CSCE 689: Special Topics in Advanced NLP

Lecture 14: Hallucinations

Kuan-Hao Huang

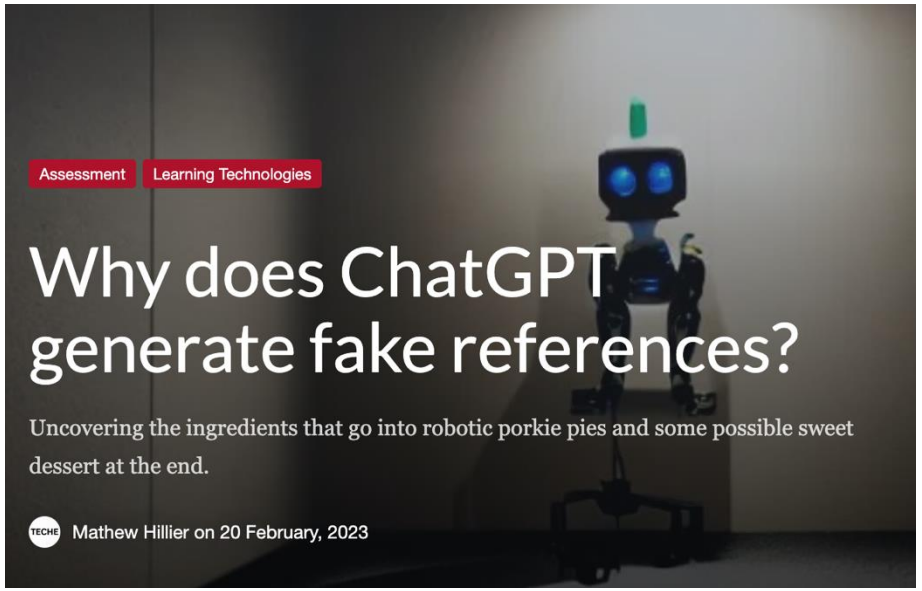
Fall 2026

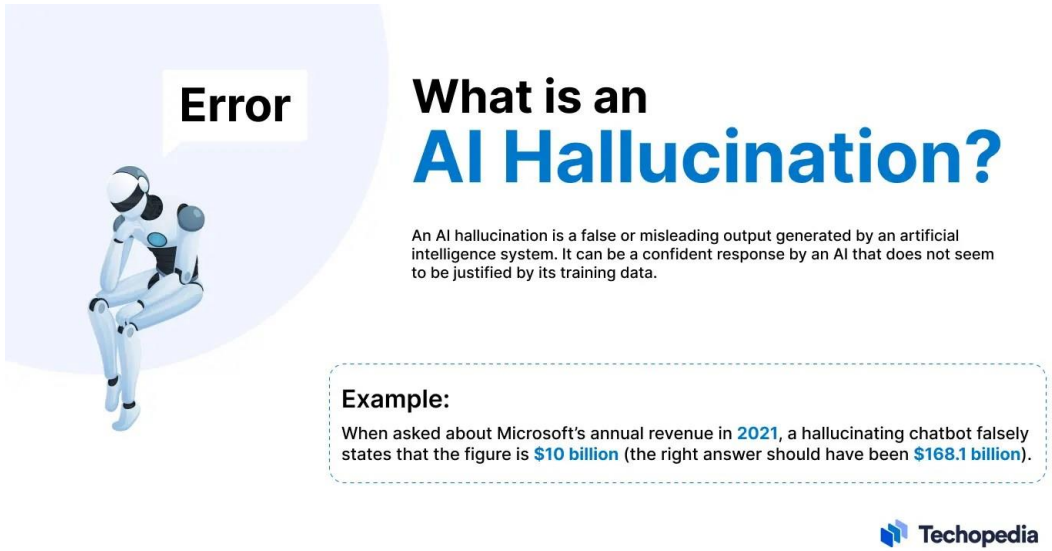


Course Project – Midterm Report

- Due: 11/5
- Page limit: 4 pages (excluding reference)
- Format: ACL style
- The report should include
  - Introduction to the topic you choose and problem definition
  - Related literature and overview of existing progress and challenges
  - Proposed solutions, novelty, and expected contributions
  - Evaluation metrics
  - Planned implementation details, including dataset, models, codebases, etc.
  - Current progress (preliminary results, baseline results, etc.)
  - Next steps and expected timeline

Hallucinations

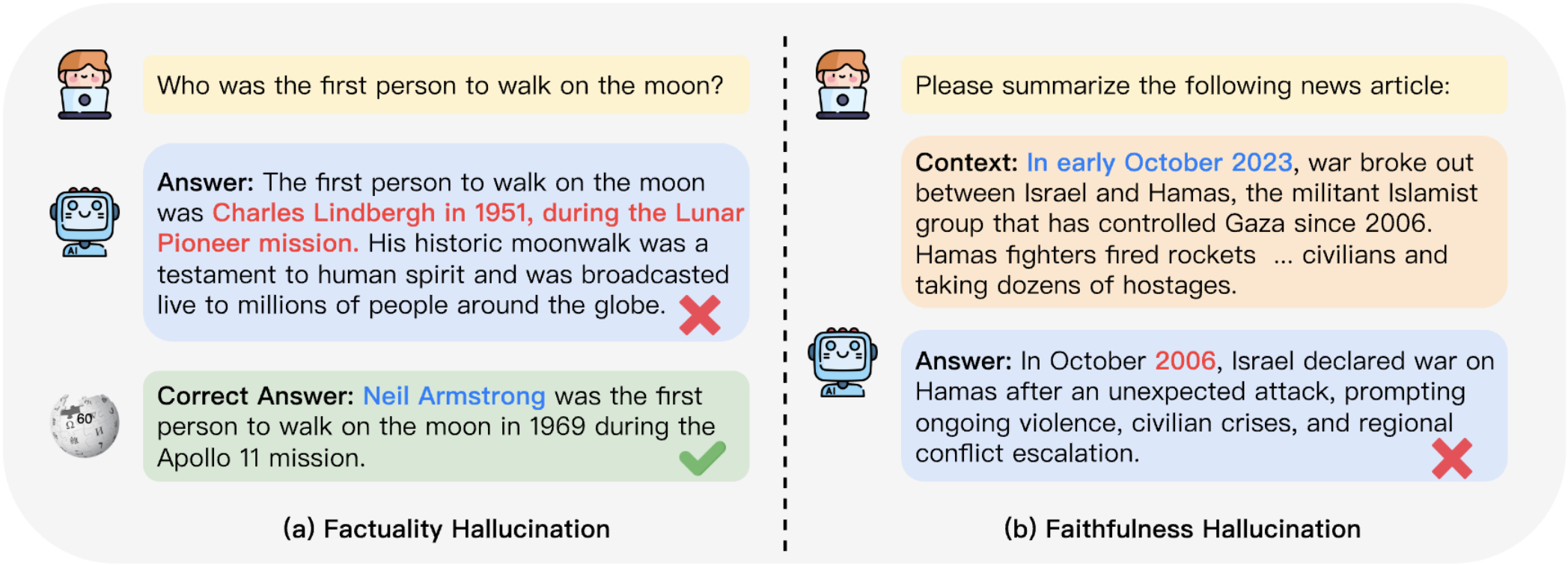




What is Hallucination?

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

What is Hallucination?

