

Large Language Model Hallucination and Solutions

Presenters:

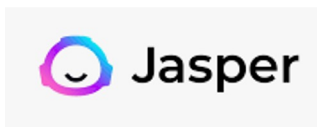
Oen McKinley & Johnny Huang & Jingyun Ma

Agenda

- ❖ **Introduction**
 - Introduction to the Large language model
- ❖ **Problem Awareness**
 - Introduction to the topic
- ❖ **Paper 1: Trusting Your Evidence: Hallucinate Less with Context-aware Decoding**
 - Details of the paper
- ❖ **Paper 2: How Language Model Hallucinations Can Snowball**
 - Details of the paper
- ❖ **Paper 3: Hallucination Detection for Generative Large Language Models by Bayesian Sequential Estimation**
 - Details of the paper
- ❖ **Paper 4: Improving Factuality and Reasoning in Language Models through Multiagent Debate**
 - Details of the paper
- ❖ **Summary**
 - Summary to the four papers
- ❖ **Future Works**
- ❖ **References**
 - References
- ❖ **Discussion**
 - Questions

Introduction

- LLMs, such as ChatGPT, Claude, and Bard AI, represent a significant leap in the field of natural language processing.



Prompt: "Python function to calculate factorial." LLM:

```
python
def factorial(n):
    return 1 if n == 0 else n * factorial(n-1)
```

复制代码

3. Translation

Prompt: "Translate 'The quick brown fox jumps over the lazy dog' to Chinese." LLM: "敏捷的棕色狐狸跳过懒狗。"

4. Summarization

Prompt: "Summarize: 'AI is transforming industries by automating processes and improving decisions.'" LLM: "AI automates processes and enhances decision-making."

Prompt: "The mist..." LLM: "...and as dawn broke, a glowing n the trees."

Problem Awareness

- Language models often generate unfaithful or hallucinated text



Once upon a time, in a faraway kingdom, there lived a beautiful princess named Snow White. She was known for her kindness and beauty, but her life was far from easy. Her wicked stepmother, the Queen, was envious of her beauty and treated her like a servant, forcing her to do all the household chores.

One day, Snow White heard of a grand ball at the royal palace. Every maiden in the kingdom was invited, and Snow White dreamed of attending. But her stepmother forbade her from going and gave her an endless list of chores to complete. Undeterred, Snow White worked hard and finished all her tasks, hoping she might still make it to the ball.

As Snow White sat weeping by the fireplace, a fairy godmother appeared. The fairy godmother, seeing Snow White's pure heart, decided to help her. With a wave of her wand, she transformed Snow White's tattered clothes into a magnificent gown and gave her a pair of glass slippers. She also conjured a beautiful carriage from a pumpkin, along with horses and footmen, so Snow White could attend the ball in style.

Story of “Snow Cinderella”

- Insufficient attention to input context
- Snowball hallucinations
- Prone to generating non-factual content

Trusting Your Evidence: Hallucinate Less with Context-aware Decoding

Weijia Shi^{1 *}

Xiaochuang Han^{1 *}

Mike Lewis²

Yulia Tsvetkov¹

Luke Zettlemoyer¹

Scott Yih²

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

Methodology : Context-aware Decoding (CAD)

- CAD adjusts the output probability distribution to favor contextually relevant information
- Uses a contrastive approach to amplify the difference between outputs with and without context

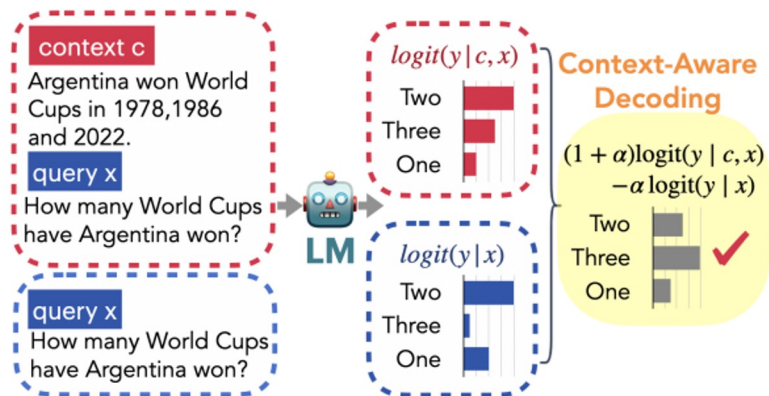


Figure 1: An illustration of context-aware decoding.

Language Model Prediction:

- \mathbf{c} : The context provided to the model
- \mathbf{x} : The input query
- $\mathbf{y}_{<t}$: The sequence of tokens generated by the model up to time step

$$y_t \sim p_{\theta}(y_t \mid \mathbf{c}, \mathbf{x}, \mathbf{y}_{<t})$$

$$\propto \exp(\text{logit}_{\theta}(y_t \mid \mathbf{c}, \mathbf{x}, \mathbf{y}_{<t}))$$

Logit Transformation:

- The probability of the model outputting is proportional to the exponential of the logit function

$$y_t \sim \tilde{p}_\theta(y_t \mid c, x, y_{<t}) \propto p_\theta(y_t \mid c, x, y_{<t}) \left(\frac{p_\theta(y_t \mid c, x, y_{<t})}{p_\theta(y_t \mid x, y_{<t})} \right)^\alpha$$

- $p_\theta(y_t \mid c, x, y_{<t})$: The probability of y_t given the context
- $p_\theta(y_t \mid x, y_{<t})$: The probability of y_t without the context

- The ratio essentially tells us how much more likely y_t becomes when the context c is included
- The exponent α is a weight that controls how much influence this adjustment has

This expression is not a valid probability distribution and needs to be normalized across all possible values of y_t .

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

$\text{logit}_\theta(y_t \mid c, x, y_{<t})$: How likely the model thinks y_t should be the next word when considering the full context

$\text{logit}_\theta(y_t \mid x, y_{<t})$: contrast the model's behavior with and without the context.

XSUM	
c	Article: Prison Link Cymru had 1,099 referrals in 2015-16 and said some ex-offenders were living rough for up to a year before finding suitable accommodation ...
x	Summarize the article in one sentence. Summary:
NQ-SWAP	
c	Tesla CEO Elon Musk is now in charge of Twitter , CNBC has learned ...
x	Who is Twitter CEO now?
MemoTrap	
c	Write a quote that ends in the word "early":
x	Better late than

Table 1: An illustration of the inputs to CAD applied to each dataset. CAD upweights the context c (in red) by sampling each token from $\text{softmax}[(1 + \alpha) \text{logit}_\theta(y_t \mid c, x, y_{<t}) - \alpha \text{logit}_\theta(y_t \mid x, y_{<t})]$.

Experimental Setup

Task Evaluated:

- Summarization (CNN-DM, XSUM)
- Knowledge Conflicts (MemoTrap, NQ-Swap)

Models Tested:

- OPT, GPT-Neo, LLaMA, FLAN-T5

Metrics:

- ROUGE-L - Measure the overlap
- BERT-Precision - Assessee outputs
- FactKB - Measure factual consistency
- Exact Match - Measure factual consistency

Result & Analysis

- CAD outperforms regular decoding in improving factual accuracy.
 - 14.3% improvement in factuality metrics for LLaMA on CNN-DM.
 - 2.9x improvement in LLaMA on knowledge conflicts QA dataset

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

CNN-DM:

A summarization task where the goal is to generate concise summaries of news articles.

XSUM: Another summarization task, but with a focus on extreme summarization, where the summaries are very short, typically a single sentence.

