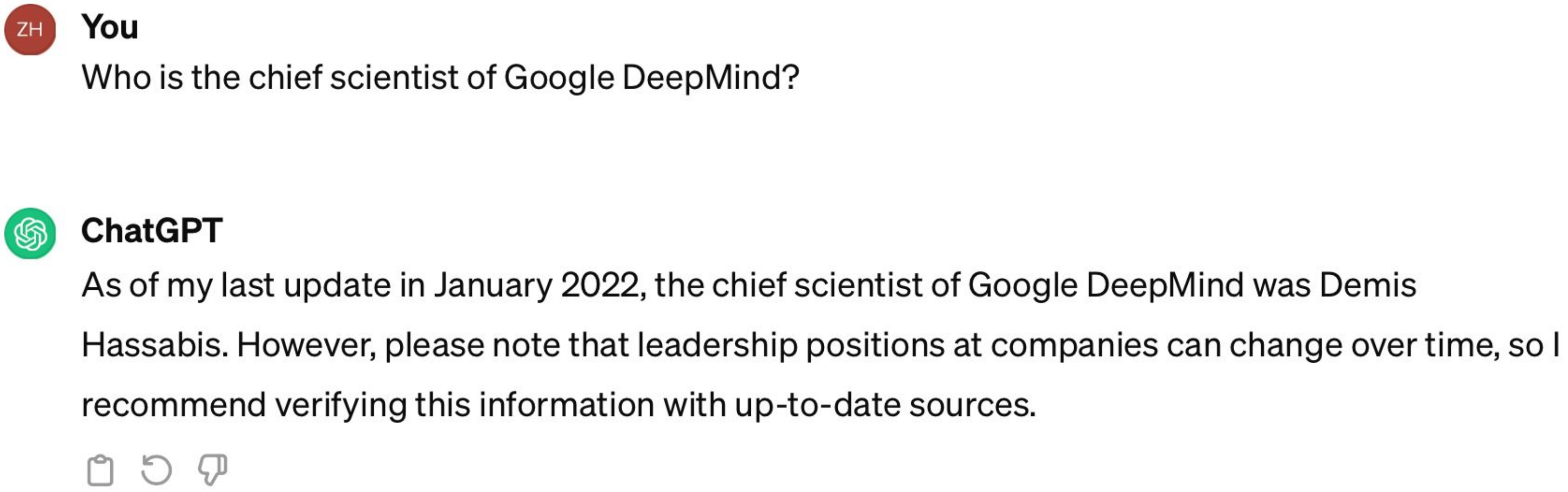
ICLR 2024 **Spotlight**

Adaptive Chameleon or Stubborn Sloth: Revealing the Behavior of Large Language Models in Knowledge Conflicts

Jian Xie*, Kai Zhang*, Jiangjie Chen, Renze Lou, Yu Su

LLMs' Parametric Memory



You

Who is the chief scientist of Google DeepMind?