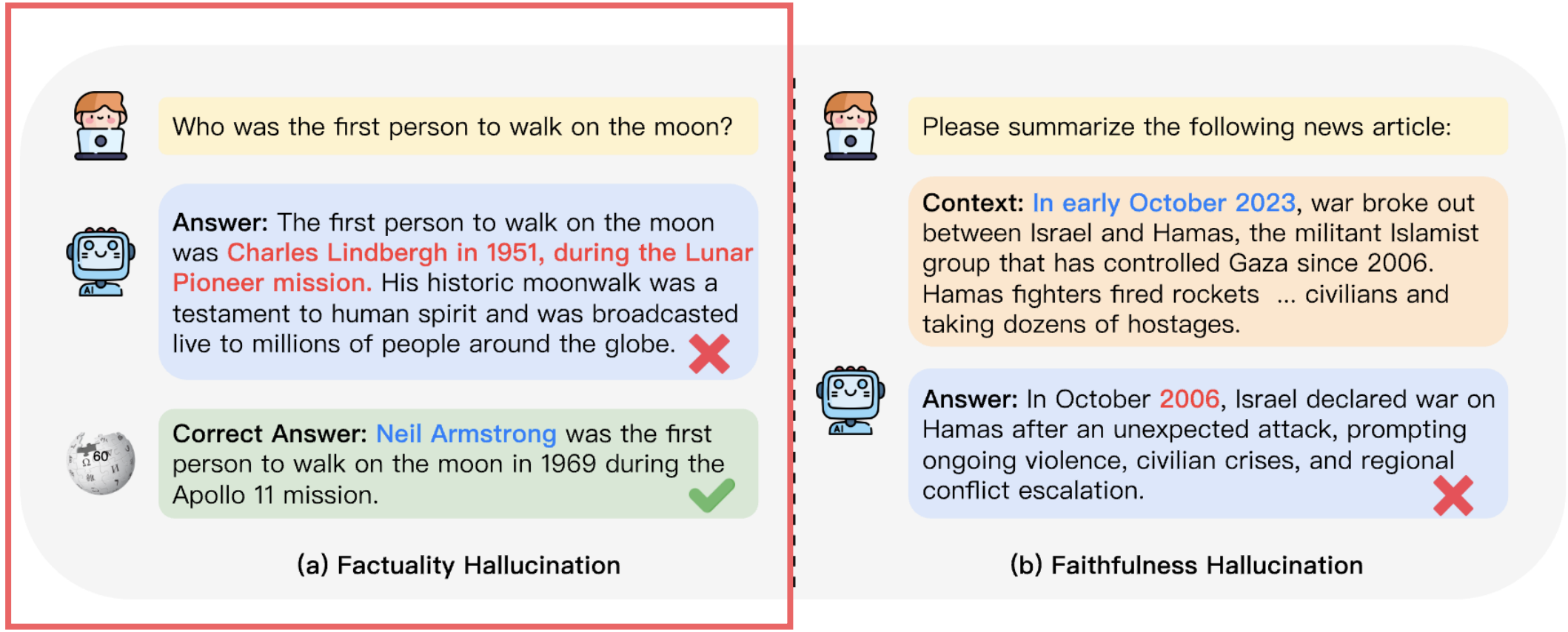


A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions, 2023

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

Types of Hallucinations



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

Step 1: Preparing Prompts

Step 2: Obtaining Generations from the LMeval

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.