Result & Analysis

- Memo Trap - Test case of memorization traps
- NQ - Question answering tasks
- NO-SWAP - Prior knowledge test

→ CAD Outperforms Regular Decoding

→ Model Performance Variations

- ◆ Larger models tend to show greater improvements with CAD

→ Implication

- ◆ An effective method

Model	Decoding	Memo.	NQ	NQ-SWAP	
OPT	13B	Reg.	32.5	29.2	18.8
		CAD	44.5	32.2	36.9
	30B	Reg.	28.4	29.4	14.7
		CAD	41.0	35.5	29.0
GPT.	3B	Reg.	22.5	31.9	19.1
		CAD	47.3	39.9	41.2
	20B	Reg.	37.1	22.8	16.1
		CAD	57.3	32.1	36.8
LLAMA	13B	Reg.	23.8	22.3	11.7
		CAD	57.1	33.6	36.7
	30B	Reg.	25.8	23.8	9.6
		CAD	50.6	34.0	37.7
FLAN	3B	Reg.	69.2	81.8	71.4
		CAD	72.2	80.3	73.3
	11B	Reg.	82.0	85.5	73.0
		CAD	88.7	82.5	77.1

Table 3: CAD outperforms the regular decoding method (Reg.) in all settings except for FLAN-T5 on NQ. Note that FLAN-T5 is trained on NQ dataset during instruction-finetuning.

Result & Analysis

- CAD Consistently Outperforms Regular Decoding

CNN-DM:
CAD shows a clear advantage

MemoTrap:
CAD maintains or slightly
improves performance

NQSWAP:
CAD shows more stable
performance across different
sizes

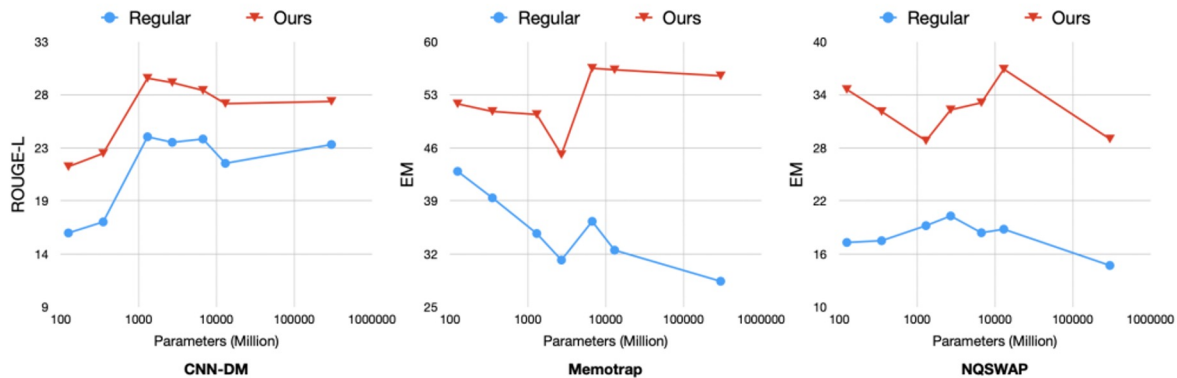
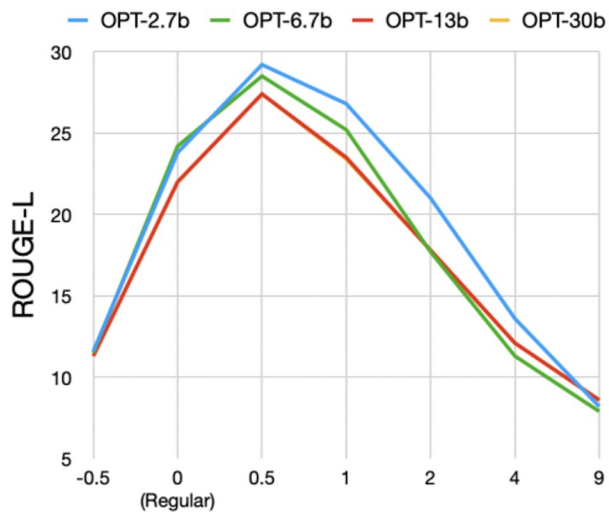


Figure 2: OPT models of varying sizes consistently benefit from CAD. The x-axis indicates the size of language models and the y-axis is the performance.

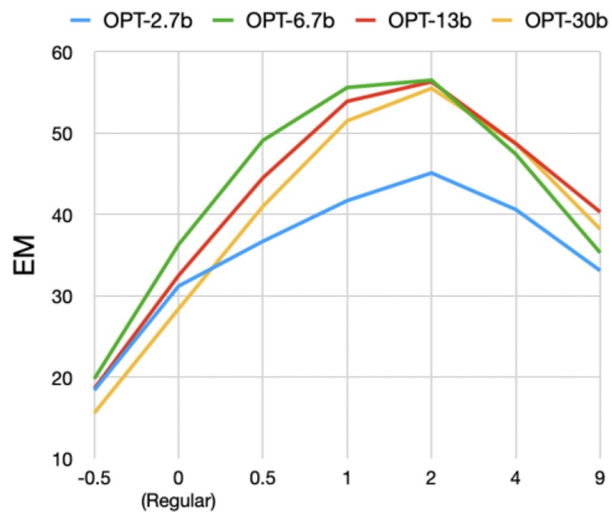
CNN-DM: Summarization task
where ROUGE-L is used as the
performance metric

Result & Analysis

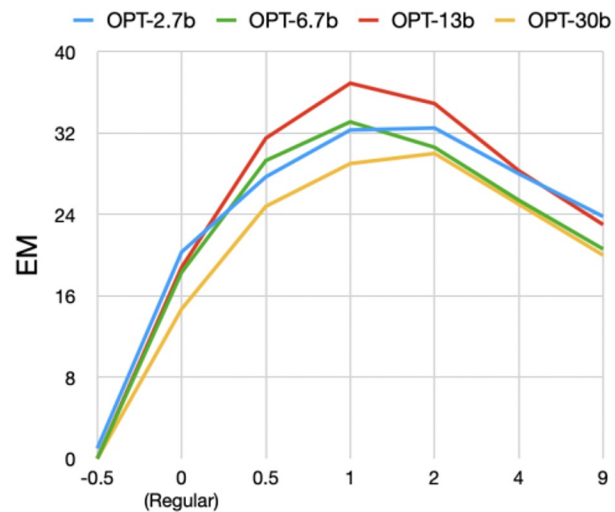
- Demonstrates that there is an optimal range for the context-aware adjustment coefficient α , typically around 0.5 to 1.0, where the performance of the model is maximized across different tasks.



coefficient
CNN-DM



coefficient
Memotrap



coefficient
NQSWAP

Conclusion

- Larger models benefit more from CAD.
 - CAD effectively mitigates hallucinations by focusing on contextual information
- CAD's performance improves with increased α , especially in knowledge conflict tasks.
 - Applicable to various language models without additional training.

Future Work

- Further exploration of CAD in different domains and model architectures.

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



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



False Claim



Snowballed
Evidence



Is 9677 divisible by 13?



No

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



Corrected
Evidence






Why Does SnowBalling Happen?

