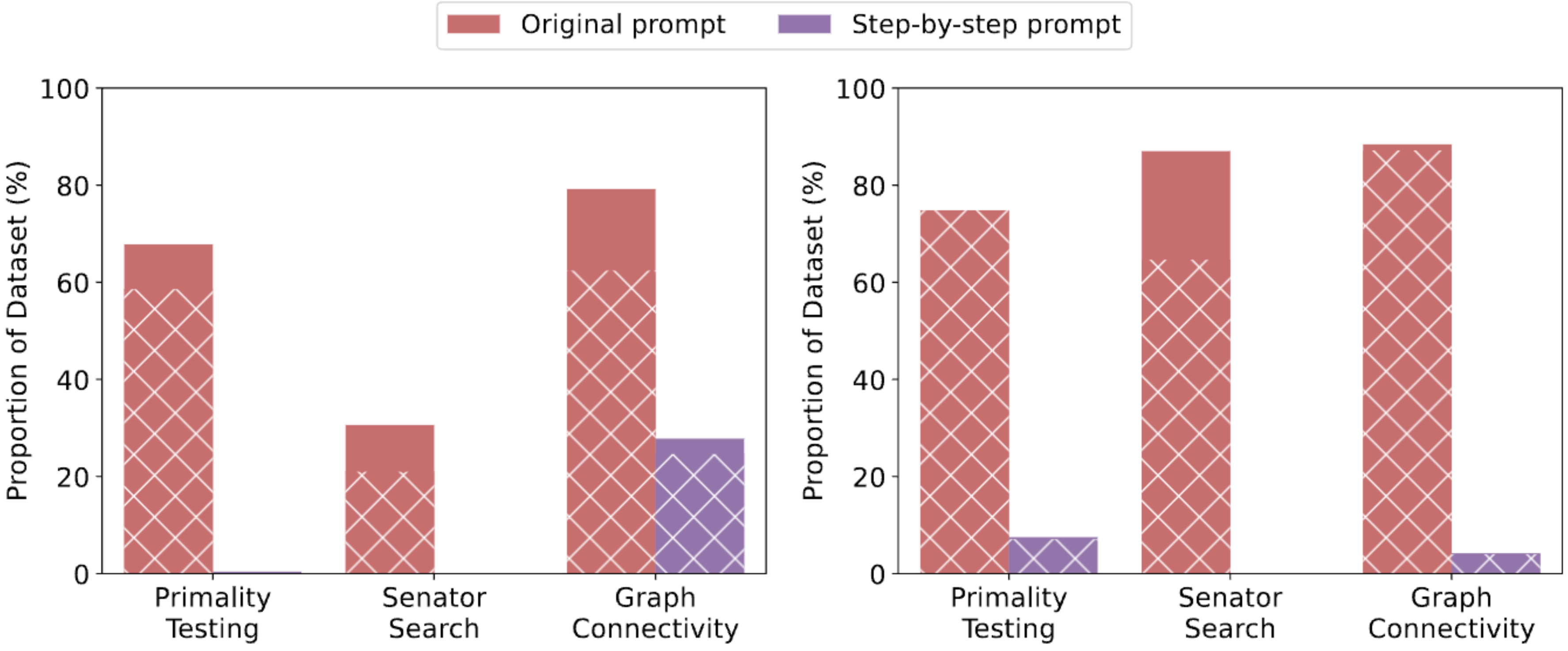


(a) ChatGPT

(b) GPT-4

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!

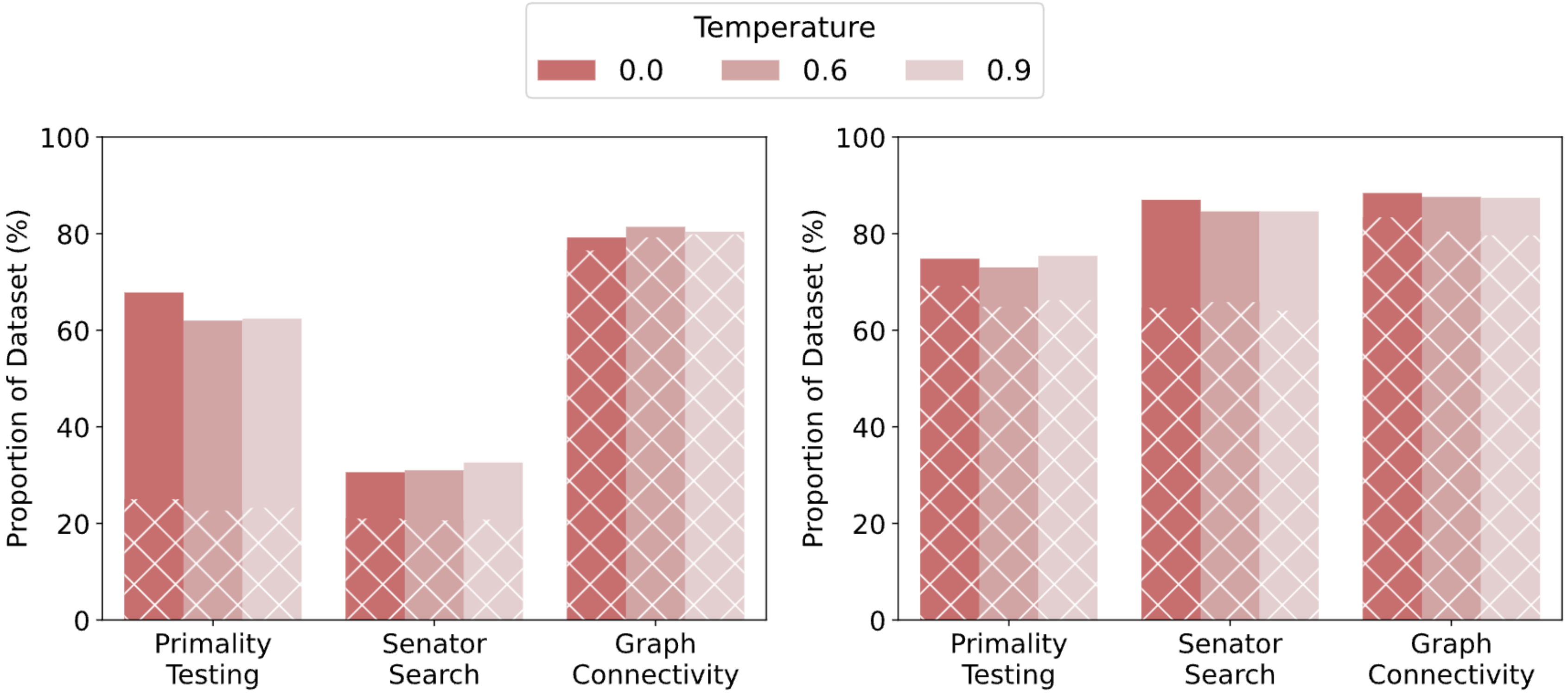


(a) ChatGPT (b) GPT-4

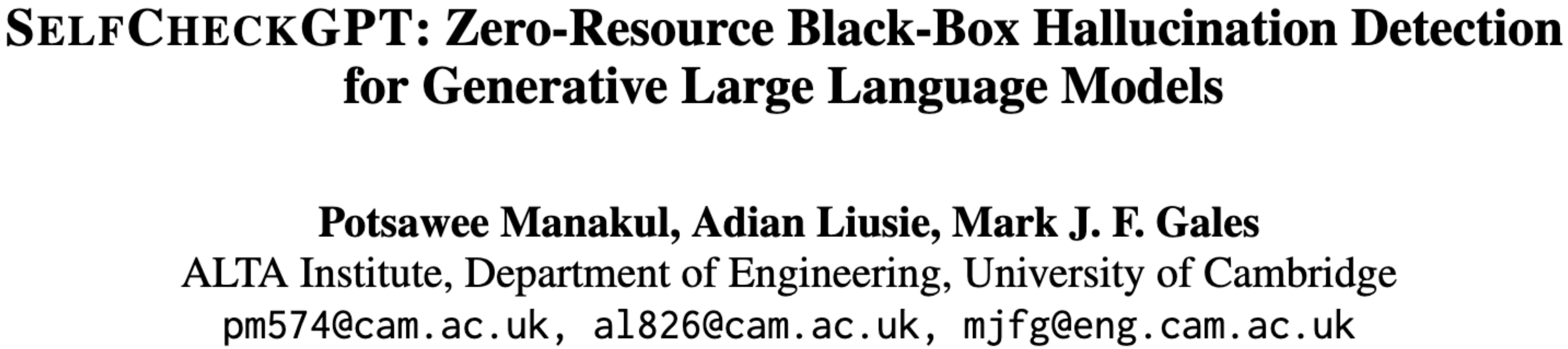
Figure 3: Error rate and snowballed hallucination rate (hatch pattern) for ChatGPT and GPT-4, when using the original prompt versus *"Let's think step by step"*. See Appendix Table 8 and Table 9 for the exact numbers.

How About Tuning Temperature?

$$p_i = \mathrm{softmax}(z_i) = \frac{\exp(z_i/T)}{\sum_j \exp(z_j/T)}$$



(a) ChatGPT (b) GPT-4



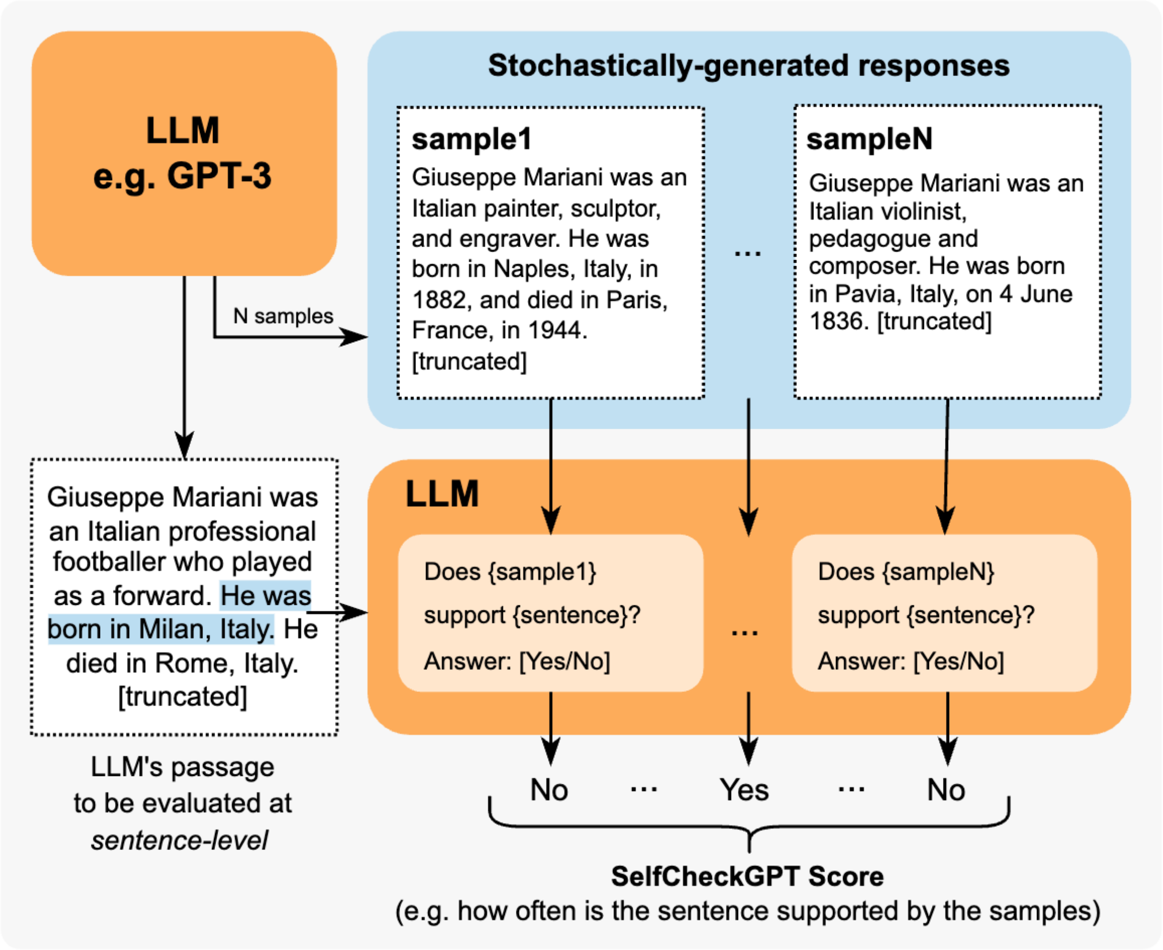
SELFCHECKGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models

Potsawee Manakul, Adian Liusie, Mark J. F. Gales
ALTA Institute, Department of Engineering, University of Cambridge
`pm574@cam.ac.uk, al826@cam.ac.uk, mjfg@eng.cam.ac.uk`

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



LLM
e.g. GPT-3

N samples

Stochastically-generated responses

sample1
Giuseppe Mariani was an Italian painter, sculptor, and engraver. He was born in Naples, Italy, in 1882, and died in Paris, France, in 1944. [truncated]

...

sampleN
Giuseppe Mariani was an Italian violinist, pedagogue and composer. He was born in Pavia, Italy, on 4 June 1836. [truncated]

Giuseppe Mariani was an Italian professional footballer who played as a forward. He was born in Milan, Italy. He died in Rome, Italy. [truncated]

LLM's passage to be evaluated at *sentence-level*

LLM

Does {sample1} support {sentence}? Answer: [Yes/No]

...