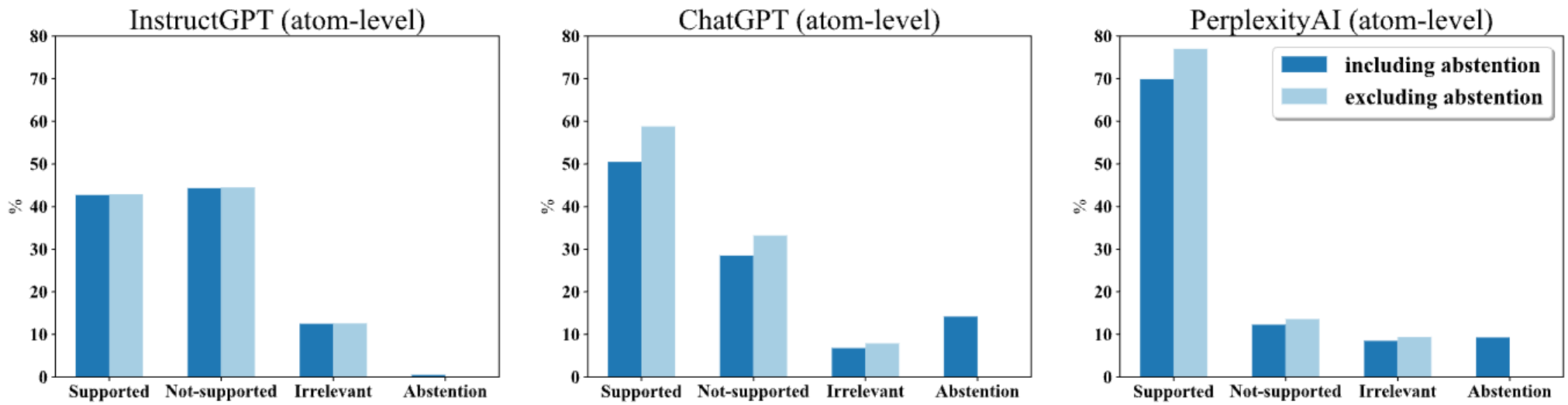
upwork

Irrelevant

Supported

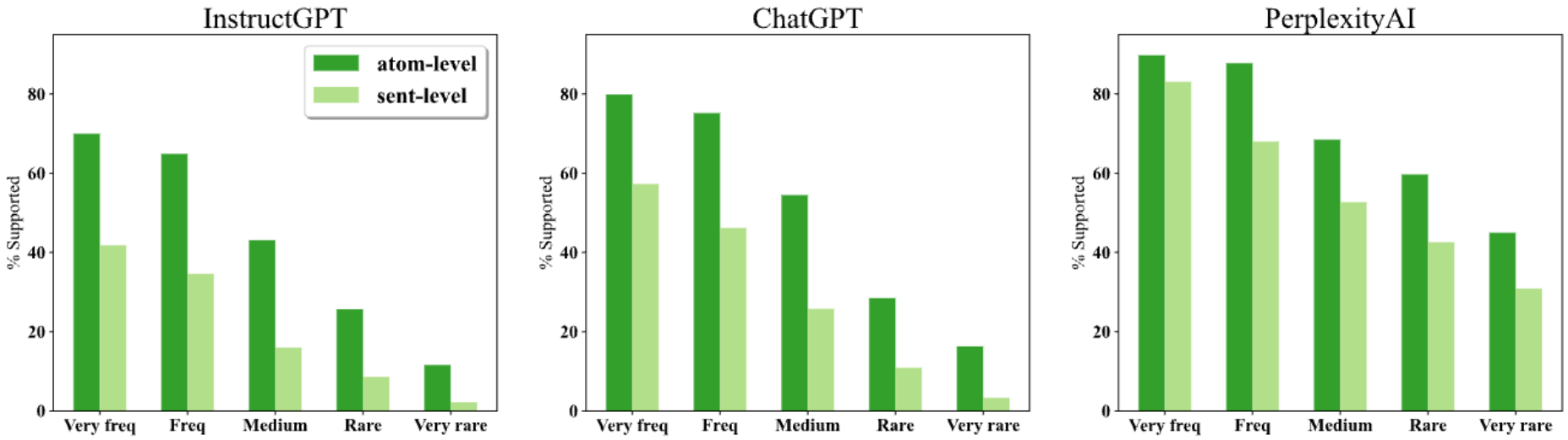
Not-supported

Results



All models make substantial precision errors

Results



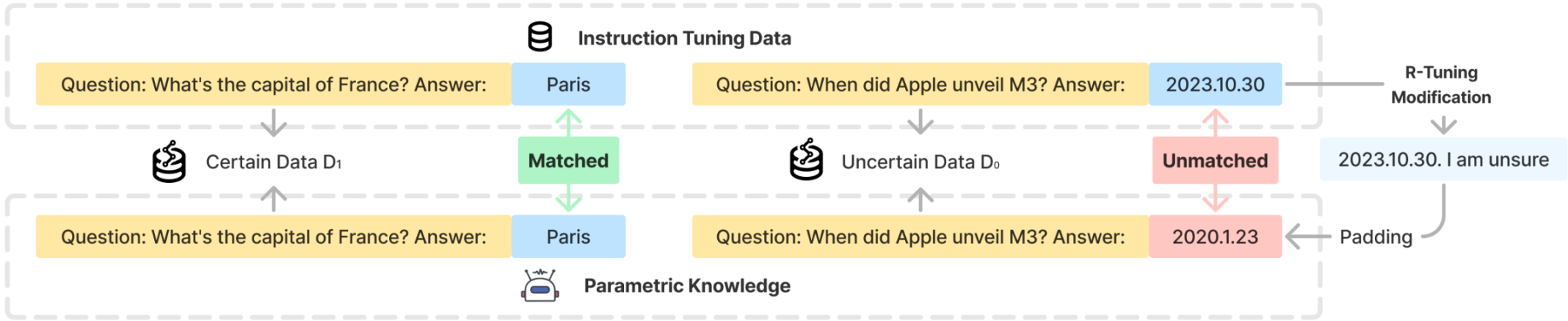
More errors as the rarity of entities increases



R-Tuning: Instructing Large Language Models to Say 'I Don't Know'

Hanning Zhang♠*, Shizhe Diao♠*, Yong Lin♠*, Yi R. Fung♡,
Qing Lian♠, Xingyao Wang♡, Yangyi Chen♡, Heng Ji♡, Tong Zhang♡
♠The Hong Kong University of Science and Technology
♡University of Illinois Urbana-Champaign
{hzhangco, sdiaoaa, ylindf, qlianab, tongzhang}@ust.hk
{yifung2, xingyao6, yangyic3, hengji}@illinois.edu

Overview

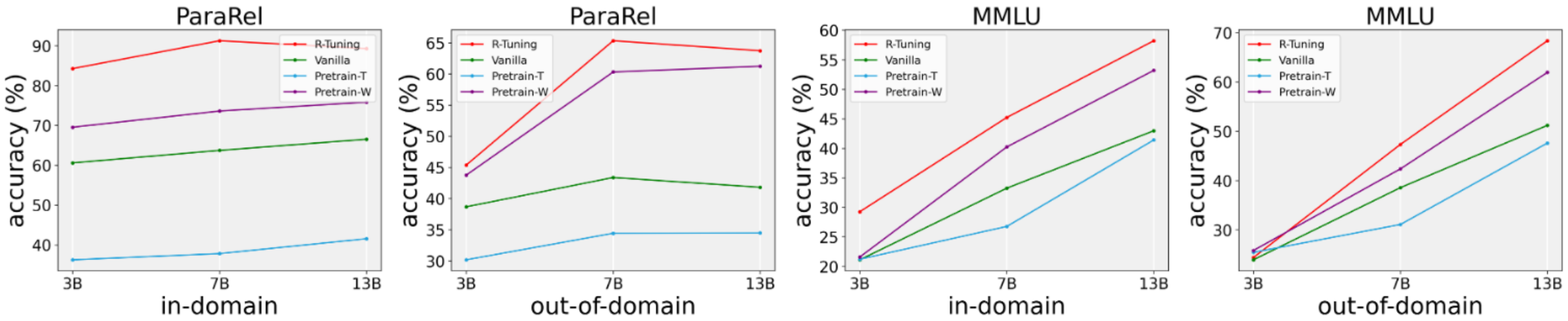

Instruction Tuning Data
Question: What's the capital of France? Answer:
Paris
Question: When did Apple unveil M3? Answer:
2023.10.30
R-Tuning Modification
2023.10.30. I am unsure
Certain Data D1
Matched
Uncertain Data D0
Unmatched
Question: What's the capital of France? Answer:
Paris
Question: When did Apple unveil M3? Answer:
2020.1.23
Padding
Parametric Knowledge

Refusal-Aware Data Construction

- Q : {Question}, A : {Answer}. {Prompt}
- Prompt = “Are you sure you accurately answered the question based on your internal knowledge?”
- Answer = “I am sure”
  - When model gives a correct answer
- Answer = “I am unsure”
  - When model gives a correct answer

Results

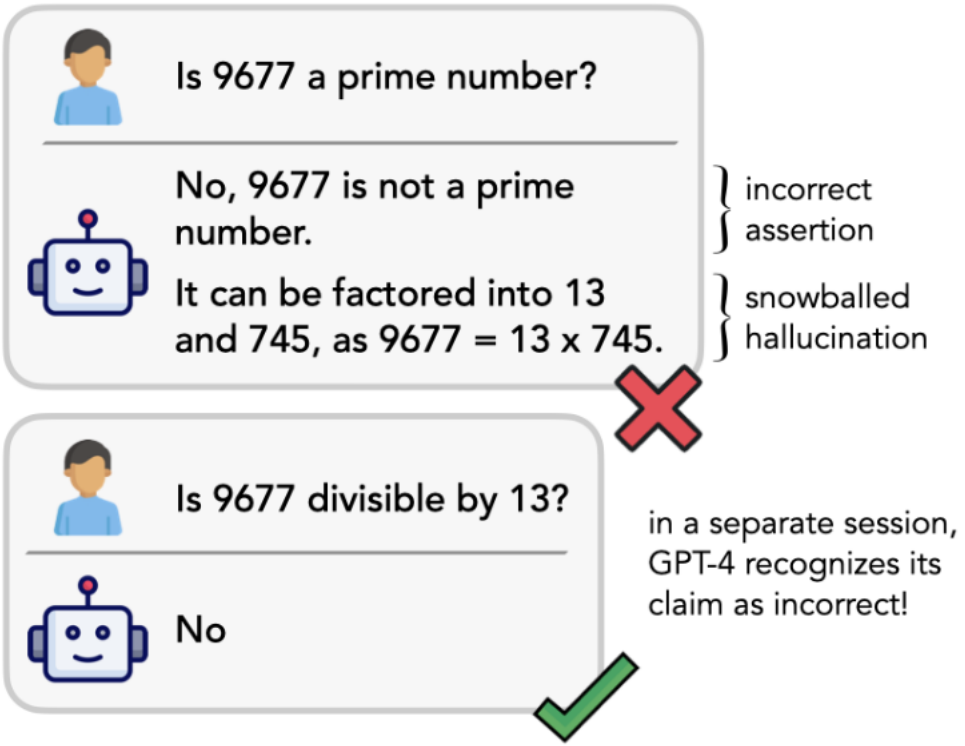
$$\text{accuracy} = \frac{\text{\# of correctly answered questions}}{\text{\# of willingly answered questions}}$$