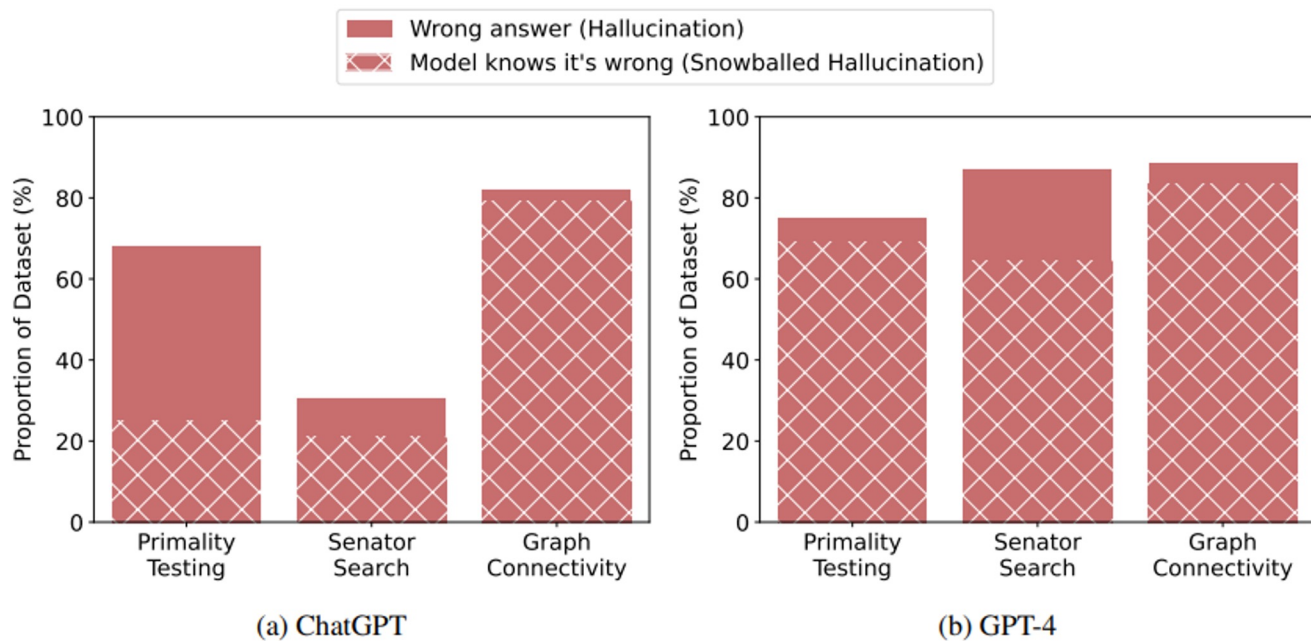
1. **Initial committal:** The prompt leads the LM to first state an answer (*before* outputting the explanation). This applies to many yes/no questions.
2. **Inherently sequential:** Transformers cannot find the answer within one timestep because of their limited reasoning abilities within one timestep.

Experiment and Dataset:

- 3 QA datasets to probe hallucination snowballing:

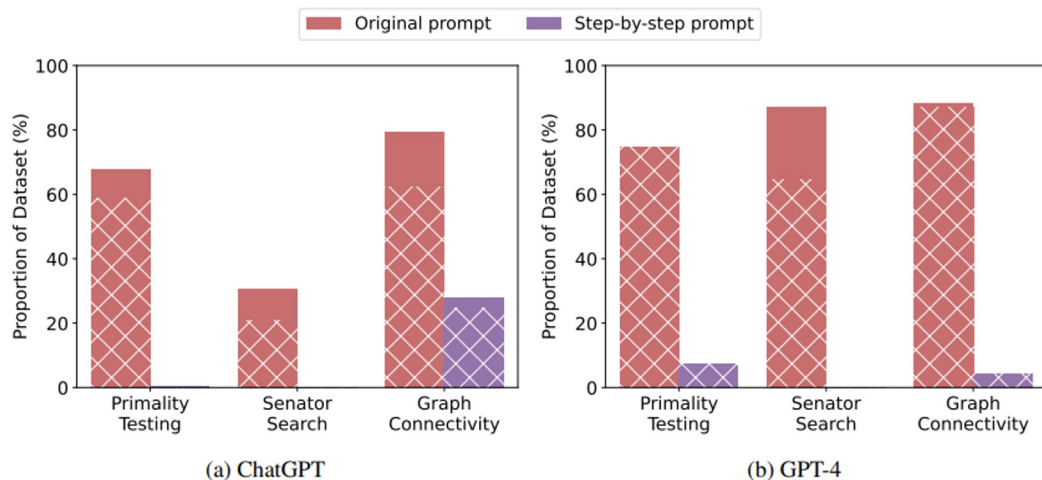
Dataset	Original Question	Verification Question
 Primality Testing	<p>User: Is 10733 a prime number?</p> <p>GPT-4: No... It can be <u>factored</u> into 3×3577.</p>	<p>User: Is 10733 divisible by 3? Answer with either Yes or No.</p> <p>GPT-4: <u>No</u></p>
 Senator Search	<p>User: Was there ever a US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

Results:



Prevention?

- Encouraging Model to generate reasoning before answer:
 - “Let’s think step by step...”



Step 3: From city E, we have three options: a flight to city N, a flight to city B, or a flight to city C.

Step 4: The only option that could potentially lead us to city M is the flight from city E to city C.

Miscellaneous Algorithmic Corrections:

- 1) Increasing temperature: At $t=$ to **0.6 ~ 0.9** the authors had the lowest error rates.
- 2) Top-k and nucleus sampling: Does **not** help as they narrow the range of tokens, causing more initial committal.
- 3) Beam search: Predicted would help tremendously

Zhang, M., Press, O., Merrill, W., Liu, A., & Smith, N. A. (2023). How language model hallucinations can snowball. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*. Paul G. Allen School of Computer Science and Engineering, University of Washington; New York University; Allen Institute for Artificial Intelligence

Takeaways

- The problem of Hallucination Snowballing
- Solutions in thought probing
- Solutions in algorithm

Hallucination Detection for Generative Large Language Models by Bayesian Sequential Estimation

Xiaohua Wang^{1,2}, Yuliang Yan^{1,2}, Longtao Huang³, Xiaoqing Zheng^{1,2,*}, Xuanjing Huang^{1,2}