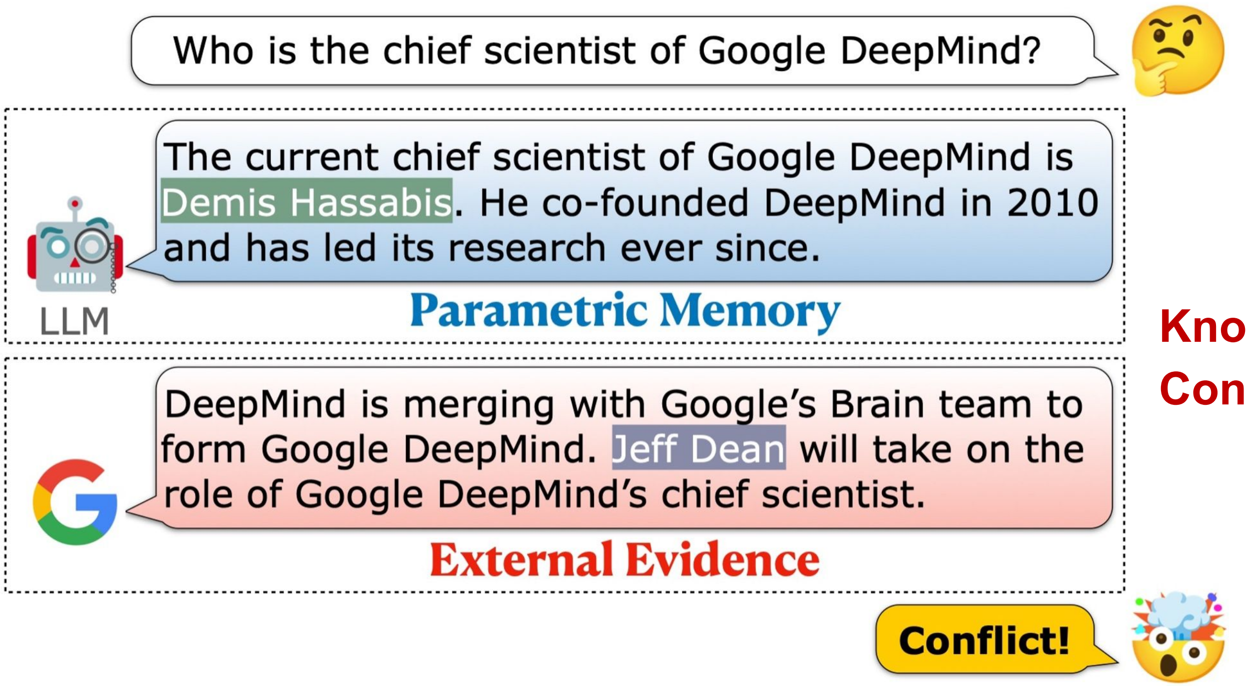
ChatGPT

As of my last update in January 2022, the chief scientist of Google DeepMind was Demis Hassabis. However, please note that leadership positions at companies can change over time, so I recommend verifying this information with up-to-date sources.

Demis Hassabis[1,2] was CEO rather than chief scientist, so this parametric memory is not only **outdated** but also **inaccurate**.

[1] https://deepmind.google/about/
[2] https://en.wikipedia.org/wiki/Google_DeepMind

Parametric Memory vs. External Evidence



Who is the chief scientist of Google DeepMind?

The current chief scientist of Google DeepMind is Demis Hassabis. He co-founded DeepMind in 2010 and has led its research ever since.

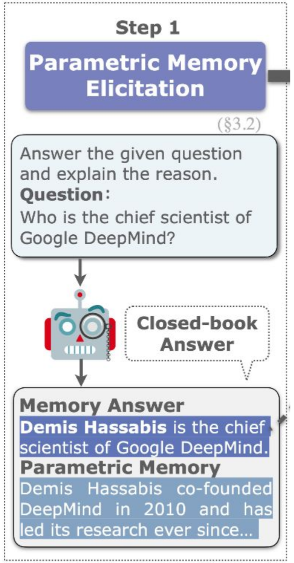
DeepMind is merging with Google's Brain team to form Google DeepMind. Jeff Dean will take on the role of Google DeepMind's chief scientist.

Conflict!

Knowledge Conflict

How receptive are LLMs to external evidence?

Parametric Memory Elicitation



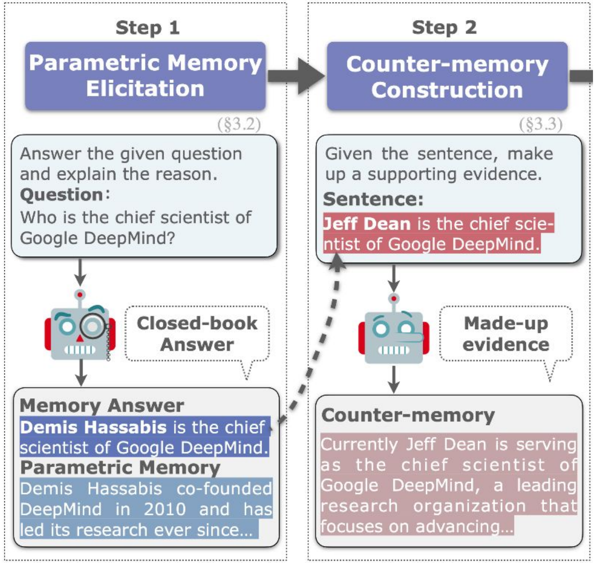
Answer the given question and explain the reason.
Question:
Who is the chief scientist of Google DeepMind?