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



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

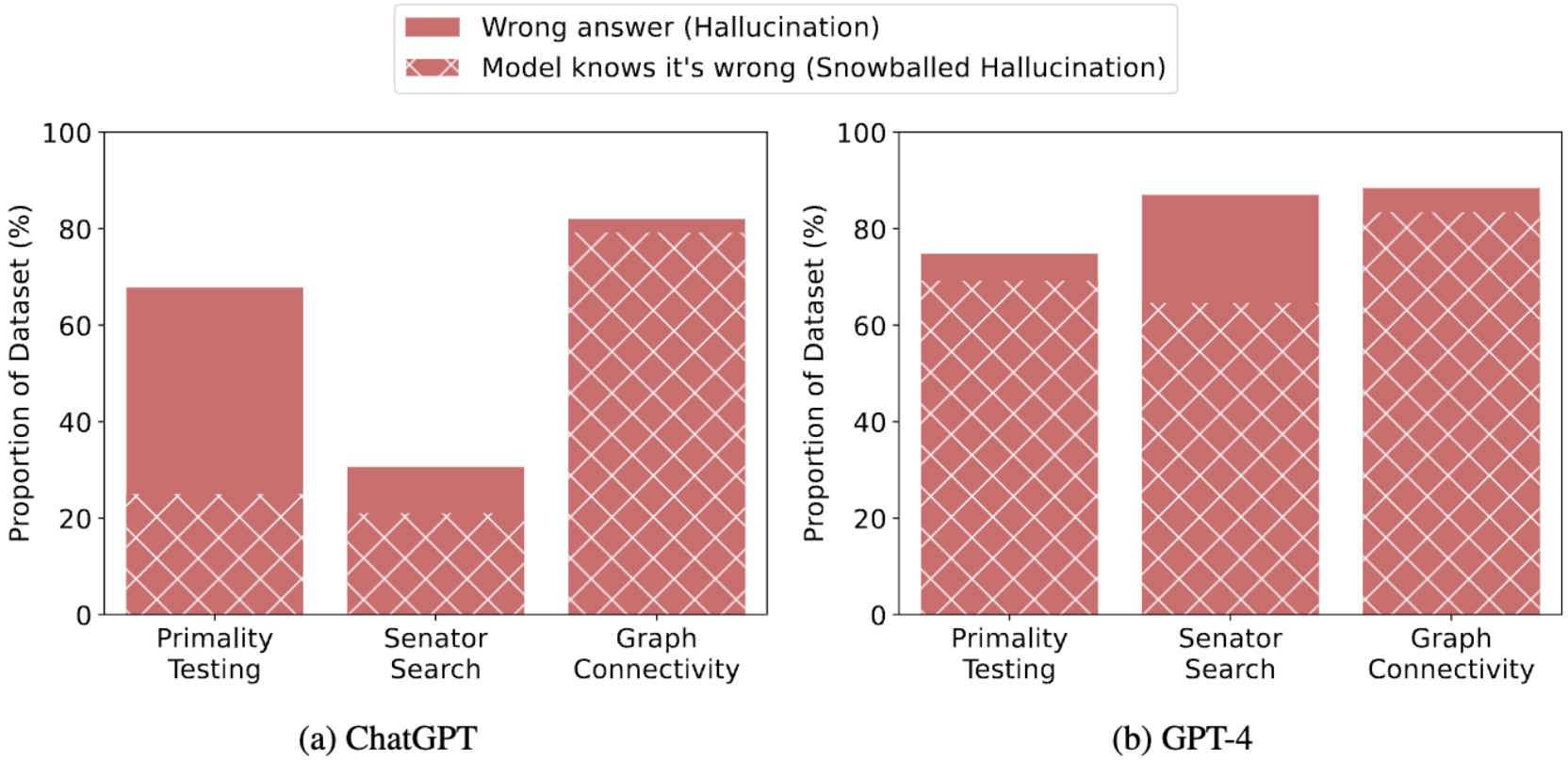


Figure 2: Percentage of hallucination and percentage of snowballed hallucination (both calculated with respect to the entire dataset) for ChatGPT and GPT-4. The precise numbers for this plot are available in Table 6 and Table 7 in the Appendix.

Results

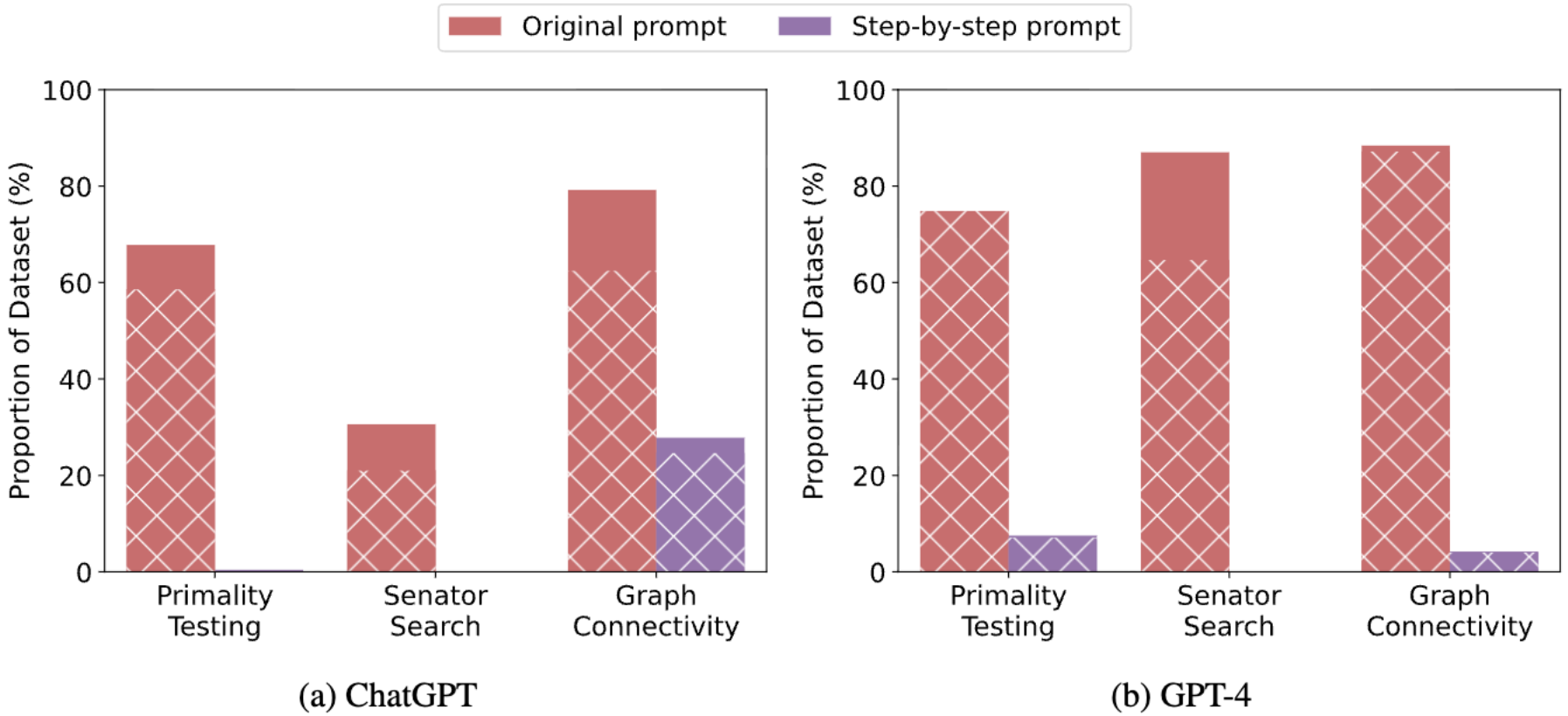


Figure 3: Error rate and snowballed hallucination rate (hatch pattern) for ChatGPT and GPT-4, when using the original prompt versus *"Let's think step by step"*. See Appendix Table 8 and Table 9 for the exact numbers.