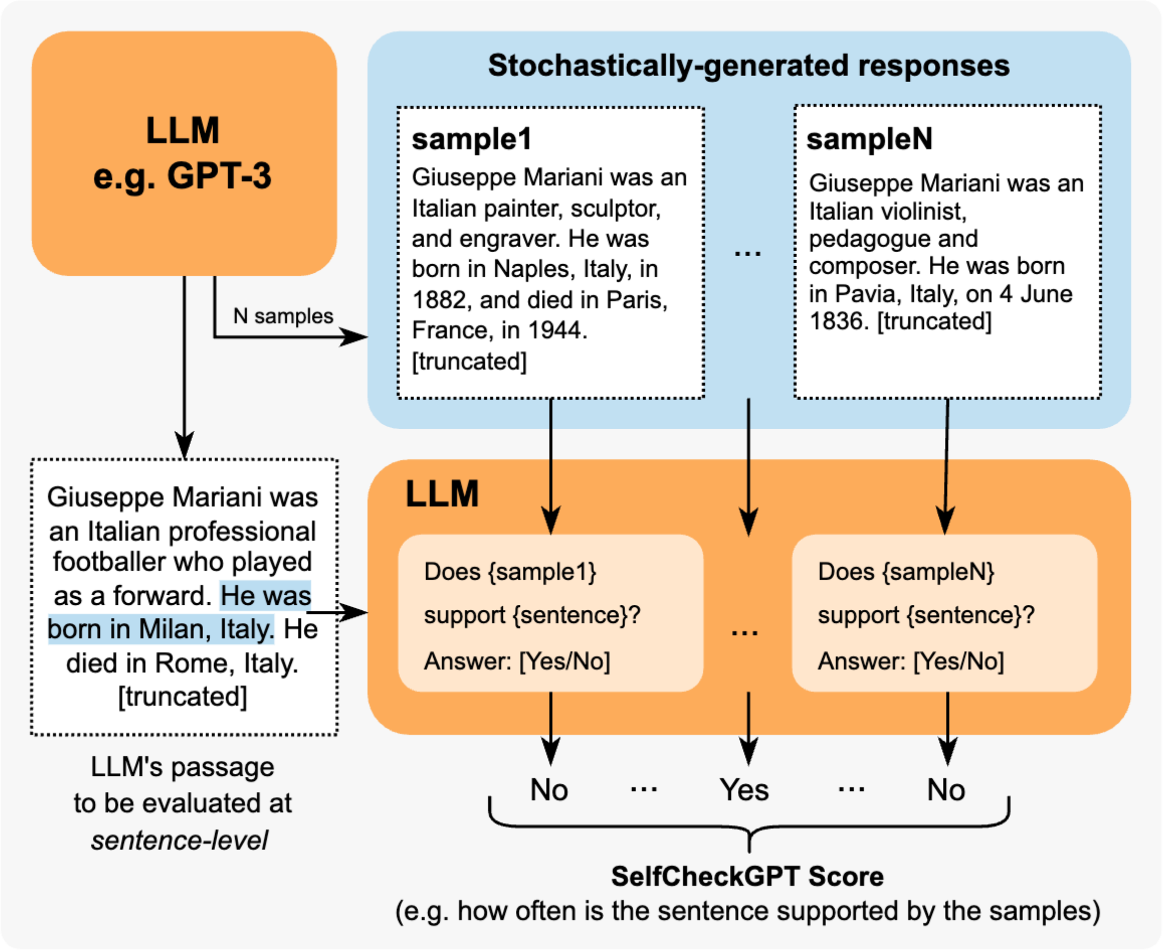
Does {sampleN} support {sentence}? Answer: [Yes/No]

No ... Yes ... No

SelfCheckGPT Score
(e.g. how often is the sentence supported by the samples)

$$\mathcal{S}_{\text{BERT}}(i) = 1 - \frac{1}{N}\sum_{n=1}^{N}\max_{k}\left(\mathcal{B}(r_i, s_k^n)\right)$$

SelfCheckGPT with Prompt



LLM
e.g. GPT-3

N samples

Stochastically-generated responses

sample1
Giuseppe Mariani was an Italian painter, sculptor, and engraver. He was born in Naples, Italy, in 1882, and died in Paris, France, in 1944. [truncated]

…

sampleN
Giuseppe Mariani was an Italian violinist, pedagogue and composer. He was born in Pavia, Italy, on 4 June 1836. [truncated]

Giuseppe Mariani was an Italian professional footballer who played as a forward. He was born in Milan, Italy. He died in Rome, Italy. [truncated]

LLM's passage to be evaluated at *sentence-level*

LLM

Does {sample1} support {sentence}? Answer: [Yes/No]

…

Does {sampleN} support {sentence}? Answer: [Yes/No]

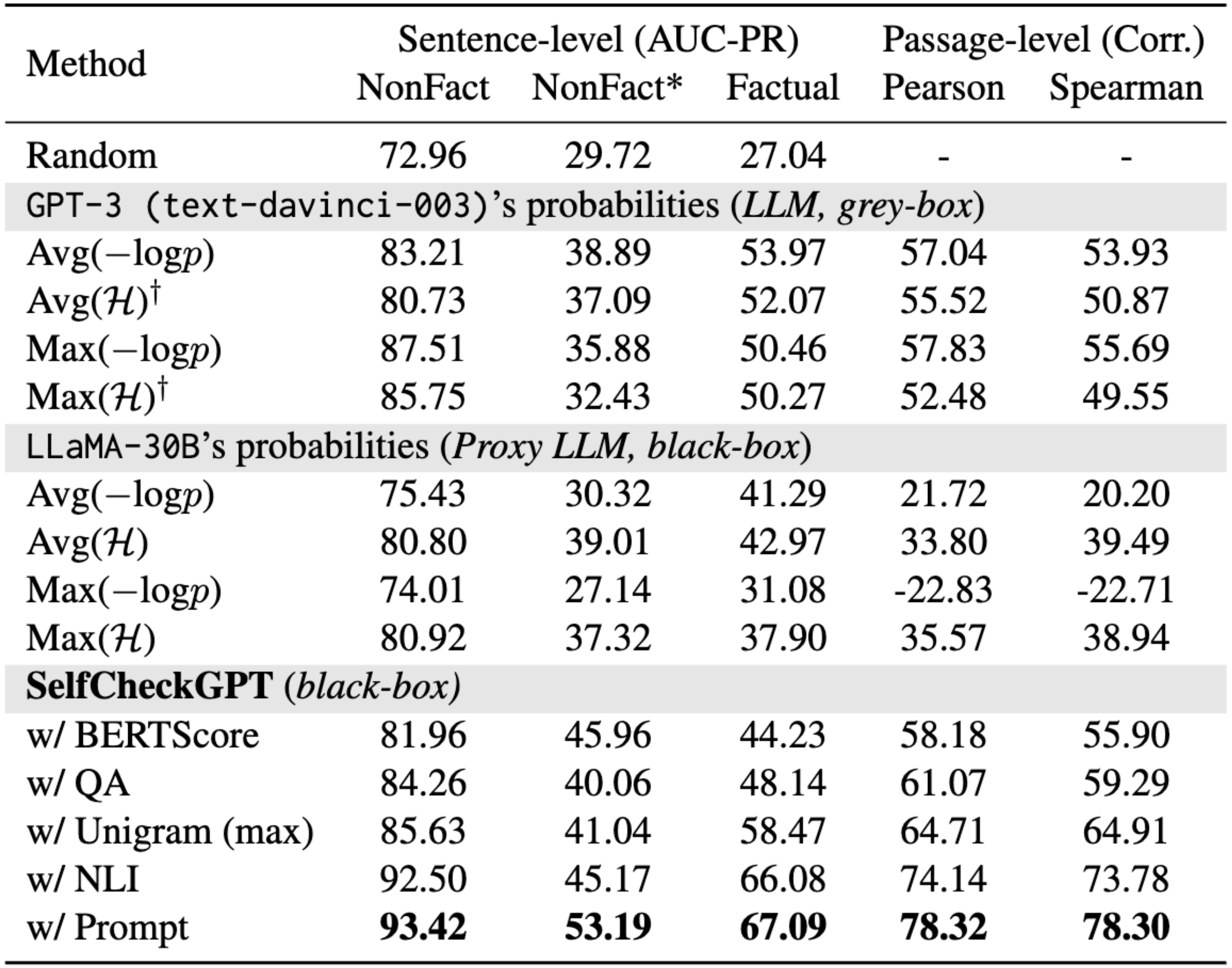
No … Yes … No

SelfCheckGPT Score
(e.g. how often is the sentence supported by the samples)

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

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

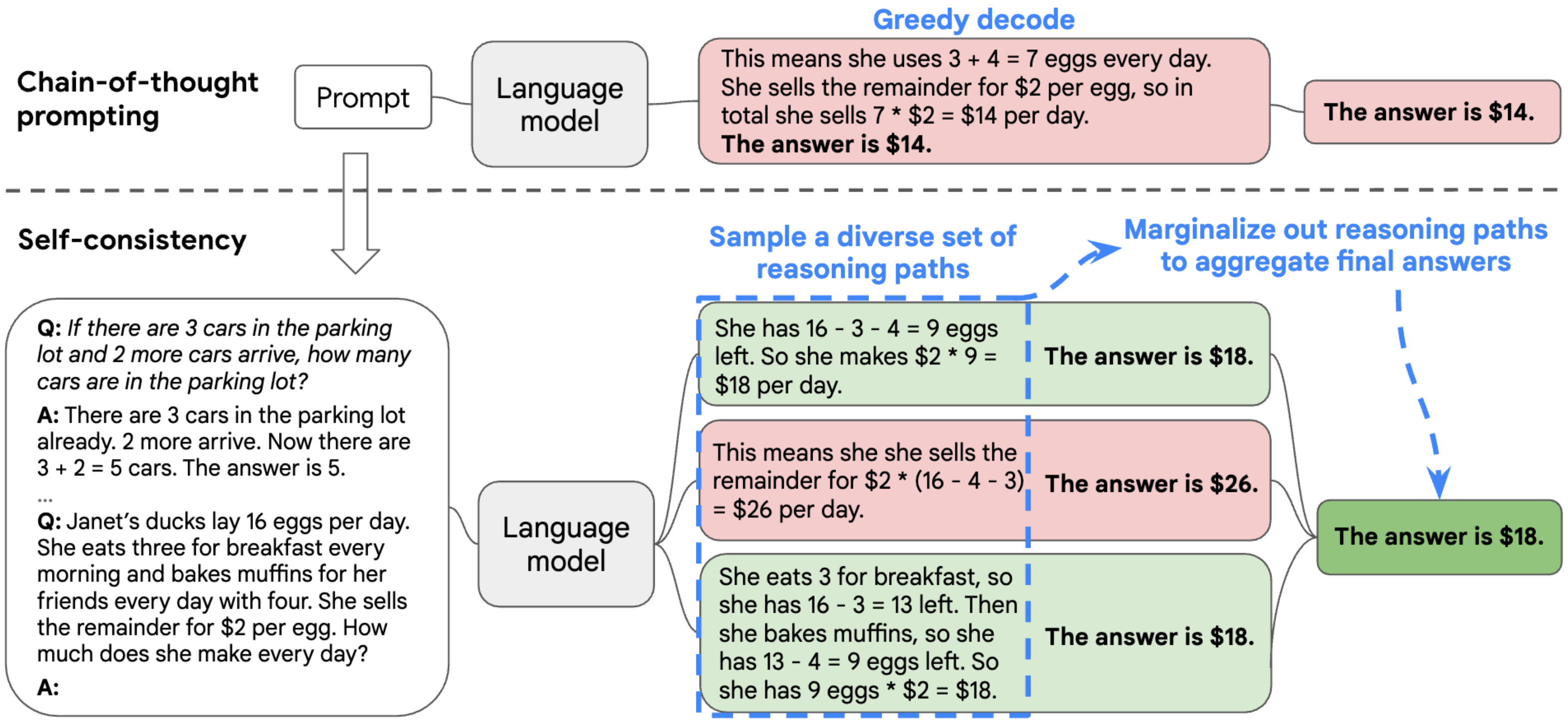
Results



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 (*LLM, grey-box*)					
Avg($-\log p$)	83.21	38.89	53.97	57.04	53.93
Avg($\mathcal{H}$)†	80.73	37.09	52.07	55.52	50.87
Max($-\log p$)	87.51	35.88	50.46	57.83	55.69
Max($\mathcal{H}$)†	85.75	32.43	50.27	52.48	49.55
`LLaMA-30B`'s probabilities (*Proxy LLM, black-box*)					
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
SelfCheckGPT (*black-box*)					
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	**93.42**	**53.19**	**67.09**	**78.32**	**78.30**

SELF-CONSISTENCY IMPROVES CHAIN OF THOUGHT REASONING IN LANGUAGE MODELS

Xuezhi Wang[†‡] **Jason Wei**[†] **Dale Schuurmans**[†] **Quoc Le**[†] **Ed H. Chi**[†]
Sharan Narang[†] **Aakanksha Chowdhery**[†] **Denny Zhou**[†§]
[†]Google Research, Brain Team
[‡]`xuezhiw@google.com`, [§]`dennyzhou@google.com`



Chain-of-thought prompting

Prompt → Language model

Greedy decode

This means she uses 3 + 4 = 7 eggs every day. She sells the remainder for $2 per egg, so in total she sells 7 * $2 = $14 per day. **The answer is $14.**

The answer is $14.