¹School of Computer Science, Fudan University, Shanghai, China

²Shanghai Key Laboratory of Intelligent Information Processing

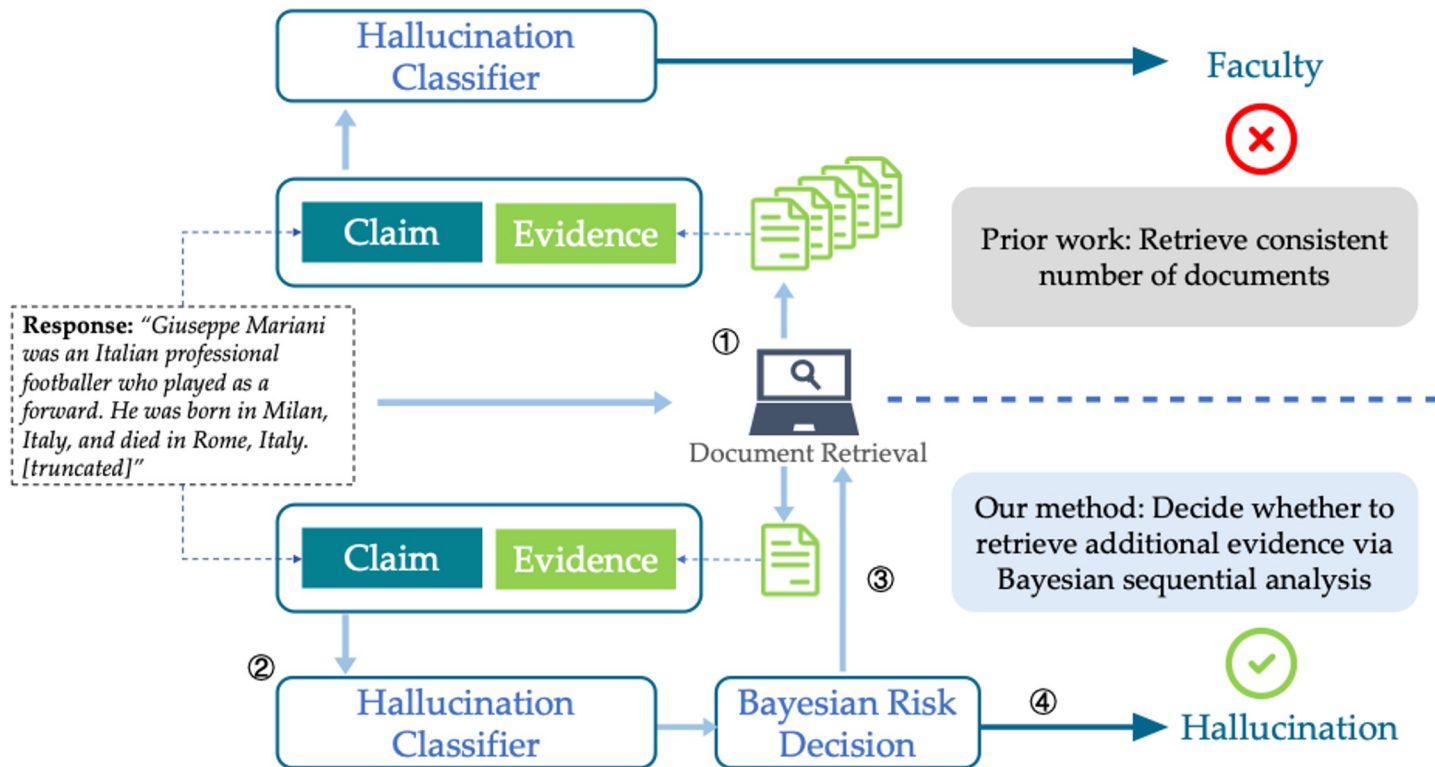
³Alibaba Group

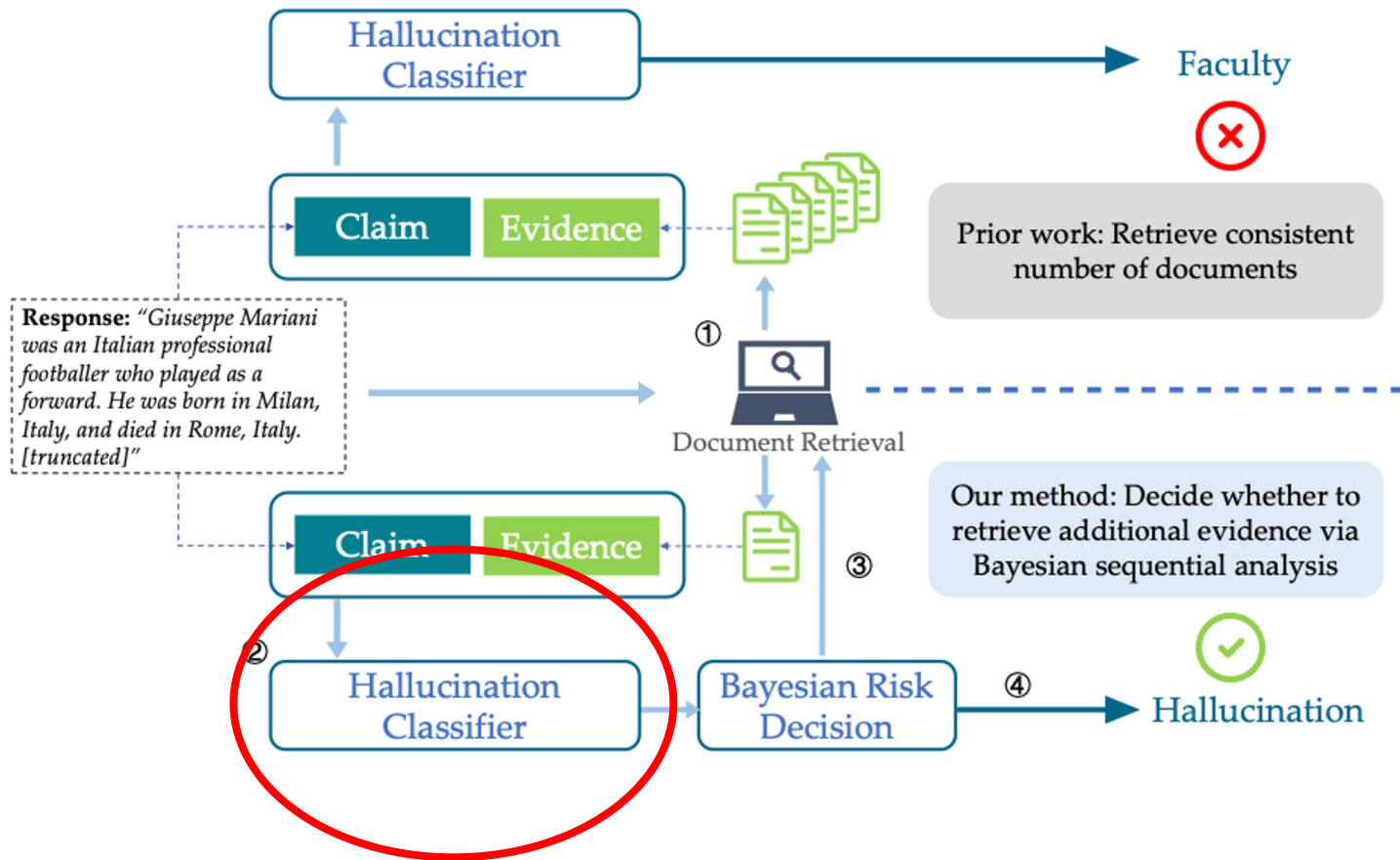
{xiaohuawang22,ylyan21}@m.fudan.edu.cn

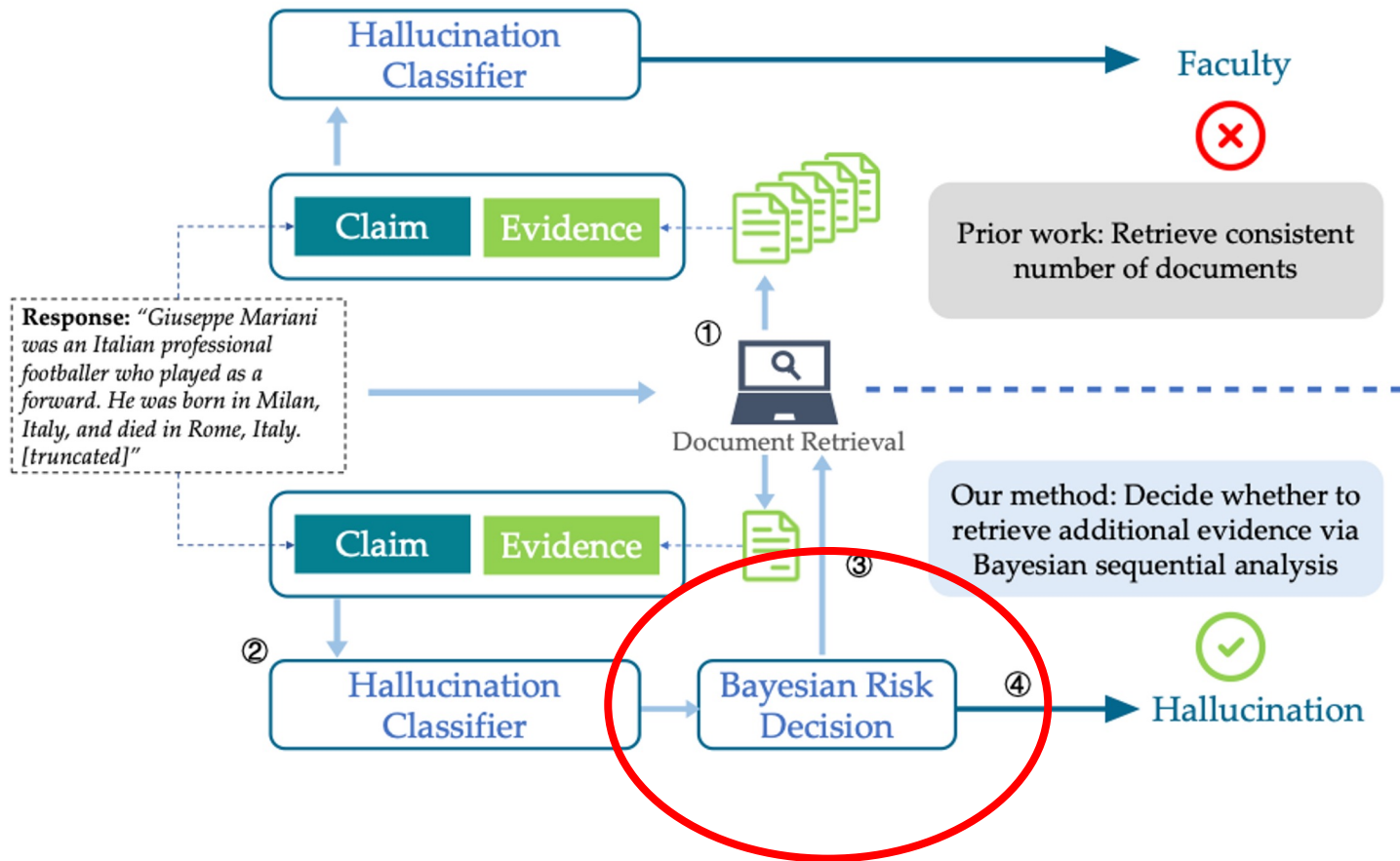
{zhengxq,xjhuang}@fudan.edu.cn

Hallucination Detection Method:

- Historic: Document retrieval to validate or challenge claims
 - Issues:
 - Hard-coded amount of documents
 - fixed context span
 - ie. texts, knowledge graphs, web results
- Solution proposed by paper:
 - Dynamically determining optimal number of external evidence sources
 - Taking into account various factors ie. complexity, ambiguity, availability of sources
 - Utilizing a **Bayesian Sequence Model** while keeping retrieval the same







Method:

- Bayesian Sequential Analysis:
 - Framework for making informed decisions based on accumulating evidence.
 - Select decisions that minimize **expected costs** utilizing grid approximation.
- C_{FA} : Cost of false alarm
- $C_{retrieve}$: Cost of retrieve
- C_M : Cost of miss

Cost Functions:

$$R_{stop}(n) = \min((1 - \pi_1(n))C_M, (1 - \pi_0(n))C_{FA}) \quad (5) \quad \leftarrow \text{Cost Stop}$$

$$R_{continue}(n) = C_{retrieve} + \sum_{f_{n+1}=0}^9 R(n+1) \cdot P(f_{n+1}) \quad (6) \quad \leftarrow \text{Cost Continue}$$

$$R_{continue}(n) = C_{retrieve} + \sum_{f_{n+1}=0}^9 R_{stop}(n+1) \cdot P(f_{n+1}) \quad \leftarrow \text{Optimized Continue}$$

$$R(n) = \min(R_{continue}(n), R_{stop}(n)) \quad (7) \quad \leftarrow \text{Decision}$$

```
1  $\{C^1, C^2, \dots, C^L\} \leftarrow \text{ClaimDecompose}(C);$ 
```

Sub claims

```
2 for  $i \leftarrow 1$  to  $L$  do
```

```
3    $n \leftarrow 1;$ 
```

```
4   while  $n \leq k$  do
```

```
5      $E^n \leftarrow \text{RetrieveDocument}(C^i);$ 
```

Document Retrieval

```
6      $f_n \leftarrow \text{CalEntailmentFeature}(E^n, C^i);$ 
```

Entailment/validity of subclaim

```
7      $\pi_1(n) \leftarrow \text{NBC}(\pi_1(n-1), f_n);$ 
```

```
8      $R_{stop}(n) \leftarrow \min((1 - \pi_1(n))C_M,$   
       $(1 - \pi_0(n))C_{FA});$ 
```

```
9      $R_{continue}(n) \leftarrow$   
       $C_{retrieve} + \mathbb{E}_{f_{n+1}}(R_{stop}(n+1));$ 
```

Bayesian Risk Decision w/ Costs

```
10    if  $R_{stop}(n) < R_{continue}(n)$  then
```

```
11      | break;
```

```
12      else
```

```
13      |  $n \leftarrow n + 1;$ 
```

```
14      end
```

```
15    end
```

```
16     $P_{factual}^i = \pi_1(n);$ 
```

```
17 end
```

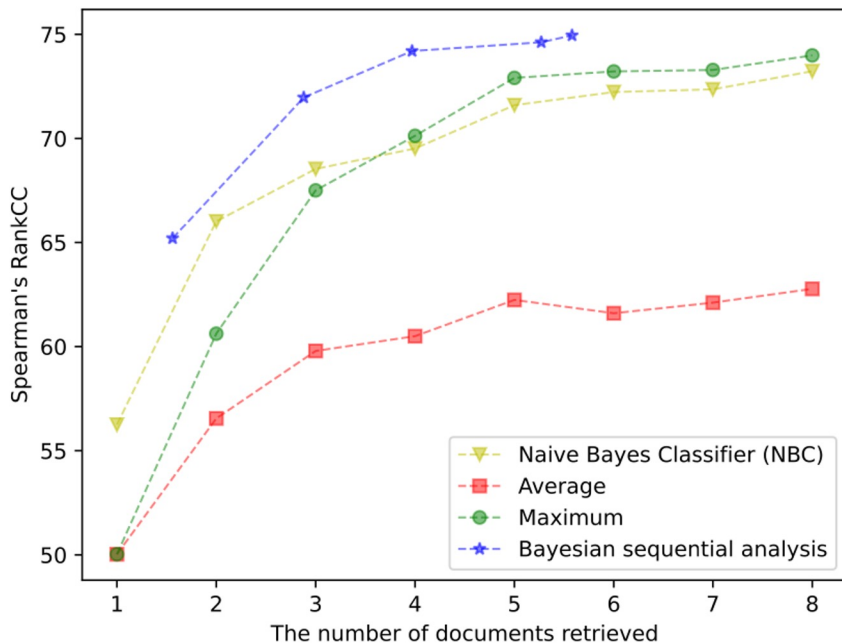
```
18  $P_{factual}(C) = \min_i P_{factual}(C^i);$ 
```

Minimum factual score

```
19 Return:  $P_{factual}(C)$ 
```


Results:

Method		Sentence-level (AUC-PR)				Passage-level (Corr.)	
		Evidence Num	Nonfact	Factual	Acc	Pearson	Spearman
Self-Detection		-	-	-	31.01	-	-
SelfCheckGPT	w/ BERTScore	20	81.96	44.23	-	58.18	55.90
	w/ QA	20	84.26	48.14	-	61.07	59.29
	w/Unigram (max)	20	85.63	58.47	-	64.71	64.91
	Combination	60	87.33	61.83	-	69.05	67.77
Our Framework	$C_M = 14, C_{FA} = 24$	3.05	82.42	57.01	80.24	71.37	64.55
	$C_M = 28, C_{FA} = 96$	6.22	86.45	61.96	82.39	81.18	74.20



Method	Sentence-level (AUC-PR)		
	Nonfact	Factual	Acc
w/o Decomposition	80.04	53.71	79.19
w Decomposition	82.42	57.01	80.24

Wang, X., Yan, Y., Huang, L., Zheng, X., & Huang, X. (2023). Hallucination detection for generative large language models by Bayesian sequential estimation. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*. Fudan University, Shanghai Key Laboratory of Intelligent Information Processing; Alibaba Group

Takeaways:

- Historic Hallucination detection consist of fixed evidence retrieval
- Using Bayesian Sequential Model is better
 - More optimization
 - Dynamic adjustments