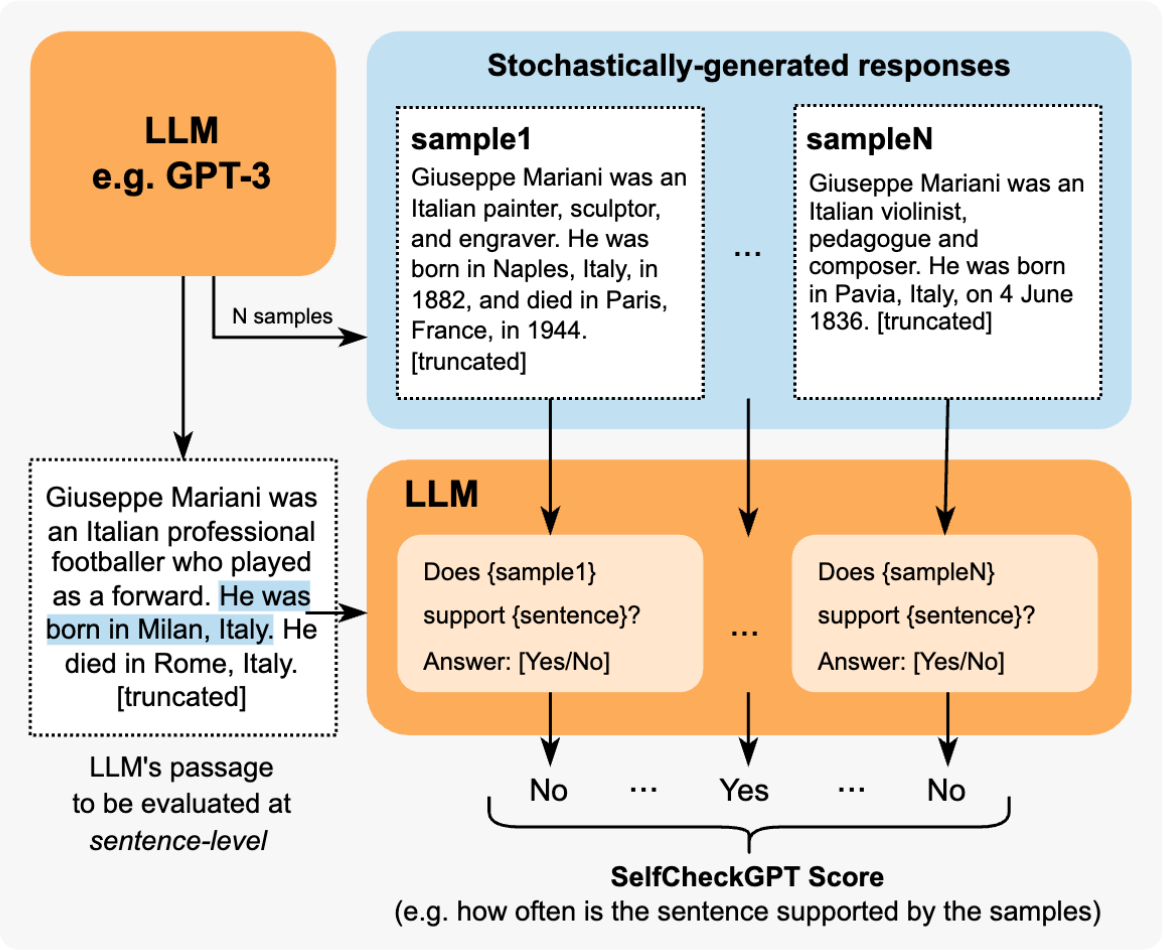
SELFCHECKGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models

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 Prompt



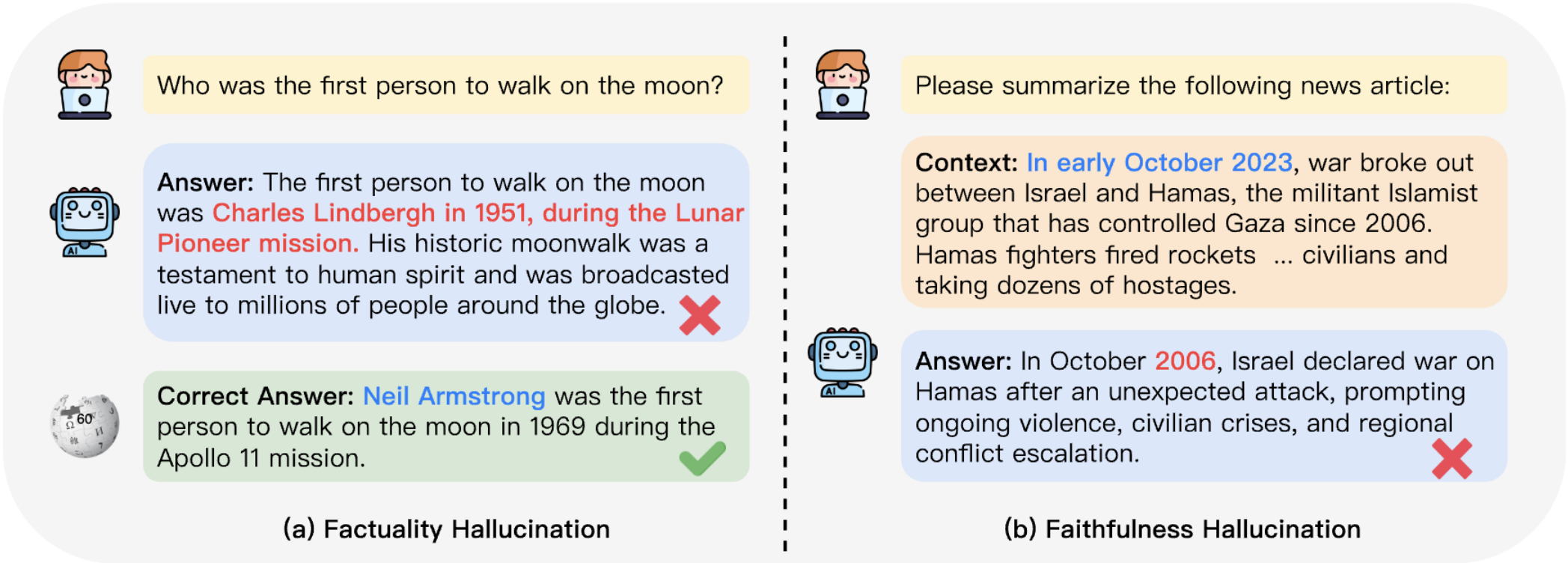

```
-----------------------------------------
Context: {}
Sentence: {}
Is the sentence supported by the context above?
Answer Yes or No:
-----------------------------------------
```