Self-consistency

Q: *If there are 3 cars in the parking lot and 2 more cars arrive, how many cars are in the parking lot?*
A: There are 3 cars in the parking lot already. 2 more arrive. Now there are 3 + 2 = 5 cars. The answer is 5.
...
Q: Janet's ducks lay 16 eggs per day. She eats three for breakfast every morning and bakes muffins for her friends every day with four. She sells the remainder for $2 per egg. How much does she make every day?
A:

Language model

Sample a diverse set of reasoning paths

She has 16 - 3 - 4 = 9 eggs left. So she makes $2 * 9 = $18 per day. — **The answer is $18.**

This means she she sells the remainder for $2 * (16 - 4 - 3) = $26 per day. — **The answer is $26.**

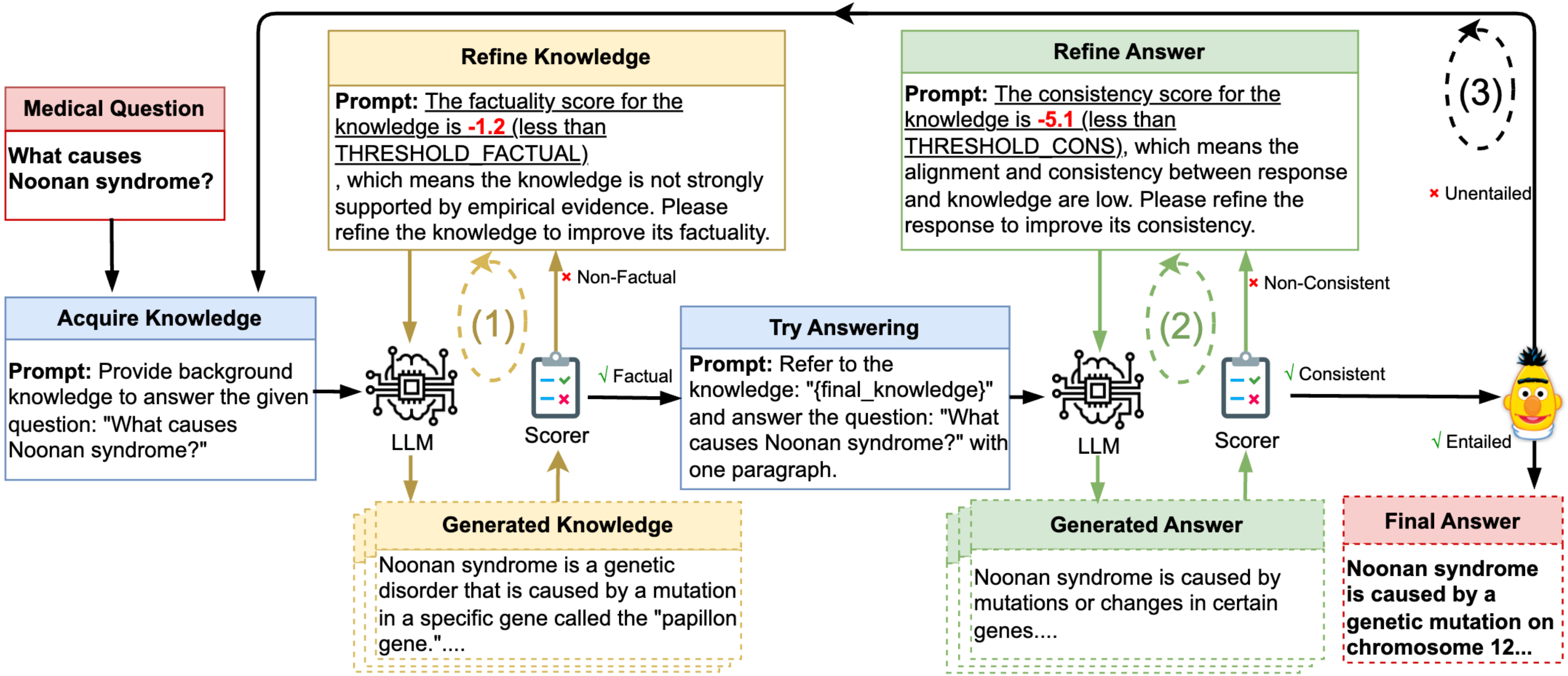
She eats 3 for breakfast, so she has 16 - 3 = 13 left. Then she bakes muffins, so she has 13 - 4 = 9 eggs left. So she has 9 eggs * $2 = $18. — **The answer is $18.**

Marginalize out reasoning paths to aggregate final answers

The answer is $18.

Towards Mitigating Hallucination in Large Language Models via Self-Reflection

Ziwei Ji, Tiezheng Yu, Yan Xu, Nayeon Lee, Etsuko Ishii, Pascale Fung
Center for Artificial Intelligence Research (CAiRE)
Hong Kong University of Science and Technology
`zjiad@connect.ust.hk, pascale@ece.ust.hk`



Medical Question

What causes Noonan syndrome?

Acquire Knowledge

Prompt: Provide background knowledge to answer the given question: "What causes Noonan syndrome?"

Refine Knowledge

Prompt: The factuality score for the knowledge is -1.2 (less than THRESHOLD_FACTUAL), which means the knowledge is not strongly supported by empirical evidence. Please refine the knowledge to improve its factuality.

Generated Knowledge

Noonan syndrome is a genetic disorder that is caused by a mutation in a specific gene called the "papillon gene."....

Try Answering

Prompt: Refer to the knowledge: "{final_knowledge}" and answer the question: "What causes Noonan syndrome?" with one paragraph.

Refine Answer

Prompt: The consistency score for the knowledge is -5.1 (less than THRESHOLD_CONS), which means the alignment and consistency between response and knowledge are low. Please refine the response to improve its consistency.

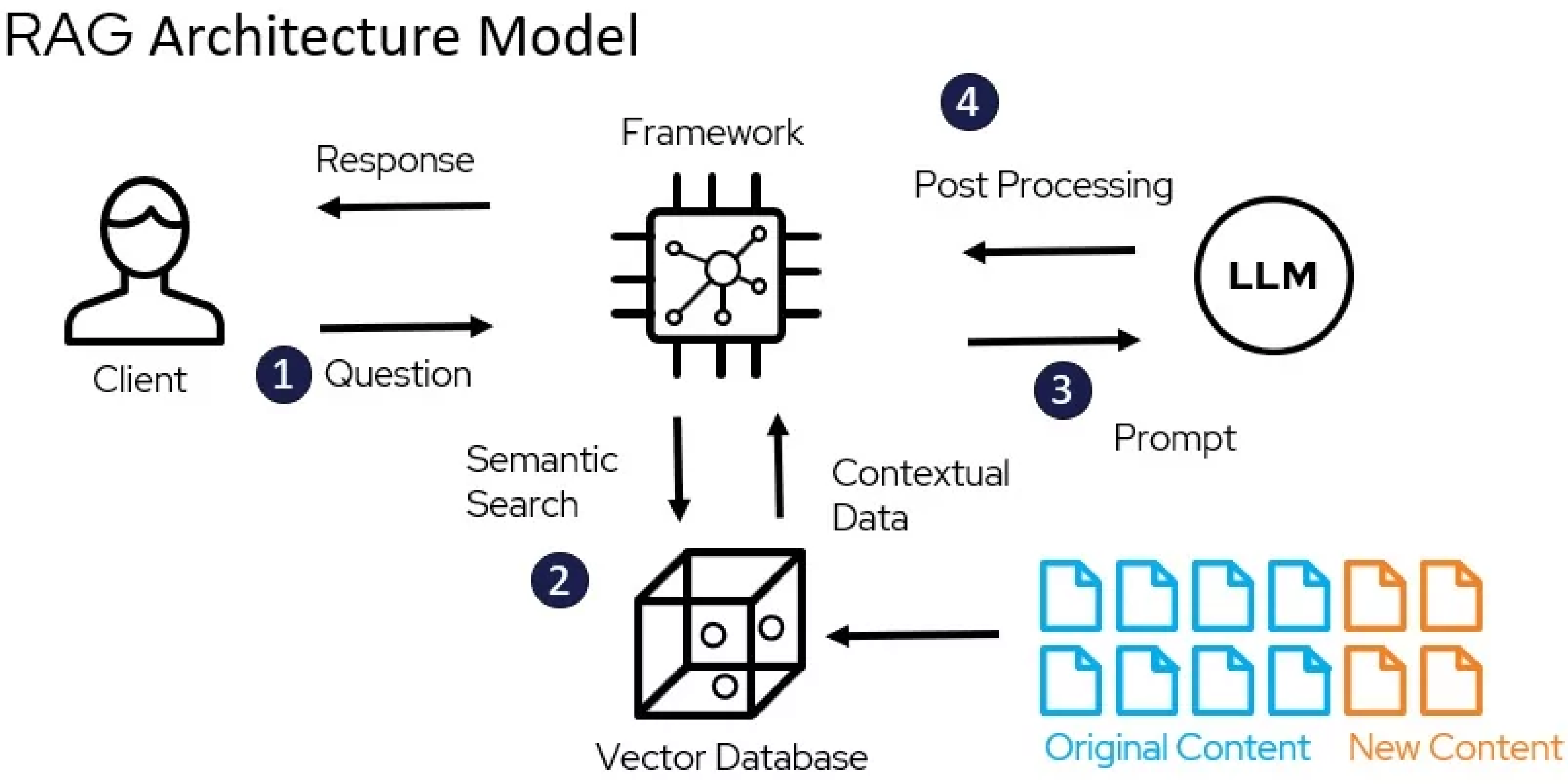
Generated Answer

Noonan syndrome is caused by mutations or changes in certain genes....

Final Answer

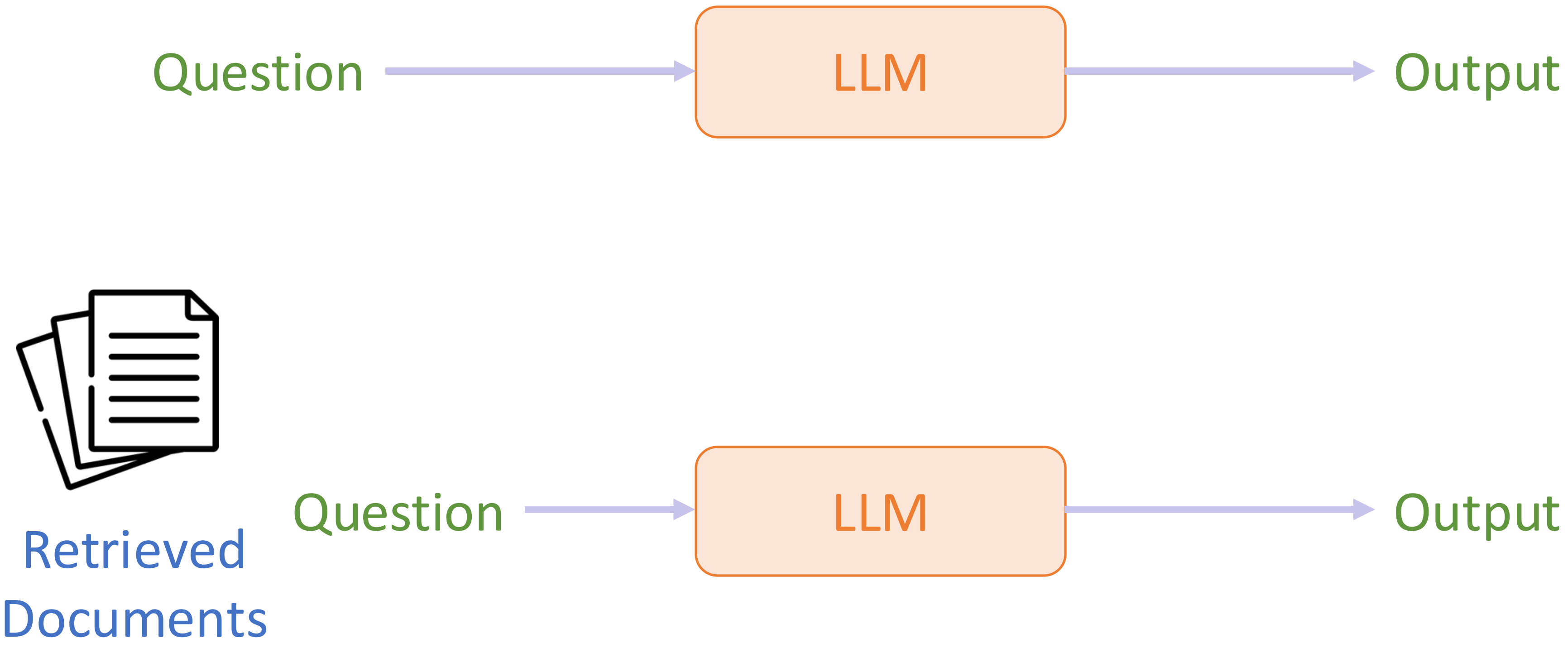
Noonan syndrome is caused by a genetic mutation on chromosome 12...