Improving Factuality and Reasoning in Language Models through Multiagent Debate

Yilun Du
MIT CSAIL
yilundu@mit.edu

Shuang Li
MIT CSAIL
lishuang@mit.edu

Antonio Torralba
MIT CSAIL
torralba@mit.edu

Joshua B. Tenenbaum
MIT CSAIL, BCS, CBMM
jbt@mit.edu

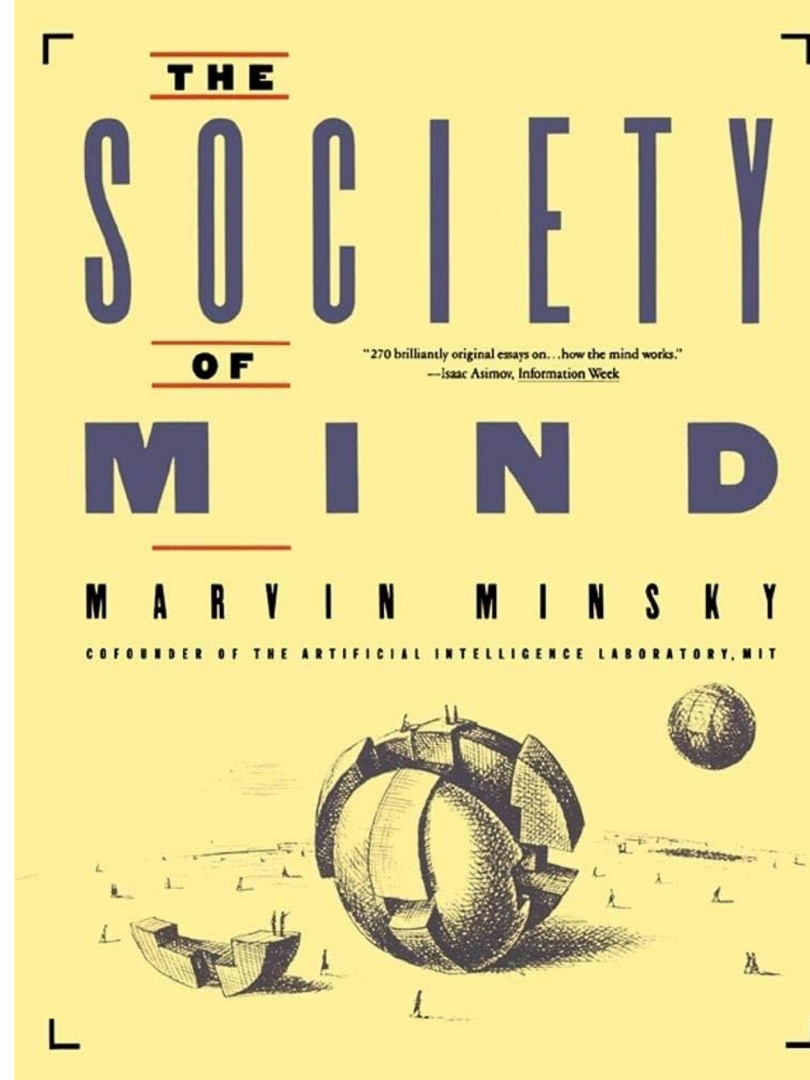
Igor Mordatch
Google Brain
imordatch@google.com

Background & Related Work

- Many works have contributed different solutions to hallucination:
 - Prompting: scratchpads, verification, chain-of-thought, self-reflection, fine-tuning
 - Training: reinforcement learning, dataset pruning, external knowledge, likelihood estimation
- All single-agent, with or without outside influence

Background & Related Work

- *The Society of Mind* by Marvin Minsky (1986)
- A philosophical framework that views a “mind” as a network of “agents”
 - Language, learning, memory, etc.
- Through collision and teamwork, these agents are the source of intelligence and consciousness



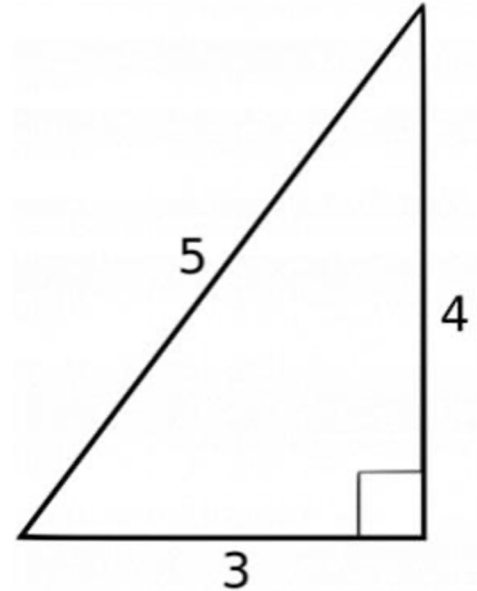
Multiagent Debate

- Du et al. reframes *Society of Mind* theory, with LLMs as agents
- Each agent creates an answer to a prompt, which is then given to the other agent(s) in a new round of prompting
- Like PHP Hinting, but the hints are generated by other agents

Multiagent Debate

- Consider solving a math problem: Find the area of this triangle
- Might think “right triangle, so area is $0.5 * h * w$ ”
 - $0.5 * 4 * 3 = 6$
- Then might double-check with Law of Cosines:
 - $\frac{1}{2} * h * w * \sin(\theta) = \frac{1}{2} * 4 * 3 * 1 = 6$
- If the answers are the same, you feel more confident

(Incidentally, the paper says the answer is 64, I think it's a typo.)



Multiagent Debate Procedure

- Initial Prompt
- Initial Answer ->
Secondary prompt

Debate Length	Prompt
Short	<i>" These are the solutions to the problem from other agents: [other answers] Based off the opinion of other agents, can you give an updated response . . ."</i>
Long	<i>" These are the solutions to the problem from other agents: [other answers] Using the opinion of other agents as additional advice, can you give an updated response . . ."</i>

Figure 3: **Prompts to induce long and short form debate.** Responses of other agents to questions are inserted in the middle of the prompt (indicated with *[other answers]*)

- Secondary Answer ->

...

- Until consensus or max rounds

Multiagent Debate Example

Question: Carla is downloading a 200 GB file. Normally she can download 2 GB/minute, but 40% of the way through the download, Windows forces a restart to install updates, which takes 20 minutes. Then Carla has to restart the download from the beginning. How long does it take to download the file?

Round 1

ChatGPT: Before the restart, Carla was able to download 40% of the file, which is $0.4 * 200 \text{ GB} = 80 \text{ GB}$. At a rate of 2 GB/minute, it would take her: $\text{time} = 80 \text{ GB} \div 2 \text{ GB/minute} = 40 \text{ minutes}$.

Because of the restart, Carla had to start over. This means she had to download the entire 200 GB file. The time it takes to download the remaining 120 GB is: $\text{time} = 120 \text{ GB} \div 2 \text{ GB/minute} = 60 \text{ minutes}$.

But Carla also had to wait 20 minutes for the Windows update to install, so the total time it took to download the file is: $\text{total time} = 40 \text{ minutes} + 60 \text{ minutes} + 20 \text{ minutes} = 120 \text{ minutes}$. ❌

Bard: Carla is downloading a 200 GB file at a speed of 2 GB/minute. This means that it will take her $200 / 2 = 100 \text{ minutes}$ to download the file.

However, Windows forces a restart to install updates, which takes 20 minutes. This means that Carla will have to restart the download from the beginning. Since Carla has already downloaded 40% of the file, she only needs to download $200 - 40 = 160 \text{ GB}$. At a speed of 2 GB/minute, it will take Carla $160 / 2 = 80 \text{ minutes}$ to download the file.

Adding the time for the Windows restart, it will take Carla $80 + 20 = 100 \text{ minutes}$ to download the file. ❌

Round 2

ChatGPT : Based on the calculations provided by the other agent, the correct answer is: Before the restart, Carla was able to download 40% of the file, which is $0.4 * 200 \text{ GB} = 80 \text{ GB}$. At a rate of 2 GB/minute, it would take her: $\text{time} = 80 \text{ GB} \div 2 \text{ GB/minute} = 40 \text{ minutes}$

However, because of the restart, Carla had to start over. This means she had to download the entire 200 GB file. The time it takes to download the entire 200 GB file is: $\text{time} = 200 \text{ GB} \div 2 \text{ GB/minute} = 100 \text{ minutes}$

And since Carla had to wait 20 minutes for the Windows update to install, the total time it took to download the file is: $\text{total time} = 40 \text{ minutes} + 100 \text{ minutes} + 20 \text{ minutes} = 160 \text{ minutes}$. ✅

Figure 11: **Debate Between chatGPT and Bard** Illustration of debate between different models.

Experiments

- Questions:
 1. To what extent does multiagent debate improve reasoning?
 2. To what extent does multiagent debate improve factual validity?
 2. What design choices enable multiagent debate to improve language generation performance?