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

Results

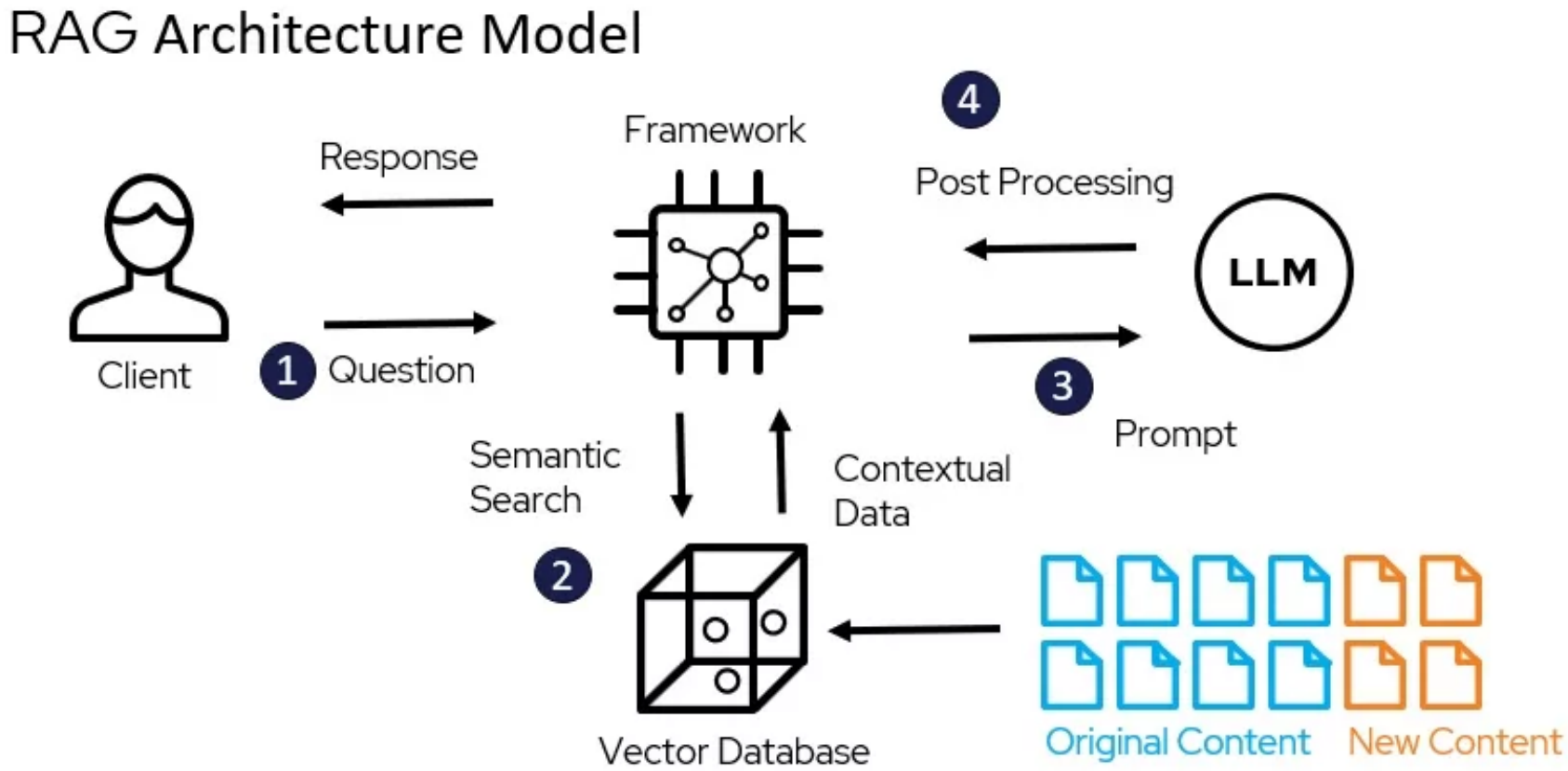
Method	Sentence-level (AUC-PR)			Passage-level (Corr.)	
	NonFact	NonFact*	Factual	Pearson	Spearman
Random	72.96	29.72	27.04	-	-
`GPT-3 (text-davinci-003)`'s probabilities (*LLM, grey-box*)					
Avg($-\log p$)	83.21	38.89	53.97	57.04	53.93
Avg($\mathcal{H}$)†	80.73	37.09	52.07	55.52	50.87
Max($-\log p$)	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**

Recap: Factuality vs. Faithfulness



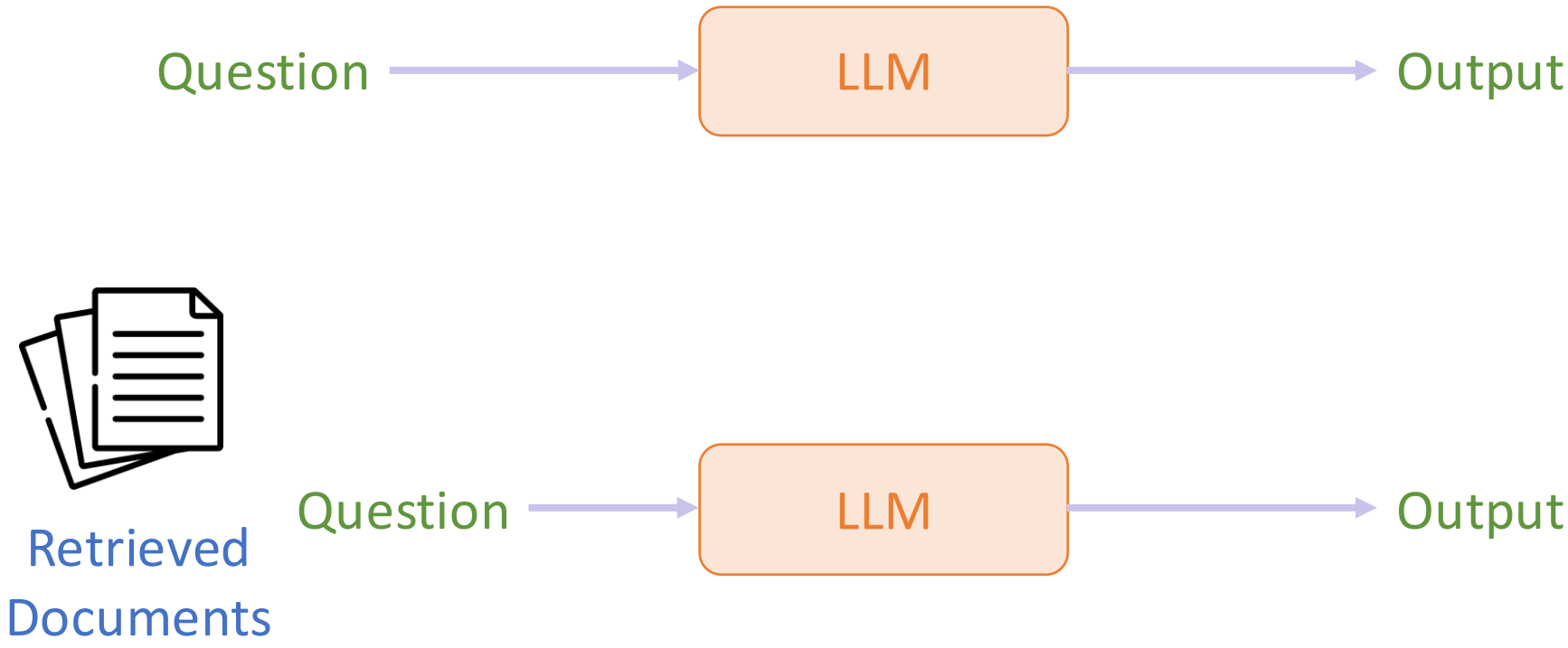
A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions, 2023

Recap: Retrieval-Augmented Generation (RAG)