Recap: Retrieval-Augmented Generation (RAG)



https://pureinsights.com/blog/2023/what-is-retrieval-augmented-generation-rag/

From Factuality to Faithfulness

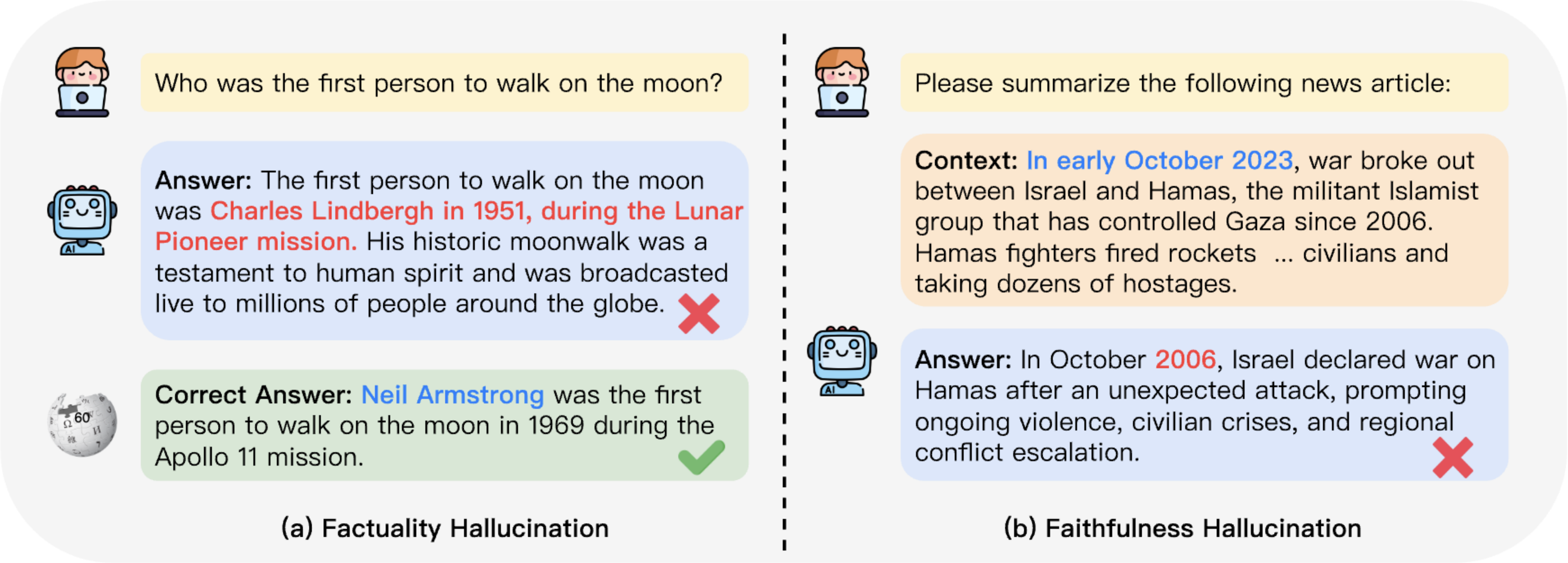


Question → LLM → Output

Retrieved Documents

Question → LLM → Output

Recap: Factuality vs. Faithfulness



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

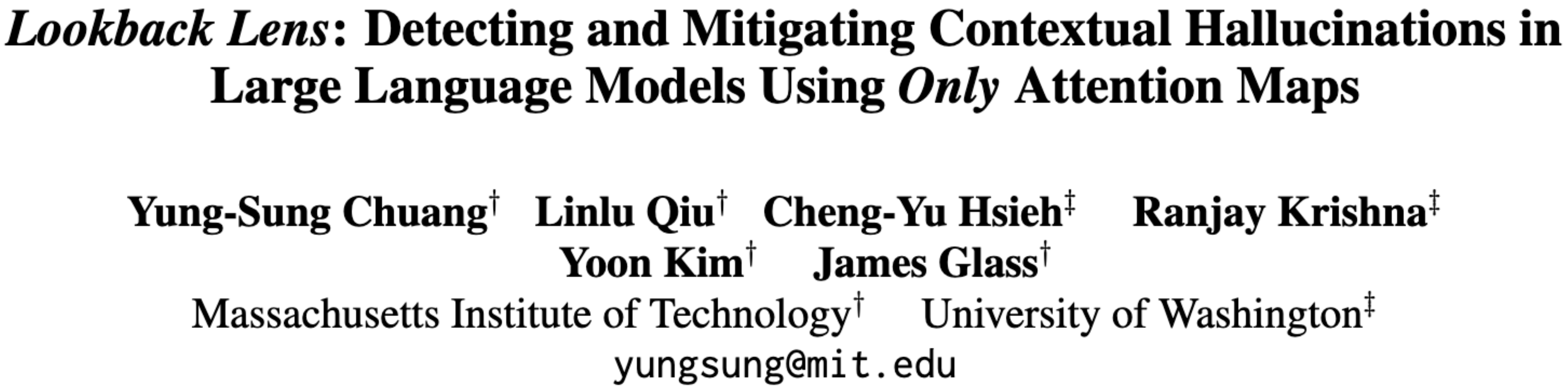
Please summarize the following news article:

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

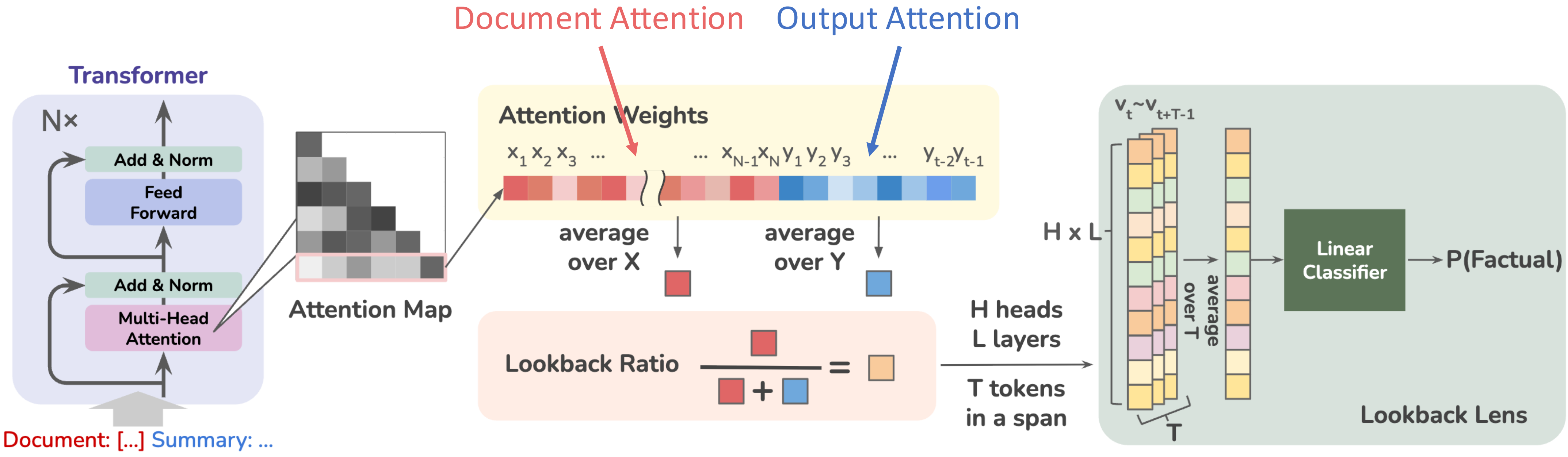
A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions, 2023



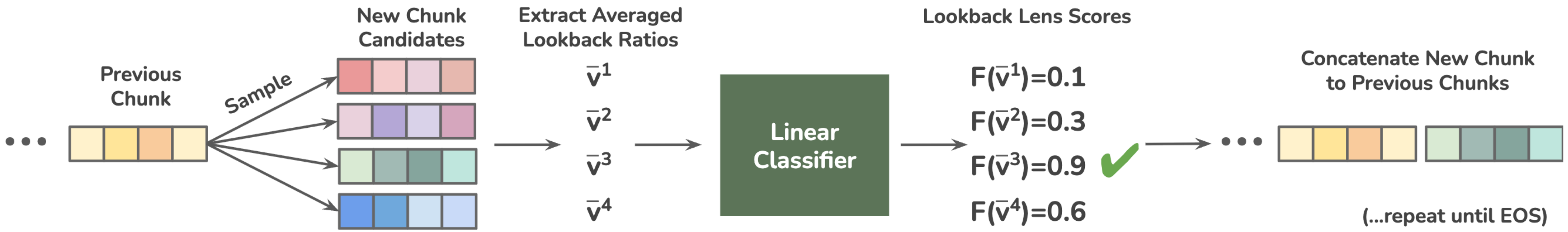
Lookback Lens: Detecting and Mitigating Contextual Hallucinations in Large Language Models Using *Only* Attention Maps

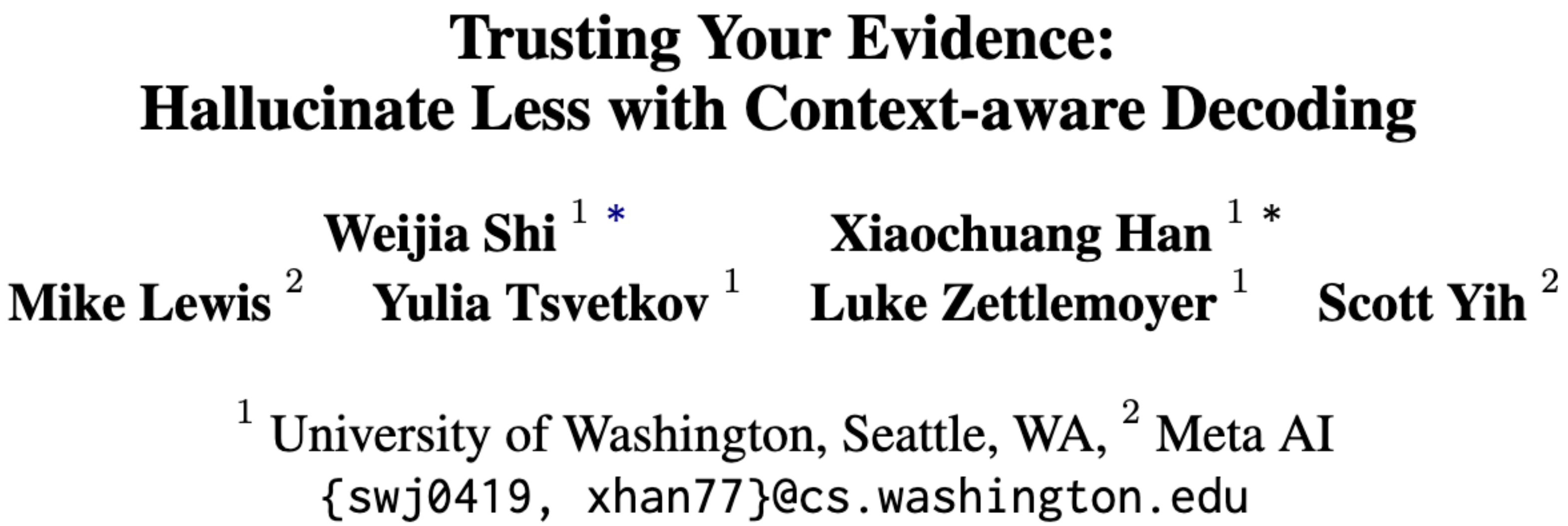
Yung-Sung Chuang[†] **Linlu Qiu**[†] **Cheng-Yu Hsieh**[‡] **Ranjay Krishna**[‡]
Yoon Kim[†] **James Glass**[†]
Massachusetts Institute of Technology[†] University of Washington[‡]
`yungsung@mit.edu`