Experiment 1 (ChatGPT)

1. Arithmetic (Accuracy)
2. Grade School Math (Accuracy)
2. Chess Move Prediction (Pawn score from Stockfish)

Experiment 1 Results - Consensus

	Question: What is the result of $10+20*23+3-11*18$?	Question: What is the result of $3+7*9+19-21*18$?
Round 1	Agent 1: 269 ❌ Agent 2: 369 ❌	Agent 1: 378 ❌ Agent 2: -351 ❌ Agent 3: -357 ❌
Round 2	Agent 1: 275 ✅ Agent 2: 275 ✅	Agent 1: -293 ✅ Agent 2: -293 ✅ Agent 3: 19 ❌
	Question: What is the result of $4+23*6+24-24*12$?	Question: What is the result of $8+14*15+20-3*26$?
Round 1	Agent 1: -244 ❌ Agent 2: -146 ❌	Agent 1: 236 ❌ Agent 2: -214 ❌ Agent 3: 210 ❌
Round 2	Agent 1: -146 ❌ Agent 2: -122 ✅	Agent 1: 160 ✅ Agent 2: 160 ✅ Agent 3: 160 ✅
Round 3	Agent 1: -122 ✅ Agent 2: -122 ✅	Agent 1: 160 ✅ Agent 2: 160 ✅ Agent 3: 160 ✅

Figure 4: **Illustration of Solving Math.** Reasoning between agents is omitted.

	Question: Regina wrote 9 novels last year. If this is 3 quarters of the number of novels she has written this year, how many novels has she written this year?	Question: Dennis uses 1 pound of butter for every dozen croissants that he makes. He needs to make 6 dozen croissants. The grocery store currently has a promotion for buy one pound of butter get one half off. If the butter costs \$4.00 a pound, how much will it cost him to purchase 6 pounds of butter?
Round 1	Agent 1: 48 ❌ Agent 2: 12 ✅	Agent 1: 18 ✅ Agent 2: 30 ❌
Round 2	Agent 1: 12 ✅ Agent 2: 12 ✅	Agent 1: 18 ✅ Agent 2: 18 ✅

Figure 5: **Illustration of Solving Grade School Math.** Reasoning between agents omitted.

Experiment 1 Results - Performance

Model	Arithmetic (%) \uparrow	Grade School Math (%) \uparrow	Chess (Δ PS) \uparrow
Single Agent	67.0 \pm 4.7	77.0 \pm 4.2	91.4 \pm 10.6
Single Agent (Reflection)	72.1 \pm 4.5	75.0 \pm 4.3	102.1 \pm 11.9
Multi-Agent (Majority)	69.0 \pm 4.6	81.0 \pm 3.9	102.2 \pm 6.2
Multi-Agent (Debate)	81.8 \pm 2.3	85.0 \pm 3.5	122.9 \pm 7.6

Table 1: **Multiagent Debate Improves Reasoning** Multi-agent debate improves the reasoning abilities of language models. Multi-agent results in the table are run with 3 agents and two rounds of debate.

Experiment 2 (ChatGPT)

1. Bullet-Point Biography (Accuracy)
2. Massive Multitask Language Understanding/MMLU (Accuracy)
2. Chess Move Prediction (Pawn score from Stockfish)

Experiment 2 Results - Consensus

Round 1	<p>Prompt: What country was the computer scientist Tomas Lozano-Perez born in?</p> <p>Agent 1: Tomas Lozano-Perez, a renowned computer scientist, was born in Spain. ❌</p>	<p>Agent 2: Tomas Lozano-Perez, a renowned computer scientist, was born in Havana, Cuba. ✅</p>
Round 2	<p>Agent 1: I apologize for the confusion. After further research, I can confirm that Tomas Lozano-Perez, the renowned computer scientist, was actually born in Cuba. Thank you for bringing this to my attention. ✅</p>	<p>Agent 2: I apologize for the conflicting responses. Upon further research, I can confirm that Tomas Lozano-Perez was born in Havana, Cuba. ✅</p>

Figure 9: **Expressing Uncertainty with Multiple Answers.** For facts that a language model is uncertain about, different language agents generate different facts. Debate causes agents to converge to one fact that is more accurate, but not necessarily always factually correct.

	<p>Question: Six positive integers are written on the faces of a cube. Each vertex is labeled with the product of the three numbers on the faces adjacent to the vertex. If the sum of the numbers on the vertices is equal to 1001, then what is the sum of the numbers written on the faces? A) 18. B) 13. C) 1001. D) 31.</p>	<p>Question: You suspect that your patient has an enlarged submandibular salivary gland. You expect the enlarged gland: A) to be palpable intraorally. B) to be palpable extraorally. C) to be palpable both intra- and extraorally. D) only to be detectable by radiographical examination.</p>				
Round 1	<p>Agent 1: A ❌</p>	<p>Agent 2: C ❌</p>	<p>Agent 3: D ✅</p>	<p>Agent 1: C ✅</p>	<p>Agent 2: B ❌</p>	<p>Agent 3: C ✅</p>
Round 2	<p>Agent 1: D ✅</p>	<p>Agent 2: D ✅</p>	<p>Agent 3: D ✅</p>	<p>Agent 1: C ✅</p>	<p>Agent 2: C ✅</p>	<p>Agent 3: C ✅</p>

Figure 8: **Illustration of MMLU.** Illustration of debate when answering factual tasks. Reasoning omitted.

Experiment 2 Results - Performance

Model	Biographies	MMLU	Chess Move Validity
Single Agent	66.0 \pm 2.2	63.9 \pm 4.8	29.3 \pm 2.6
Single Agent (Reflection)	68.3 \pm 2.9	57.7 \pm 5.0	38.8 \pm 2.9
Multi-Agent (Debate)	73.8 \pm 2.3	71.1 \pm 4.6	45.2 \pm 2.9

Table 2: **Multiagent Debate Improves Factual Accuracy** Multi-agent debate improves the factual accuracy.

Experiment 3

1. Varying number of agents
2. Varying number of rounds
2. ChatGPT and Bard

Experiment 3 Results - Performance

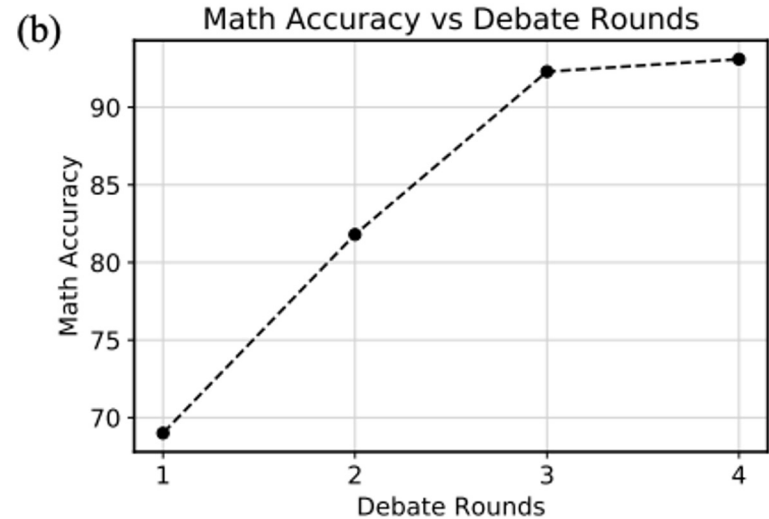
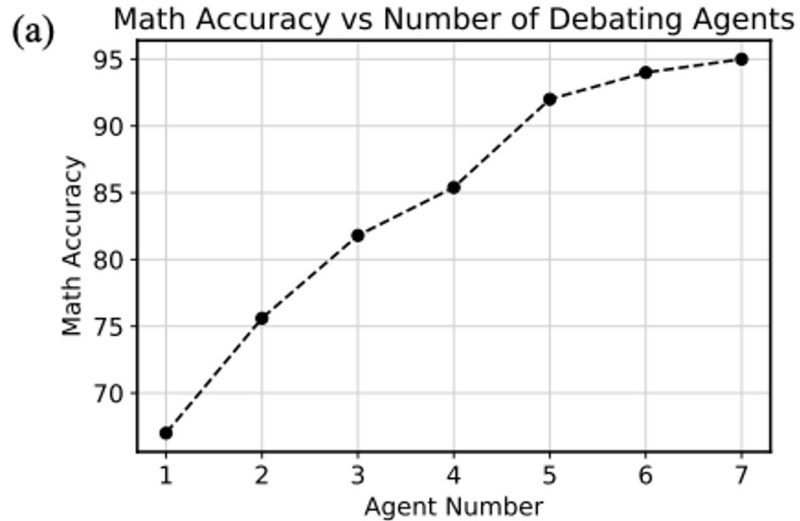


Figure 10: (a) **Performance with Increased Agents.** Performance improves as the number of underlying agents involved in debate increases. (b) **Performance with Increased Rounds.** Performance rises as the number of rounds of underlying debate increases.