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

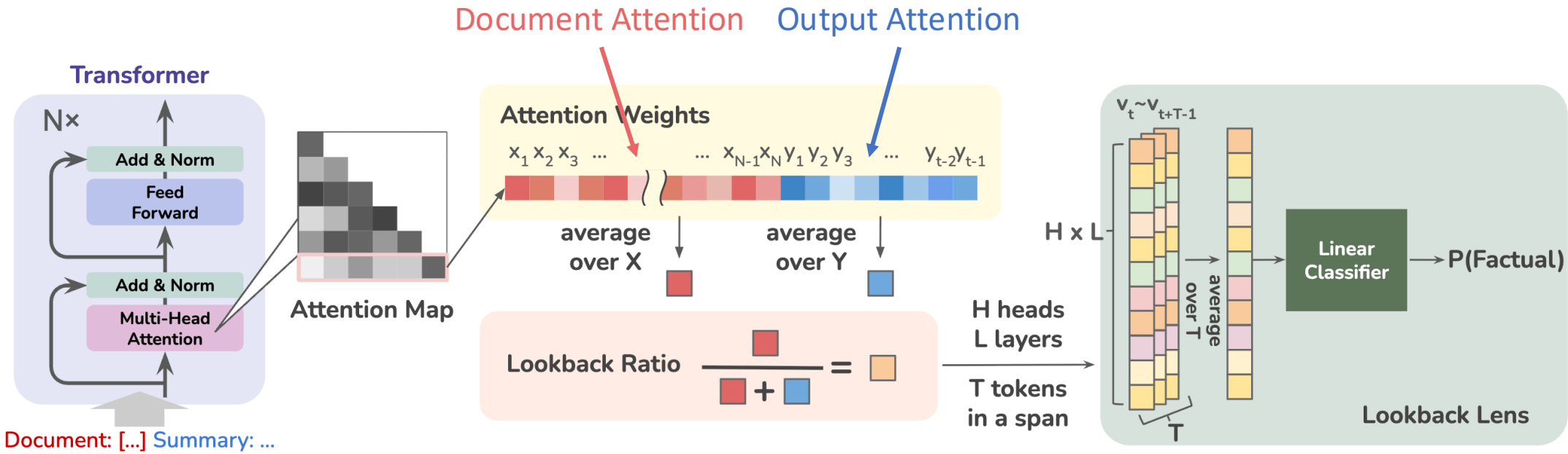
From Factuality to Faithfulness



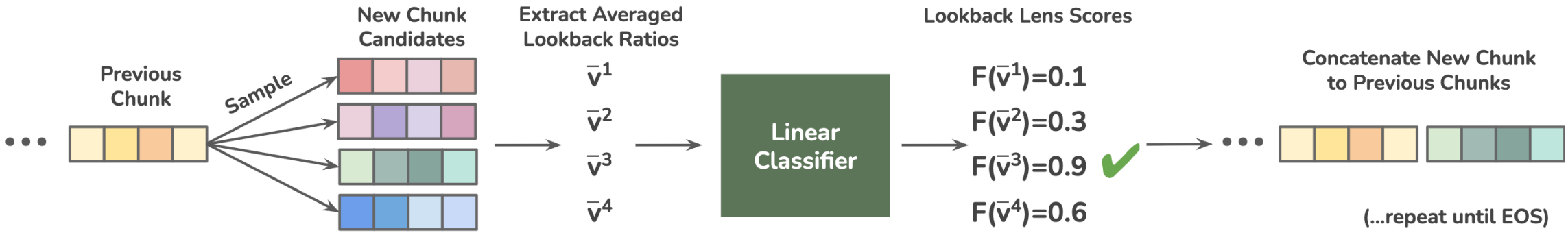
Lookback Lens: Detecting and Mitigating Contextual Hallucinations in Large Language Models Using *Only* Attention Maps

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

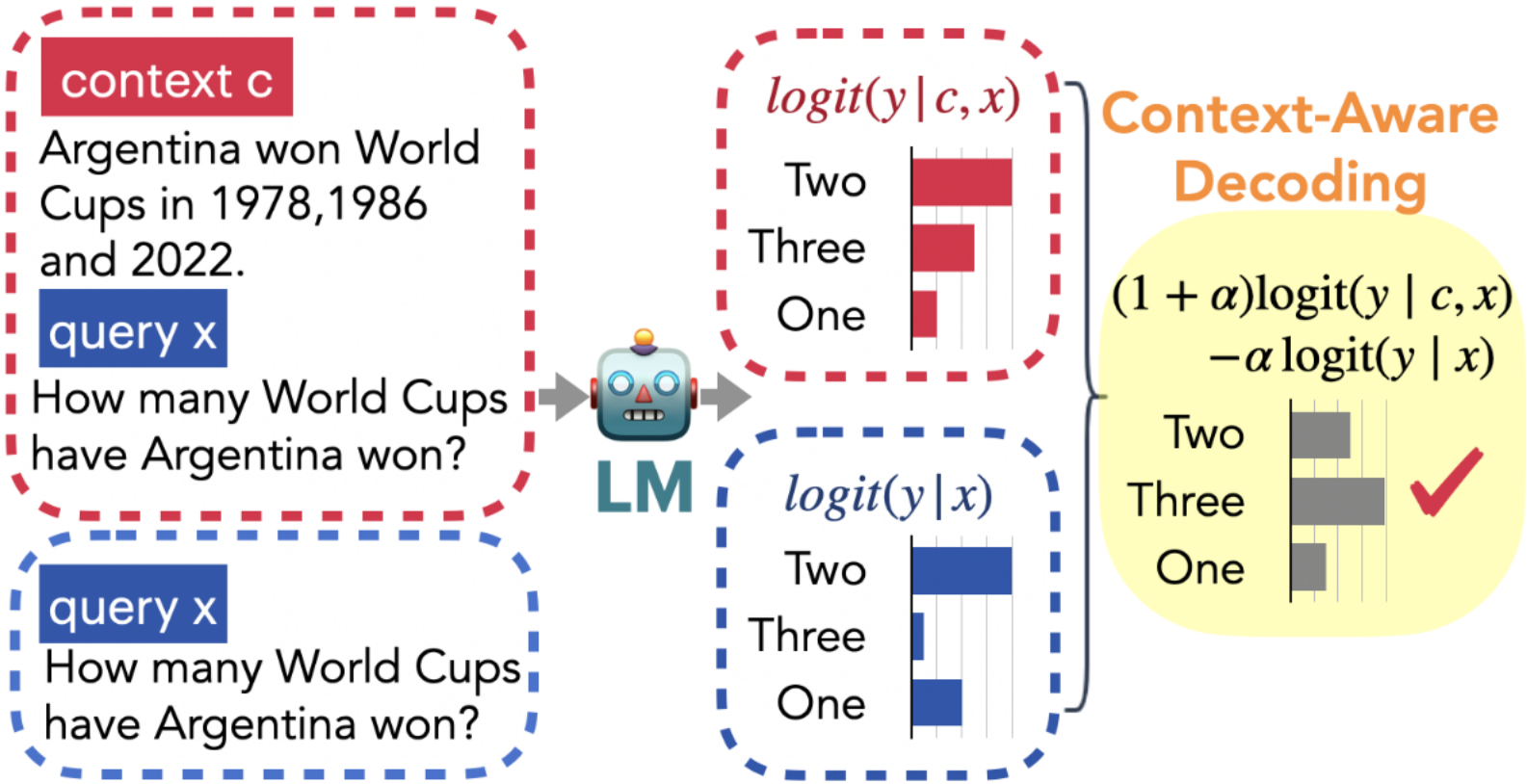
Weijia Shi [1] * **Xiaochuang Han** [1] *
Mike Lewis [2] **Yulia Tsvetkov** [1] **Luke Zettlemoyer** [1] **Scott Yih** [2]