Lookback Ratio



Lookback Lens Guided Decoding



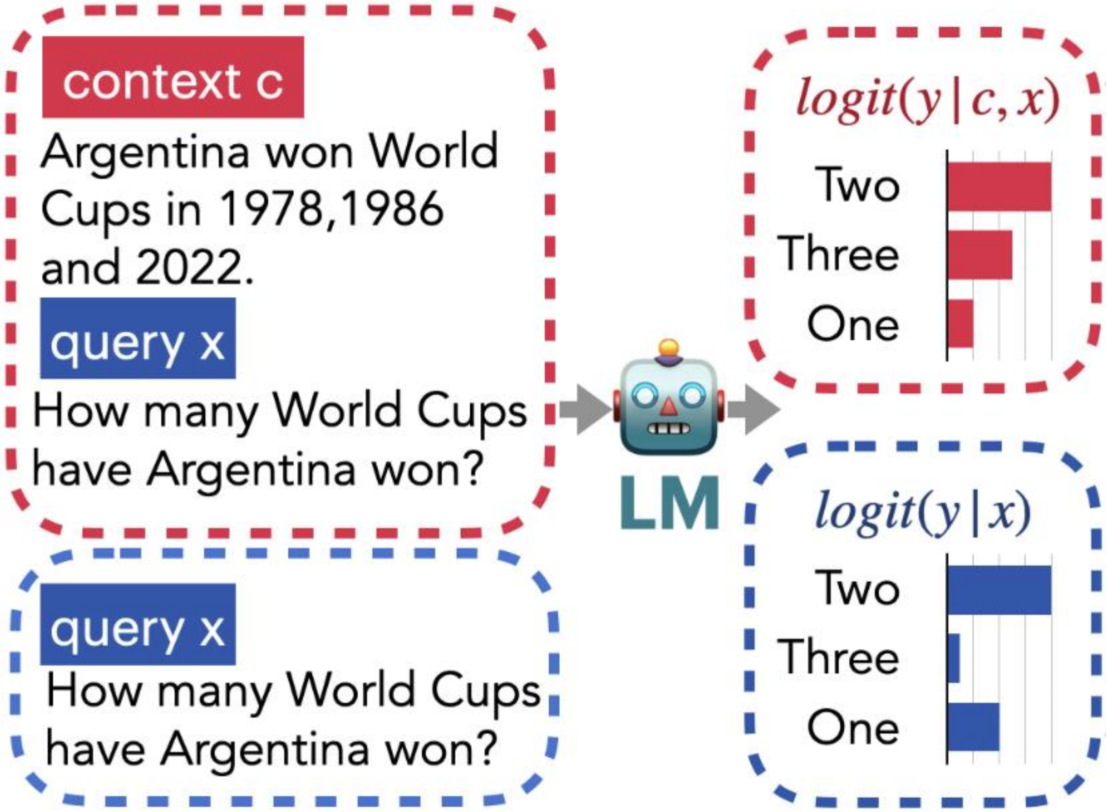


Trusting Your Evidence:
Hallucinate Less with Context-aware Decoding

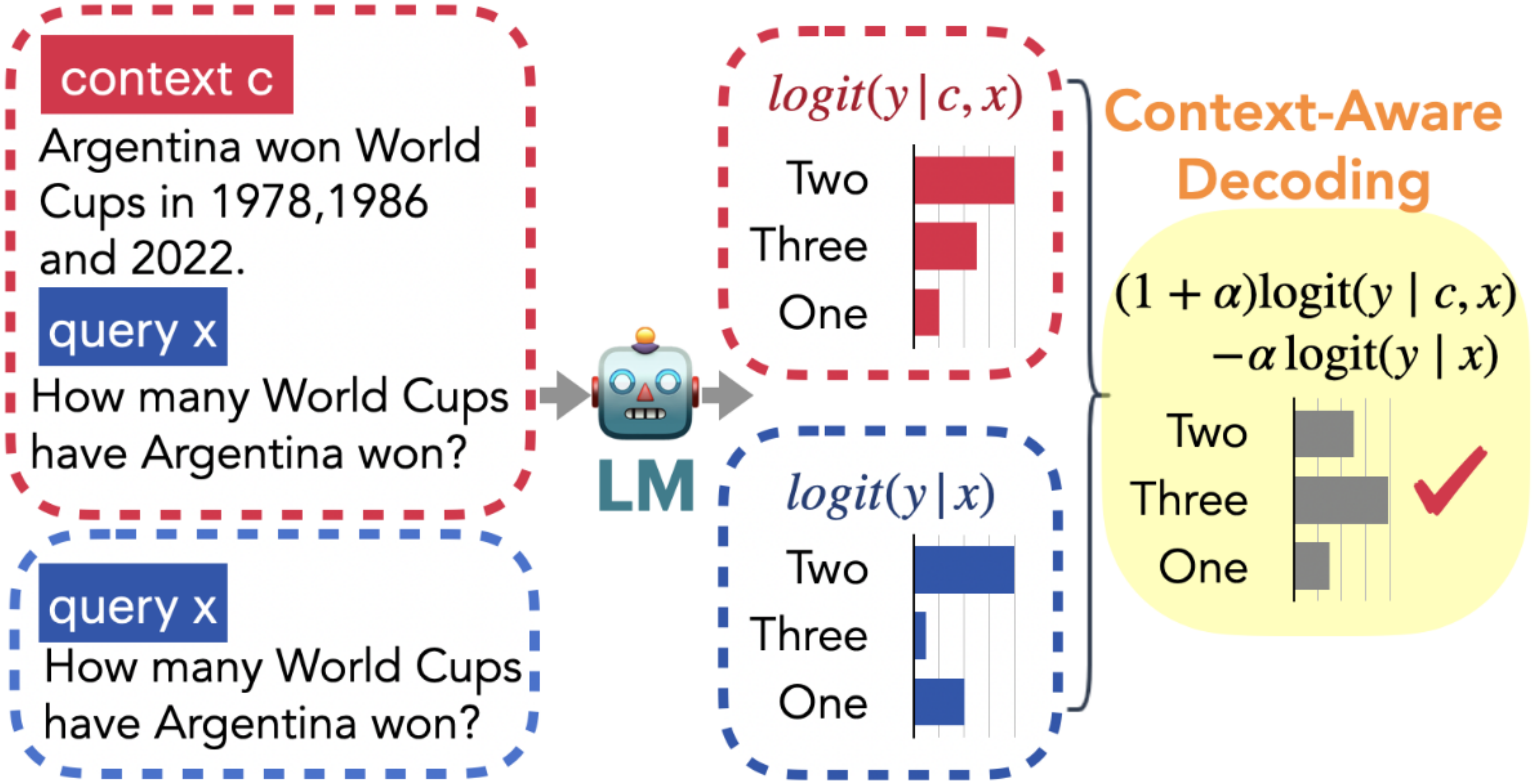
Weijia Shi [1] * **Xiaochuang Han** [1] *
Mike Lewis [2] **Yulia Tsvetkov** [1] **Luke Zettlemoyer** [1] **Scott Yih** [2]

[1] University of Washington, Seattle, WA, [2] Meta AI
`{swj0419, xhan77}@cs.washington.edu`

Context-Aware Decoding

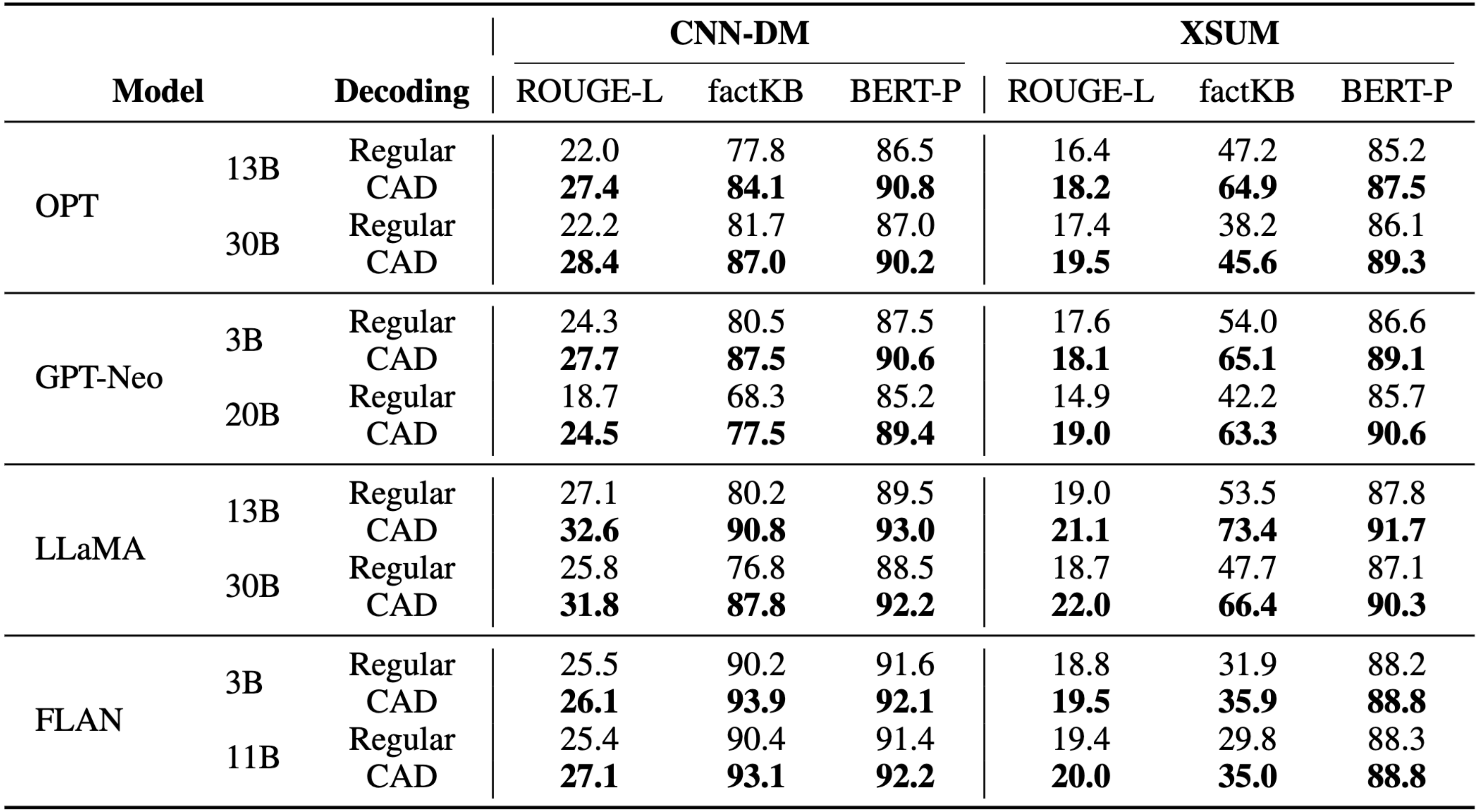

context c
Argentina won World Cups in 1978,1986 and 2022.
query x
How many World Cups have Argentina won?
query x
How many World Cups have Argentina won?
LM
$logit(y|c,x)$
Two
Three
One
$logit(y|x)$
Two
Three
One

Context-Aware Decoding

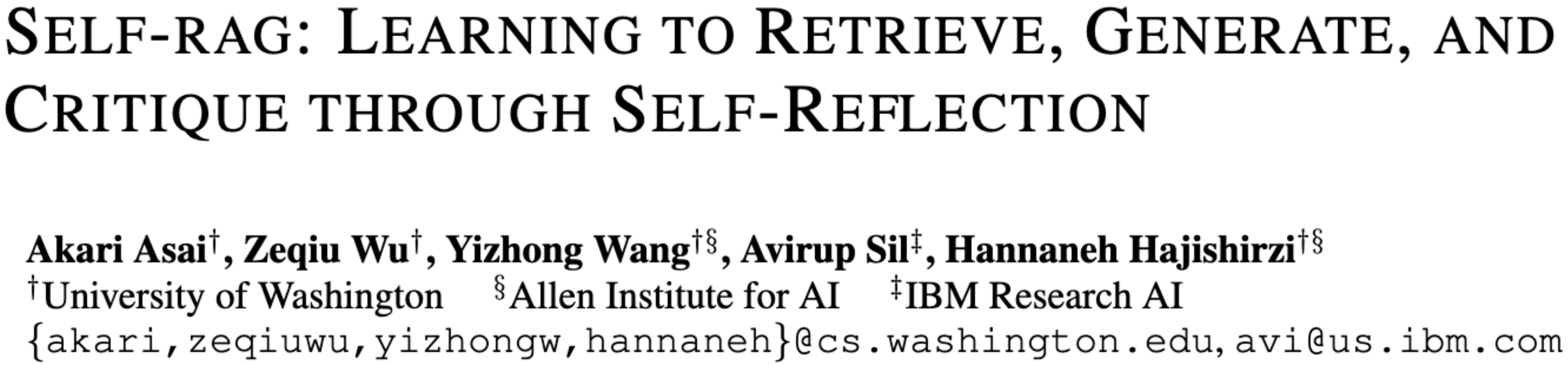

context c
Argentina won World Cups in 1978,1986 and 2022.
query x
How many World Cups have Argentina won?
query x
How many World Cups have Argentina won?
LM
$logit(y|c,x)$
Two
Three
One
$logit(y|x)$
Two
Three
One
Context-Aware Decoding
$(1+\alpha)\text{logit}(y \mid c,x)$
$-\alpha\,\text{logit}(y \mid x)$
Two
Three
One

$$y_t \sim \text{softmax}[(1+\alpha)\,\text{logit}_\theta(y_t \mid \boldsymbol{c},\boldsymbol{x},\boldsymbol{y}_{<t}) - \alpha\,\text{logit}_\theta(y_t \mid \boldsymbol{x},\boldsymbol{y}_{<t})]$$