Experiment 3 Results - Performance

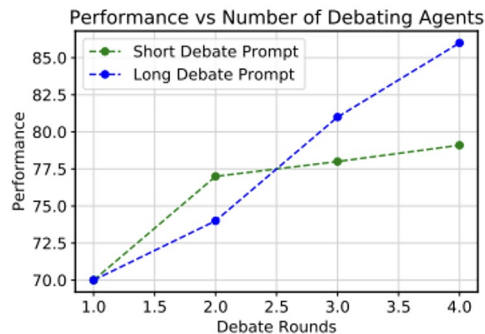


Figure 12: **Performance vs Debate Length.** Prompts which induce longer debate improve performance.

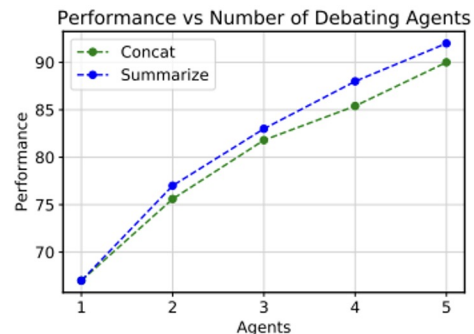


Figure 13: **Effect of Summarization.** When there are many agents in a debate, responses from other agents may be first summarized and then given as context, reducing context length. This operation improves performance.

Analysis - Performance

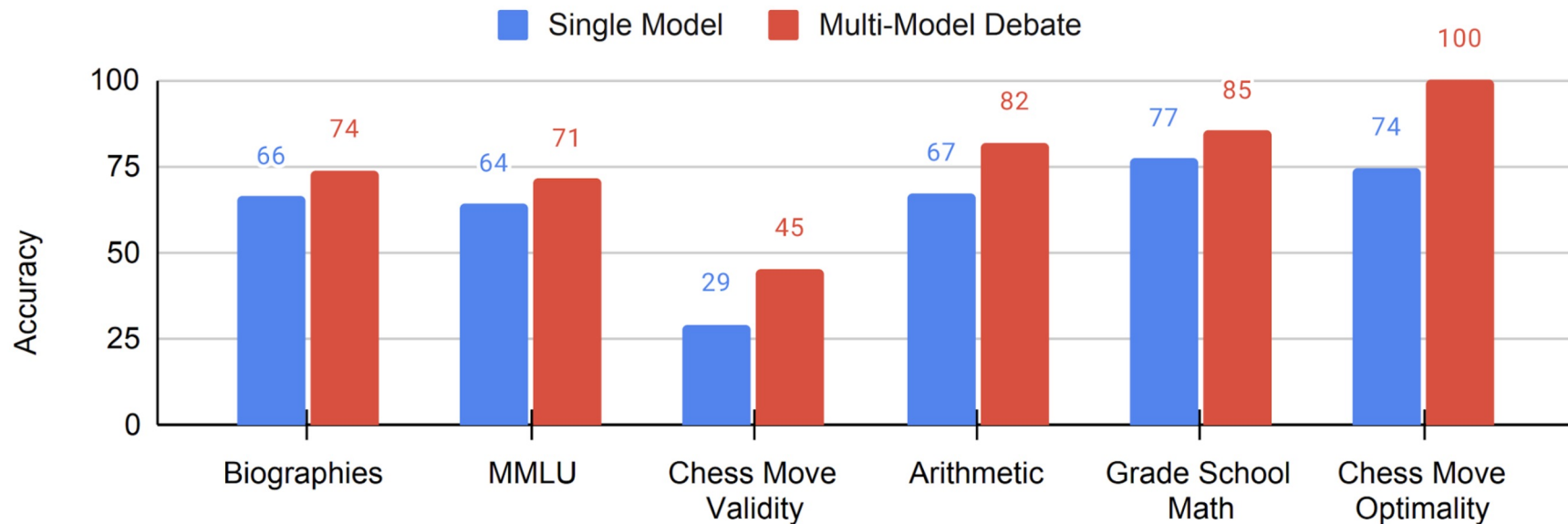


Figure 1: **Multiagent Debate Improves Reasoning and Factual Accuracy.** Accuracy of traditional inference and our multi-agent debate over six benchmarks (chess move optimality reported as a normalized score)

Analysis – Performance

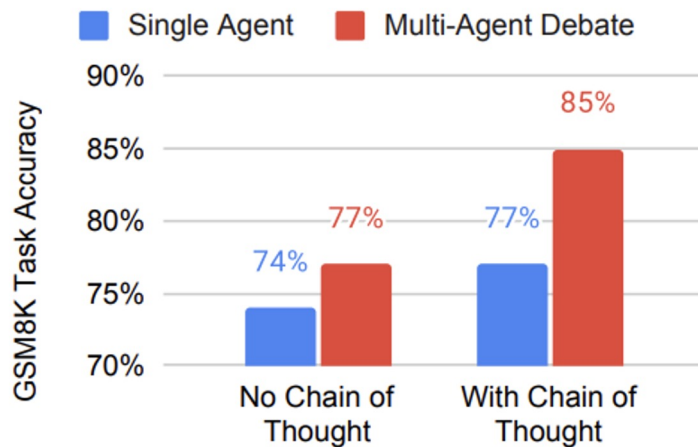


Figure 6: **Synergy with Other Methods.** Performance of debate increases with use of Chain of Thought prompting.

Discussion

- Simple and effective
- However, more computationally expensive
- More difficult to process all input, especially with more rounds
- Convergence \neq Correct

Wrap Up

Hallucination

- LLMs are known to regurgitate inaccurate information, often that is not even present in the input document
- Known as “hallucination”

Summary – Four Papers

1. Context-Aware Decoding is shown to mitigate hallucinations
1. Hallucinations can snowball into more mistakes that otherwise would not occur
 - Better prompts or algorithms
2. Bayesian Sequential Estimation is also useful for hallucination
1. Multiple agents can come together for a more accurate answer, both for hallucination and reasoning

Future Work - Challenges

- Even in the best performance, LLMs can still hallucinate
- This can impact trust and adoption of these models, even in areas where they excel

Future Work - Challenges

- Difficult to run more complex experiments, as fact-checking is resource-intensive and ambiguous (Havana vs. Spain)
- Many of the largest/powerful models are black-box and therefore hard to theoretically analyze

Future Work - Challenges

- All strategies defined here increase computation, time, or both
- This can also cause these methods to gain less adoption

Future Work - Opportunities

- Need to explore different application scenarios
- There is an opportunity to explore the role of the input data in hallucination

Future Work - Opportunities

- Need to explore the relationship between hallucination and creativity
- In psychology, there is plenty of work on the interplay between hallucination and imagination, this may be an opportunity to open up creativity as an emergent property

References

- Shi, Weijia, et al. "Trusting your evidence: Hallucinate less with context-aware decoding." *arXiv preprint arXiv:2305.14739* (2023).
- Zhang, Muru, et al. "How language model hallucinations can snowball." *arXiv preprint arXiv:2305.13534* (2023).
- Wang, Xiaohua, et al. "Hallucination detection for generative large language models by bayesian sequential estimation." *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 2023.
- Du, Yilun, et al. "Improving factuality and reasoning in language models through multiagent debate." *arXiv preprint arXiv:2305.14325* (2023).

Thank you!

Any Questions?