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

Context-Aware Decoding



Context-Aware Decoding



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

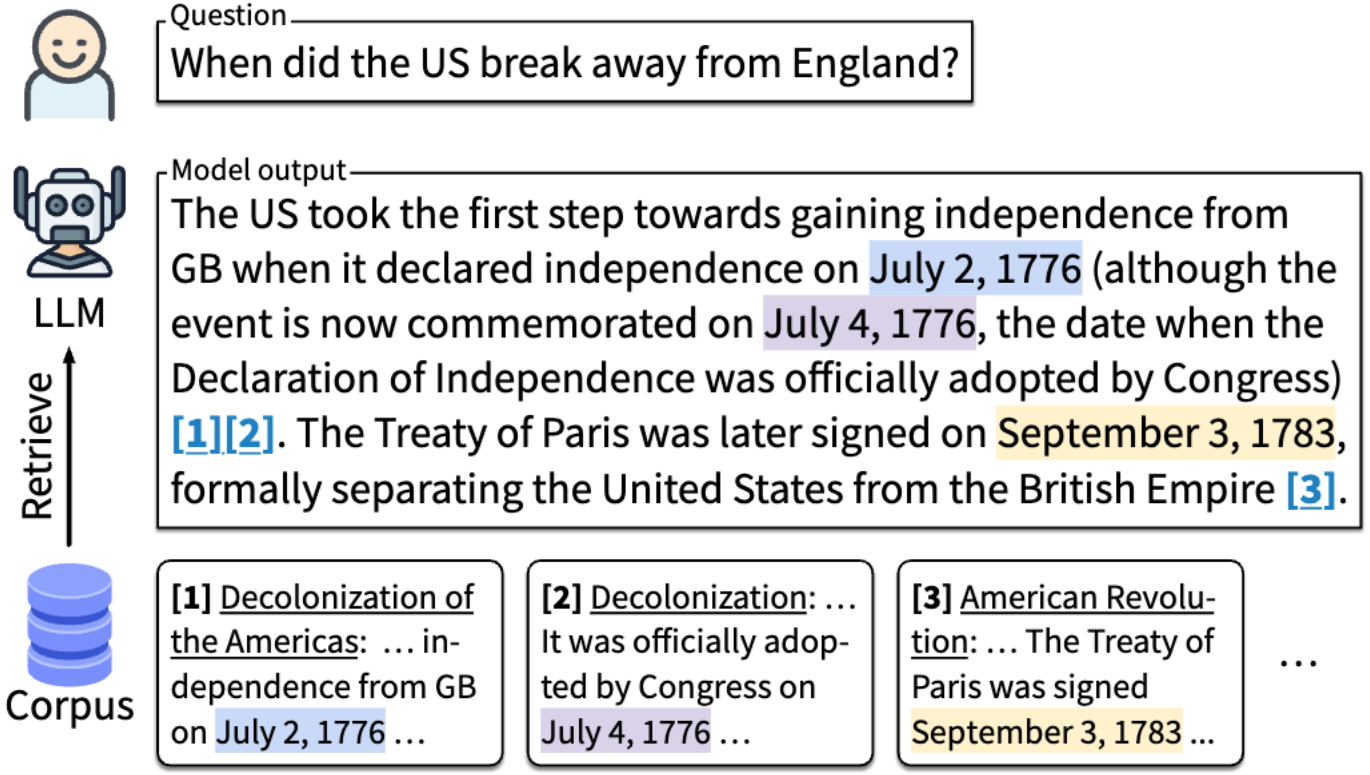
Results

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**

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



Examples

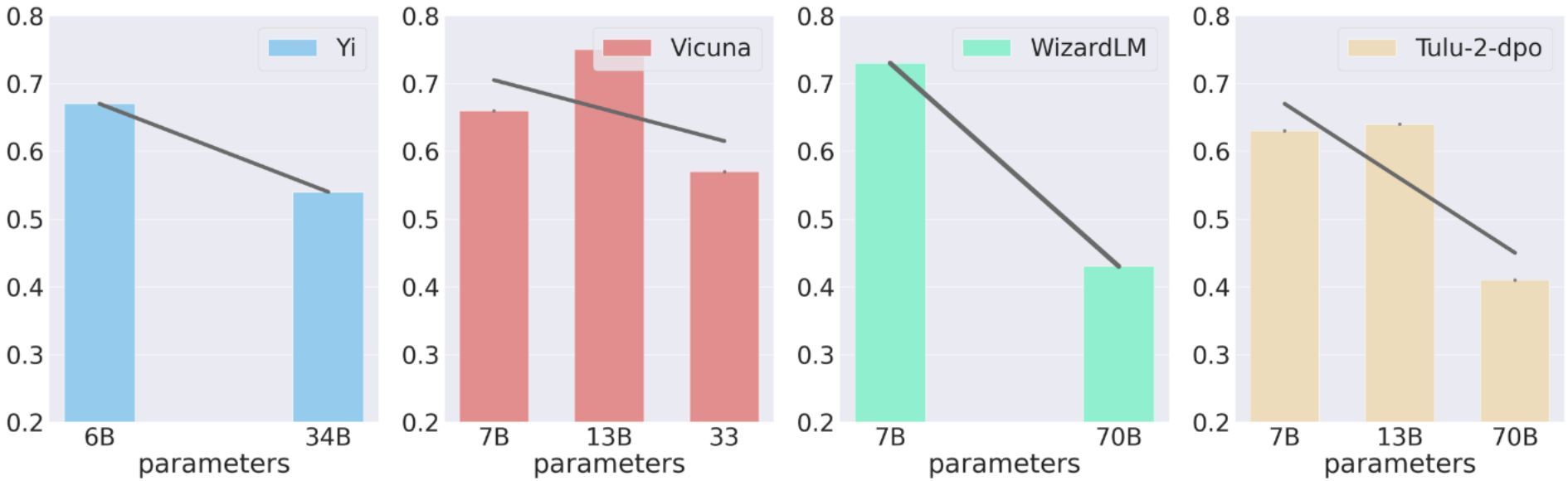
Dataset	Corpus (#passages)	Question type	Example
ASQA	Wikipedia (21M)	Factoid (ambiguous)	**Q**: When did the US break away from England? **A**: The US declared independence on July 2, 1776 [1][2] ... The Treaty of Paris was later signed on September 3, 1783 [3].
QAMPARI	Wikipedia (21M)	Factoid (list)	**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]

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



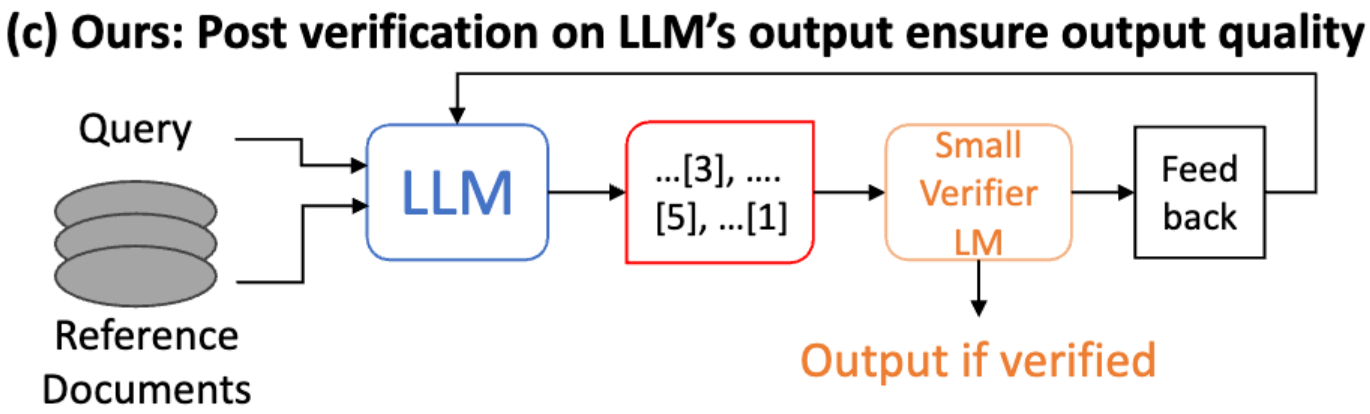
I-Hung Hsu[1]*, **Zifeng Wang**[2], **Long T. Le**[2], **Lesly Miculicich**[2], **Nanyun Peng**[3], **Chen-Yu Lee**[2], **Tomas Pfister**[2]
[1]University of Southern California, [2]Google Cloud AI Research,
[3]University of California, Los Angeles

Sensitivity to Retrieved Documents



Small LMs are more sensitive

Overview



Key idea: using small LMs as verifiers

Key Idea



Retrieved Documents

Question → Large LM → Output [2,3,5]



Retrieved Documents

Question → Small LM → Output [2,3,5]

Results

Method	Method Type	# Main LM Call	# Verifier LM Call	DPR as retriever					GTR as retriever				
				Fluency	Correct.	Citation		Average	Fluency	Correct.	Citation		Average
				mauve	EM Rec.	Prec.	Rec.		mauve	EM Rec.	Prec.	Rec.	
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)	Preprocess	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)	Postprocess	≤ 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)	Preprocess	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)	Postprocess	≤ 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

Questions?