Results



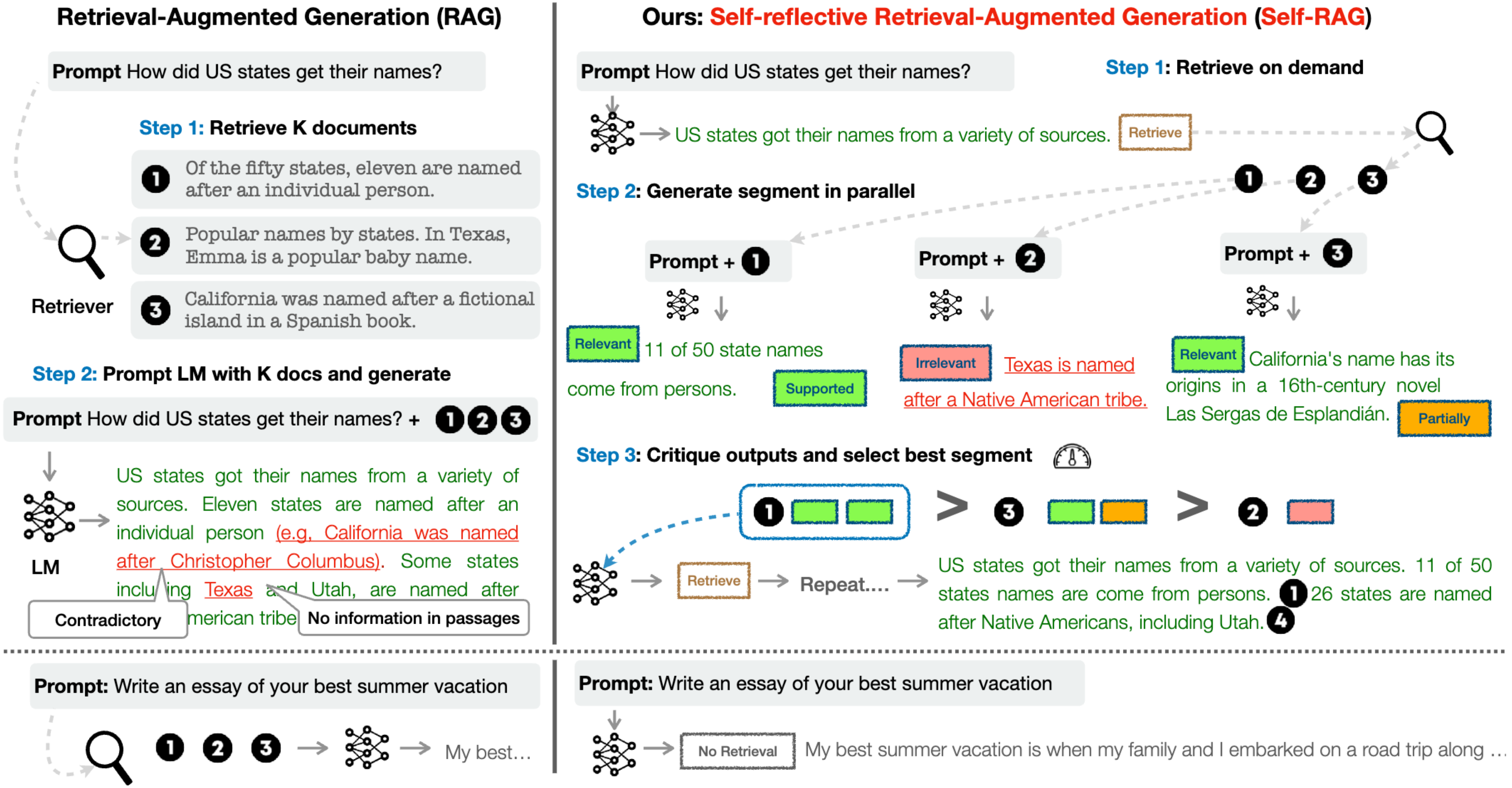
Model		**Decoding**	**CNN-DM**			**XSUM**		
			ROUGE-L	factKB	BERT-P	ROUGE-L	factKB	BERT-P
OPT	13B	Regular	22.0	77.8	86.5	16.4	47.2	85.2
		CAD	**27.4**	**84.1**	**90.8**	**18.2**	**64.9**	**87.5**
	30B	Regular	22.2	81.7	87.0	17.4	38.2	86.1
		CAD	**28.4**	**87.0**	**90.2**	**19.5**	**45.6**	**89.3**
GPT-Neo	3B	Regular	24.3	80.5	87.5	17.6	54.0	86.6
		CAD	**27.7**	**87.5**	**90.6**	**18.1**	**65.1**	**89.1**
	20B	Regular	18.7	68.3	85.2	14.9	42.2	85.7
		CAD	**24.5**	**77.5**	**89.4**	**19.0**	**63.3**	**90.6**
LLaMA	13B	Regular	27.1	80.2	89.5	19.0	53.5	87.8
		CAD	**32.6**	**90.8**	**93.0**	**21.1**	**73.4**	**91.7**
	30B	Regular	25.8	76.8	88.5	18.7	47.7	87.1
		CAD	**31.8**	**87.8**	**92.2**	**22.0**	**66.4**	**90.3**
FLAN	3B	Regular	25.5	90.2	91.6	18.8	31.9	88.2
		CAD	**26.1**	**93.9**	**92.1**	**19.5**	**35.9**	**88.8**
	11B	Regular	25.4	90.4	91.4	19.4	29.8	88.3
		CAD	**27.1**	**93.1**	**92.2**	**20.0**	**35.0**	**88.8**



SELF-RAG: LEARNING TO RETRIEVE, GENERATE, AND CRITIQUE THROUGH SELF-REFLECTION

Akari Asai[†], Zeqiu Wu[†], Yizhong Wang[†§], Avirup Sil[‡], Hannaneh Hajishirzi[†§]
[†]University of Washington [§]Allen Institute for AI [‡]IBM Research AI
`{akari,zeqiuwu,yizhongw,hannaneh}@cs.washington.edu`, `avi@us.ibm.com`

Self-Reflective Retrieval-Augmented Generation



Retrieval-Augmented Generation (RAG)

Prompt How did US states get their names?

Step 1: Retrieve K documents

Retriever

❶ Of the fifty states, eleven are named after an individual person.

❷ Popular names by states. In Texas, Emma is a popular baby name.

❸ California was named after a fictional island in a Spanish book.

Step 2: Prompt LM with K docs and generate

Prompt How did US states get their names? + ❶❷❸

LM

US states got their names from a variety of sources. Eleven states are named after an individual person (e.g, California was named after Christopher Columbus). Some states including Texas and Utah, are named after Native American tribe.

Contradictory

No information in passages

Ours: Self-reflective Retrieval-Augmented Generation (Self-RAG)

Prompt How did US states get their names?

Step 1: Retrieve on demand

US states got their names from a variety of sources. Retrieve

❶ ❷ ❸

Step 2: Generate segment in parallel

Prompt + ❶

Relevant 11 of 50 state names come from persons. Supported

Prompt + ❷

Irrelevant Texas is named after a Native American tribe.

Prompt + ❸

Relevant California's name has its origins in a 16th-century novel Las Sergas de Esplandián. Partially

Step 3: Critique outputs and select best segment

❶ 🟩🟩 > ❸ 🟩🟧 > ❷ 🟥

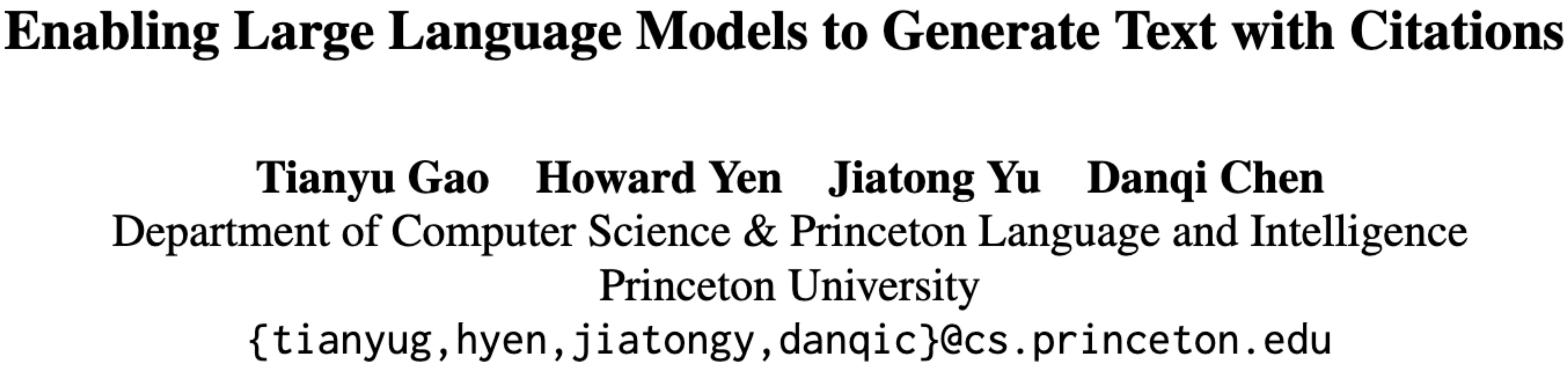
→ Retrieve → Repeat.... → US states got their names from a variety of sources. 11 of 50 states names are come from persons. ❶ 26 states are named after Native Americans, including Utah. ❹

Prompt: Write an essay of your best summer vacation

❶ ❷ ❸ → → My best…

Prompt: Write an essay of your best summer vacation

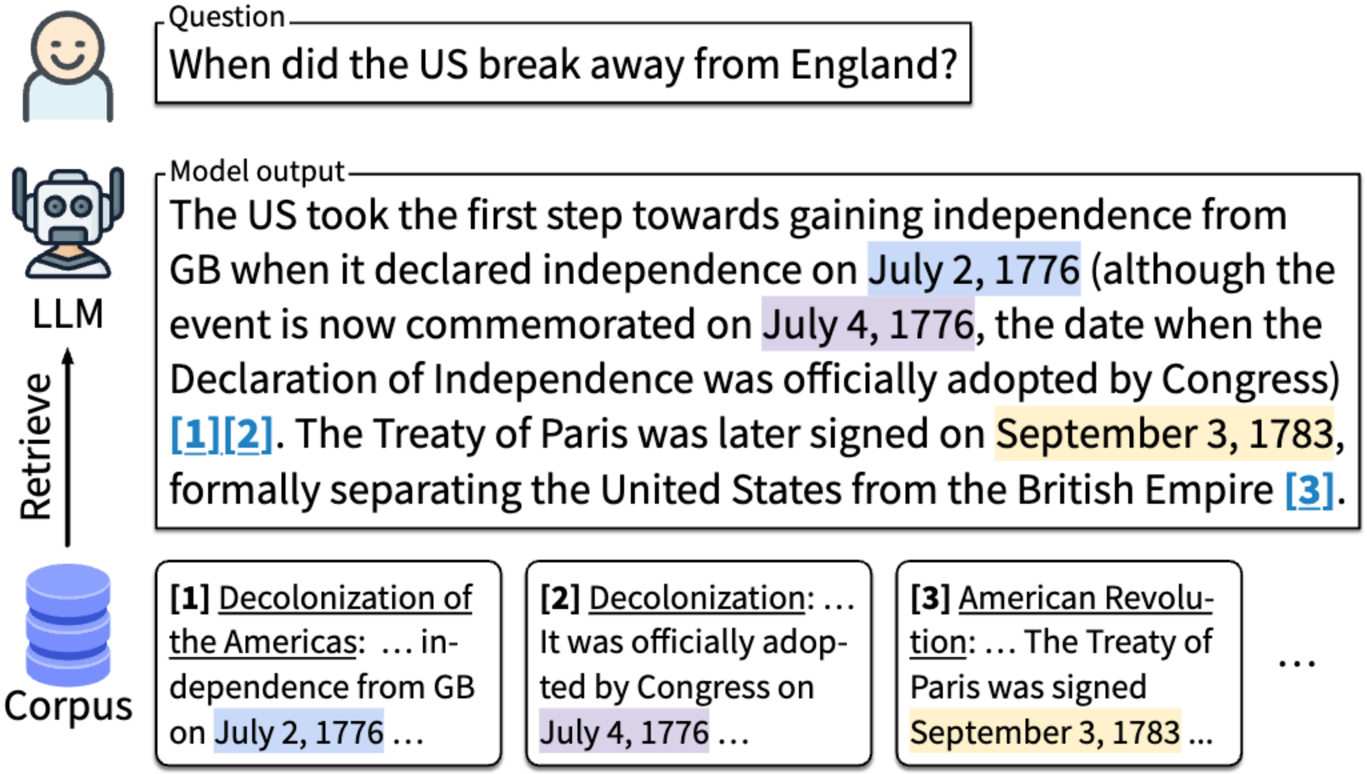
→ No Retrieval My best summer vacation is when my family and I embarked on a road trip along …



Enabling Large Language Models to Generate Text with Citations

Tianyu Gao **Howard Yen** **Jiatong Yu** **Danqi Chen**
Department of Computer Science & Princeton Language and Intelligence
Princeton University
`{tianyug,hyen,jiatongy,danqic}@cs.princeton.edu`

Grounded Generation



When did the US break away from England?

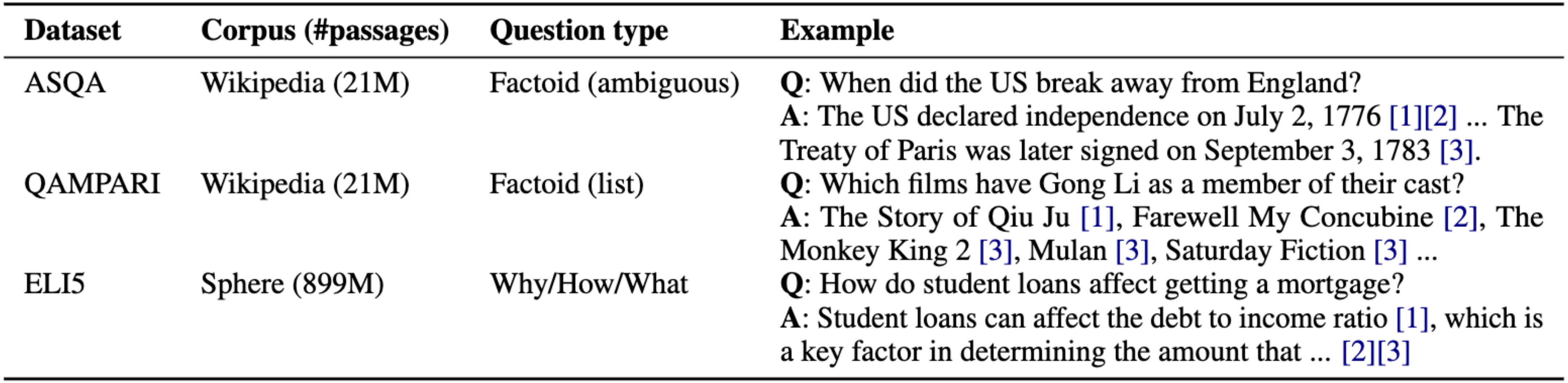
The US took the first step towards gaining independence from GB when it declared independence on July 2, 1776 (although the event is now commemorated on July 4, 1776, the date when the Declaration of Independence was officially adopted by Congress) [1][2]. The Treaty of Paris was later signed on September 3, 1783, formally separating the United States from the British Empire [3].

[1] Decolonization of the Americas: … independence from GB on July 2, 1776 …

[2] Decolonization: … It was officially adopted by Congress on July 4, 1776 …

[3] American Revolution: … The Treaty of Paris was signed September 3, 1783 ...

Examples

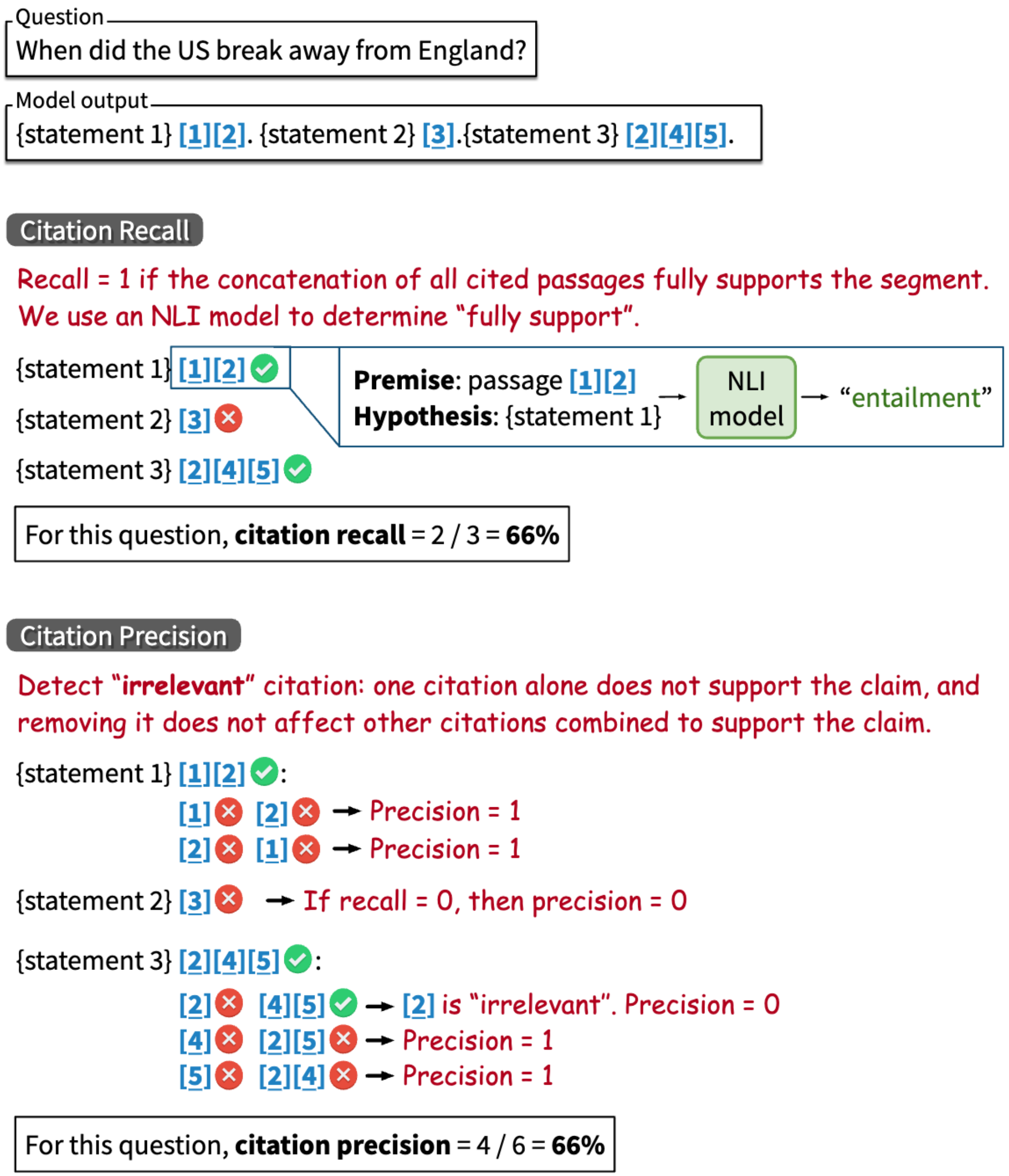


Dataset	Corpus (#passages)	Question type	Example
ASQA	Wikipedia (21M)	Factoid (ambiguous)	**Q**: When did the US break away from England? **A**: The US declared independence on July 2, 1776 [1][2] ... The Treaty of Paris was later signed on September 3, 1783 [3].
QAMPARI	Wikipedia (21M)	Factoid (list)	**Q**: Which films have Gong Li as a member of their cast? **A**: The Story of Qiu Ju [1], Farewell My Concubine [2], The Monkey King 2 [3], Mulan [3], Saturday Fiction [3] ...
ELI5	Sphere (899M)	Why/How/What	**Q**: How do student loans affect getting a mortgage? **A**: Student loans can affect the debt to income ratio [1], which is a key factor in determining the amount that ... [2][3]

Evaluation Goal

- **Fluency**: whether the model's generated text is fluent and coherent
- **Correctness**: whether the answer is accurate and covers all aspects of interest
- **Citation quality**: whether the answer is well supported by the cited passages and no irrelevant passages are cited

Citation Quality



Question
When did the US break away from England?

Model output
{statement 1} [1][2]. {statement 2} [3].{statement 3} [2][4][5].

Citation Recall

Recall = 1 if the concatenation of all cited passages fully supports the segment.
We use an NLI model to determine "fully support".

{statement 1} [1][2] ✓
{statement 2} [3] ✗
{statement 3} [2][4][5] ✓

Premise: passage [1][2] → NLI model → "entailment"
Hypothesis: {statement 1}

For this question, **citation recall** = 2 / 3 = **66%**

Citation Precision

Detect "**irrelevant**" citation: one citation alone does not support the claim, and removing it does not affect other citations combined to support the claim.

{statement 1} [1][2] ✓:
[1] ✗ [2] ✗ → Precision = 1
[2] ✗ [1] ✗ → Precision = 1

{statement 2} [3] ✗ → If recall = 0, then precision = 0

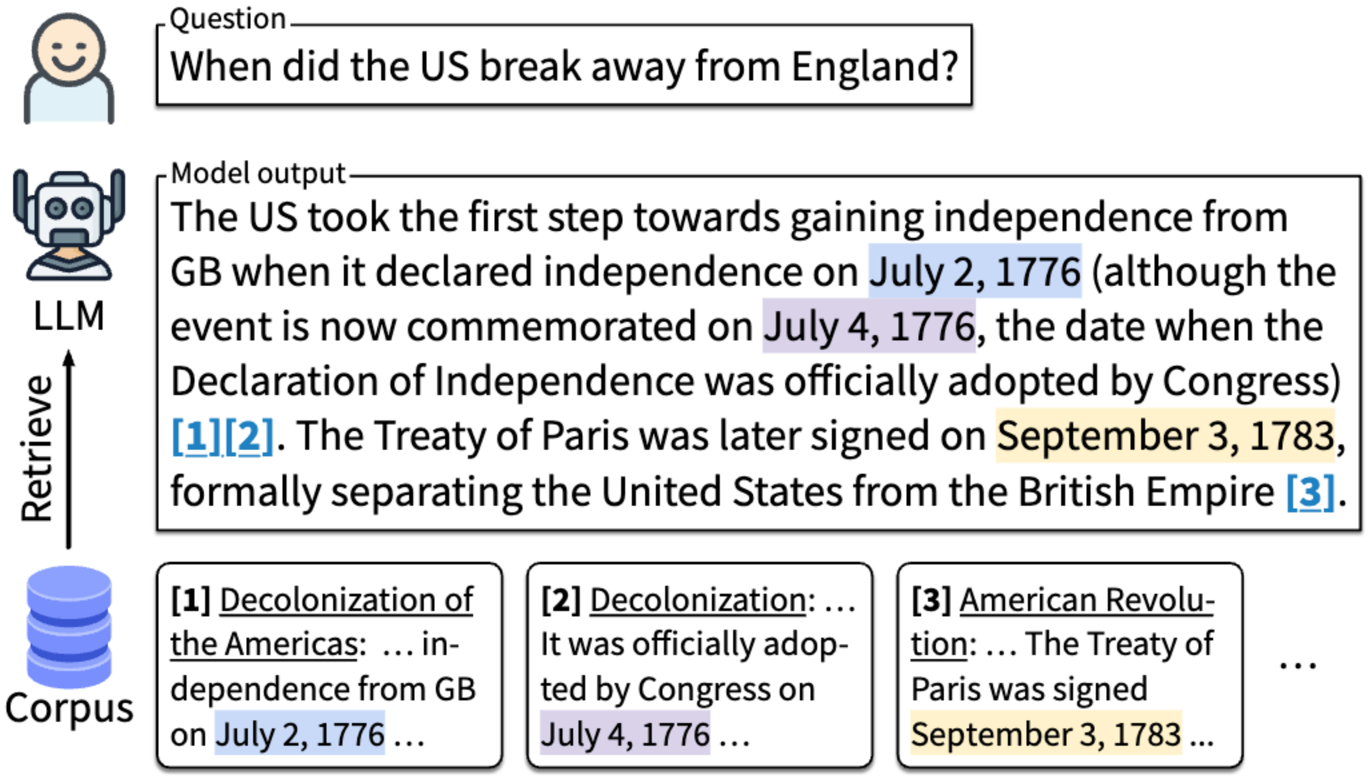
{statement 3} [2][4][5] ✓:
[2] ✗ [4][5] ✓ → [2] is "irrelevant". Precision = 0
[4] ✗ [2][5] ✗ → Precision = 1
[5] ✗ [2][4] ✗ → Precision = 1

For this question, **citation precision** = 4 / 6 = **66%**

CaLM: Contrasting Large and Small Language Models to Verify Grounded Generation

I-Hung Hsu[1]*, **Zifeng Wang**[2], **Long T. Le**[2], **Lesly Miculicich**[2],
Nanyun Peng[3], **Chen-Yu Lee**[2], **Tomas Pfister**[2]
[1]University of Southern California, [2]Google Cloud AI Research,
[3]University of California, Los Angeles

Grounded Generation



When did the US break away from England?

The US took the first step towards gaining independence from GB when it declared independence on July 2, 1776 (although the event is now commemorated on July 4, 1776, the date when the Declaration of Independence was officially adopted by Congress) [1][2]. The Treaty of Paris was later signed on September 3, 1783, formally separating the United States from the British Empire [3].

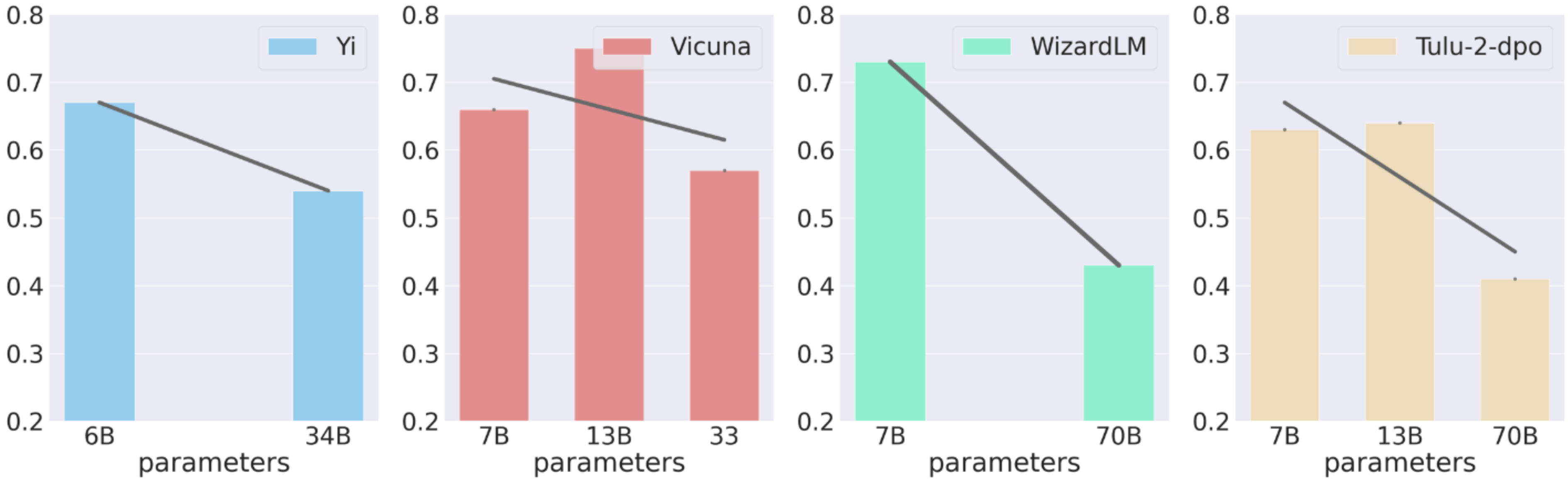
[1] Decolonization of the Americas: … independence from GB on July 2, 1776 …

[2] Decolonization: … It was officially adopted by Congress on July 4, 1776 …

[3] American Revolution: … The Treaty of Paris was signed September 3, 1783 ...

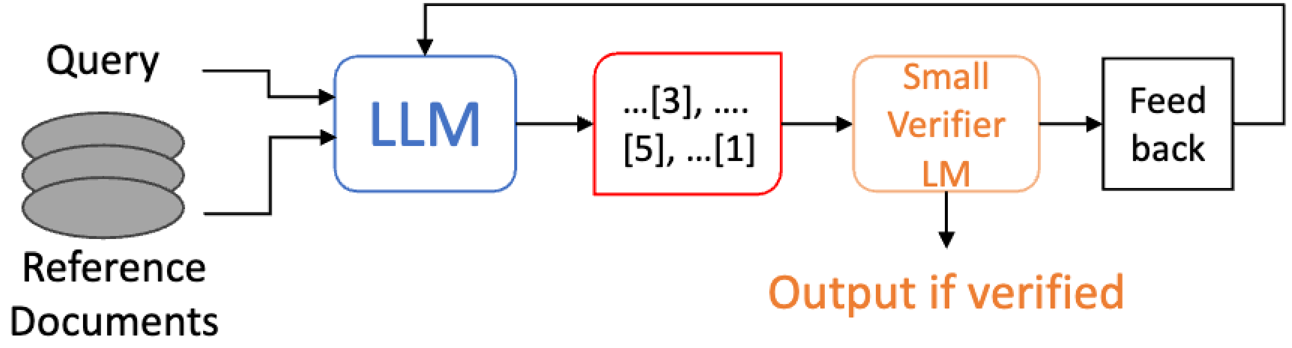
…

Sensitivity to Retrieved Documents



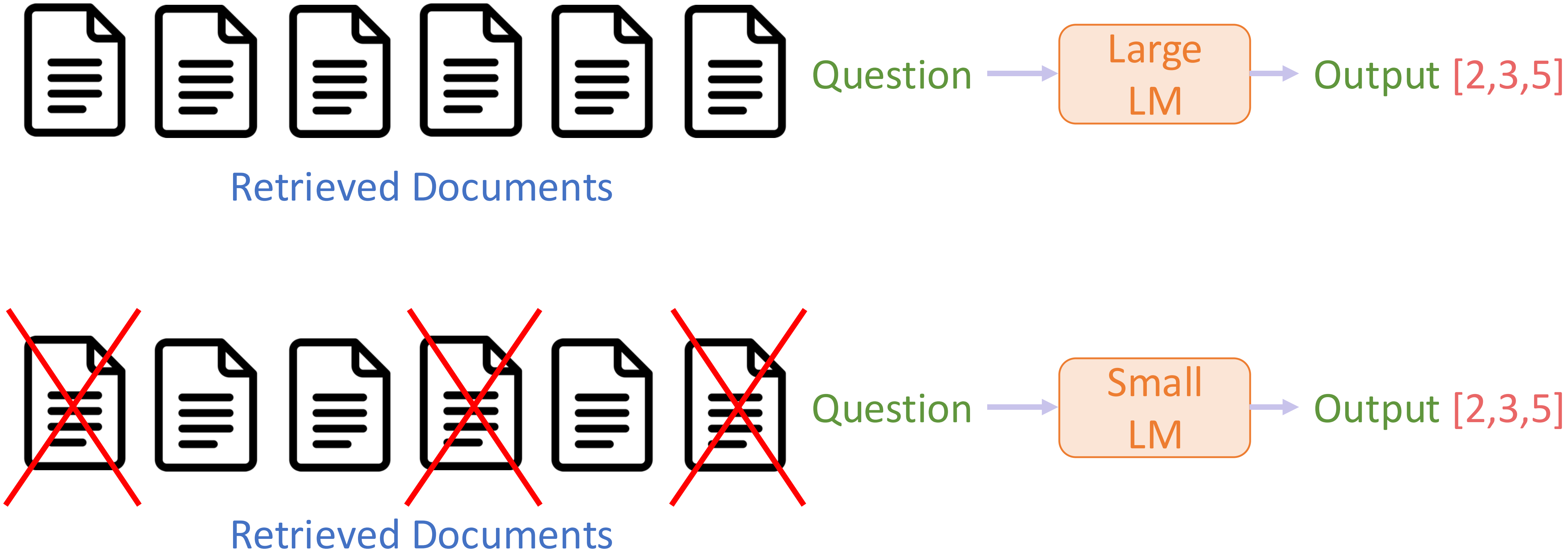
Small LMs are more sensitive to documents

Overview

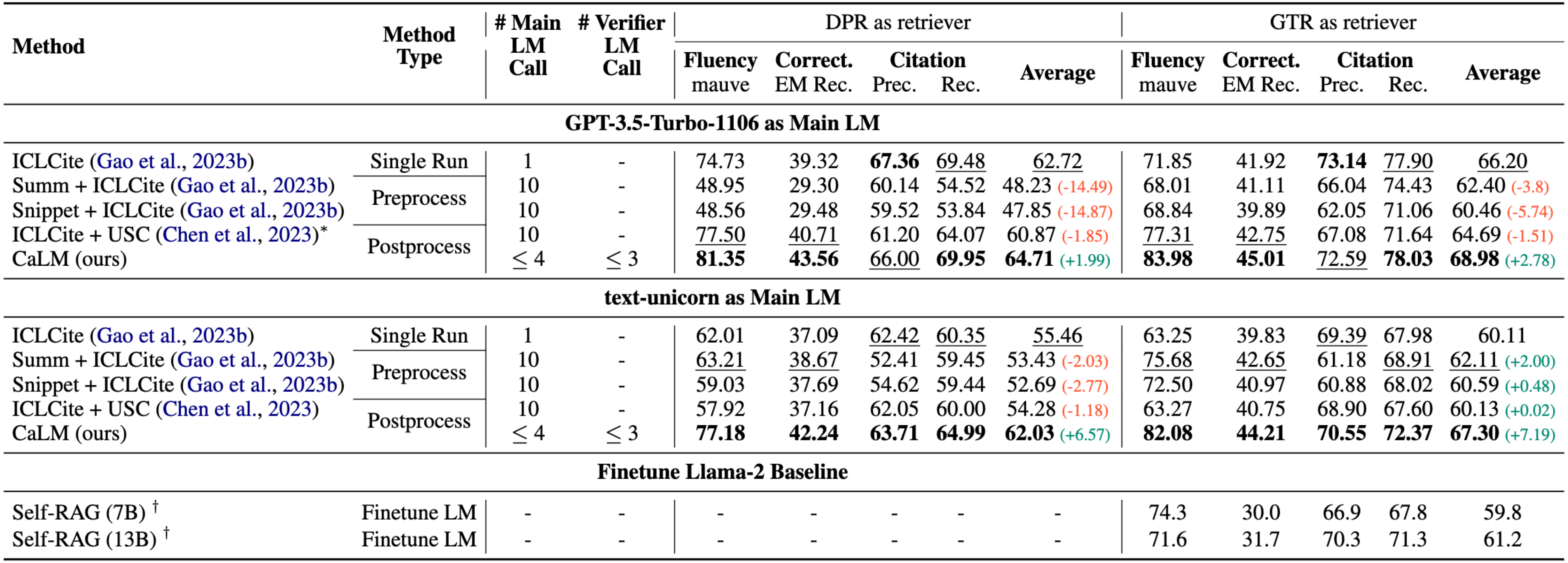

(c) Ours: Post verification on LLM's output ensure output quality

Key idea: using small LMs as verifiers

Key Idea


Question
Large LM
Output [2,3,5]
Retrieved Documents
Question
Small LM
Output [2,3,5]
Retrieved Documents

Results



Method	**Method Type**	**# Main LM Call**	**# Verifier LM Call**	DPR as retriever					GTR as retriever				
				Fluency mauve	**Correct.** EM Rec.	**Citation** Prec.	**Citation** Rec.	**Average**	**Fluency** mauve	**Correct.** EM Rec.	**Citation** Prec.	**Citation** Rec.	**Average**
GPT-3.5-Turbo-1106 as Main LM													
ICLCite (Gao et al., 2023b)	Single Run	1	-	74.73	39.32	**67.36**	69.48	62.72	71.85	41.92	**73.14**	77.90	66.20
Summ + ICLCite (Gao et al., 2023b)	Preprocess	10	-	48.95	29.30	60.14	54.52	48.23 (-14.49)	68.01	41.11	66.04	74.43	62.40 (-3.8)
Snippet + ICLCite (Gao et al., 2023b)		10	-	48.56	29.48	59.52	53.84	47.85 (-14.87)	68.84	39.89	62.05	71.06	60.46 (-5.74)
ICLCite + USC (Chen et al., 2023)*	Postprocess	10	-	77.50	40.71	61.20	64.07	60.87 (-1.85)	77.31	42.75	67.08	71.64	64.69 (-1.51)
CaLM (ours)		≤ 4	≤ 3	**81.35**	**43.56**	66.00	**69.95**	**64.71** (+1.99)	**83.98**	**45.01**	72.59	**78.03**	**68.98** (+2.78)
text-unicorn as Main LM													
ICLCite (Gao et al., 2023b)	Single Run	1	-	62.01	37.09	62.42	60.35	55.46	63.25	39.83	69.39	67.98	60.11
Summ + ICLCite (Gao et al., 2023b)	Preprocess	10	-	63.21	38.67	52.41	59.45	53.43 (-2.03)	75.68	42.65	61.18	68.91	62.11 (+2.00)
Snippet + ICLCite (Gao et al., 2023b)		10	-	59.03	37.69	54.62	59.44	52.69 (-2.77)	72.50	40.97	60.88	68.02	60.59 (+0.48)
ICLCite + USC (Chen et al., 2023)	Postprocess	10	-	57.92	37.16	62.05	60.00	54.28 (-1.18)	63.27	40.75	68.90	67.60	60.13 (+0.02)
CaLM (ours)		≤ 4	≤ 3	**77.18**	**42.24**	**63.71**	**64.99**	**62.03** (+6.57)	**82.08**	**44.21**	**70.55**	**72.37**	**67.30** (+7.19)
Finetune Llama-2 Baseline													
Self-RAG (7B) †	Finetune LM	-	-	-	-	-	-	-	74.3	30.0	66.9	67.8	59.8
Self-RAG (13B) †	Finetune LM	-	-	-	-	-	-	-	71.6	31.7	70.3	71.3	61.2