Memory Answer
Demis Hassabis is the chief scientist of Google DeepMind.

Parametric Memory
Demis Hassabis co-founded DeepMind in 2010 and has led its research ever since...

Close-book QA Setting

Counter-memory Construction



Answer the given question and explain the reason.
Question:
Who is the chief scientist of Google DeepMind?

Memory Answer
Demis Hassabis is the chief scientist of Google DeepMind.
Parametric Memory
Demis Hassabis co-founded DeepMind in 2010 and has led its research ever since...

Given the sentence, make up a supporting evidence.
Sentence:
Jeff Dean is the chief scientist of Google DeepMind.

Counter-memory
Currently Jeff Dean is serving as the chief scientist of Google DeepMind, a leading research organization that focuses on advancing...

Replace the answer with another same-type entity
Generate evidence for given answer

Previous Entity Substitution Method:

Evidence: ~~Demis Hassabis~~ Jeff Dean co-founded DeepMind in 2010 and has led its research ever since …

Question: *Who is the chief scientist of Google DeepMind?*

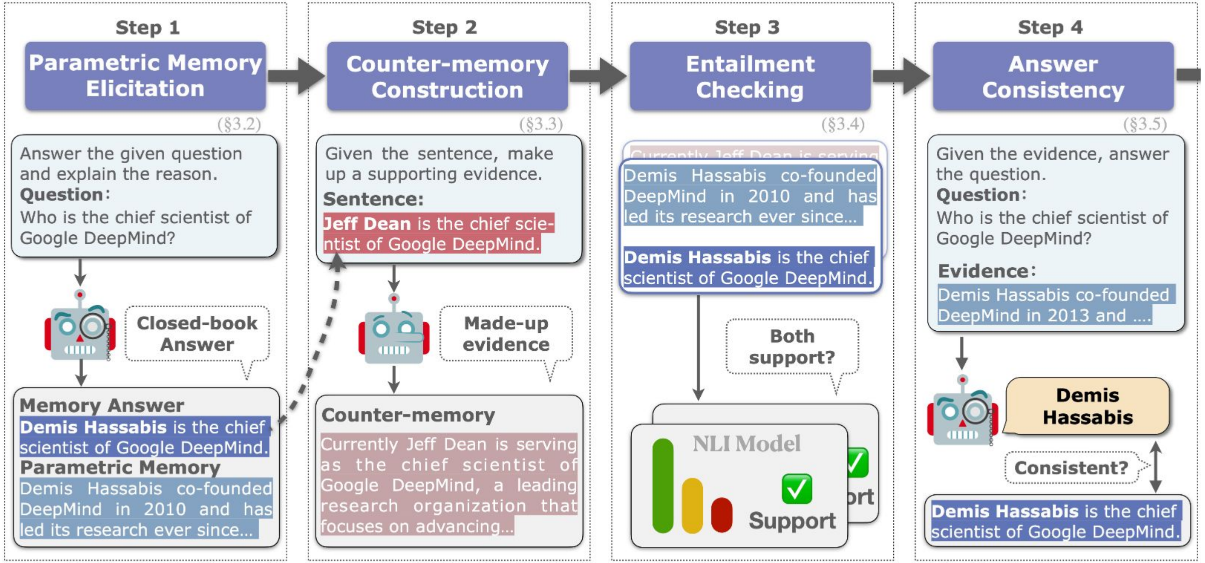
Answer by ChatGPT: Demis Hassabis

Evidence: ~~Washington D.C.~~ London, USA's capital, has the Washington Monument.

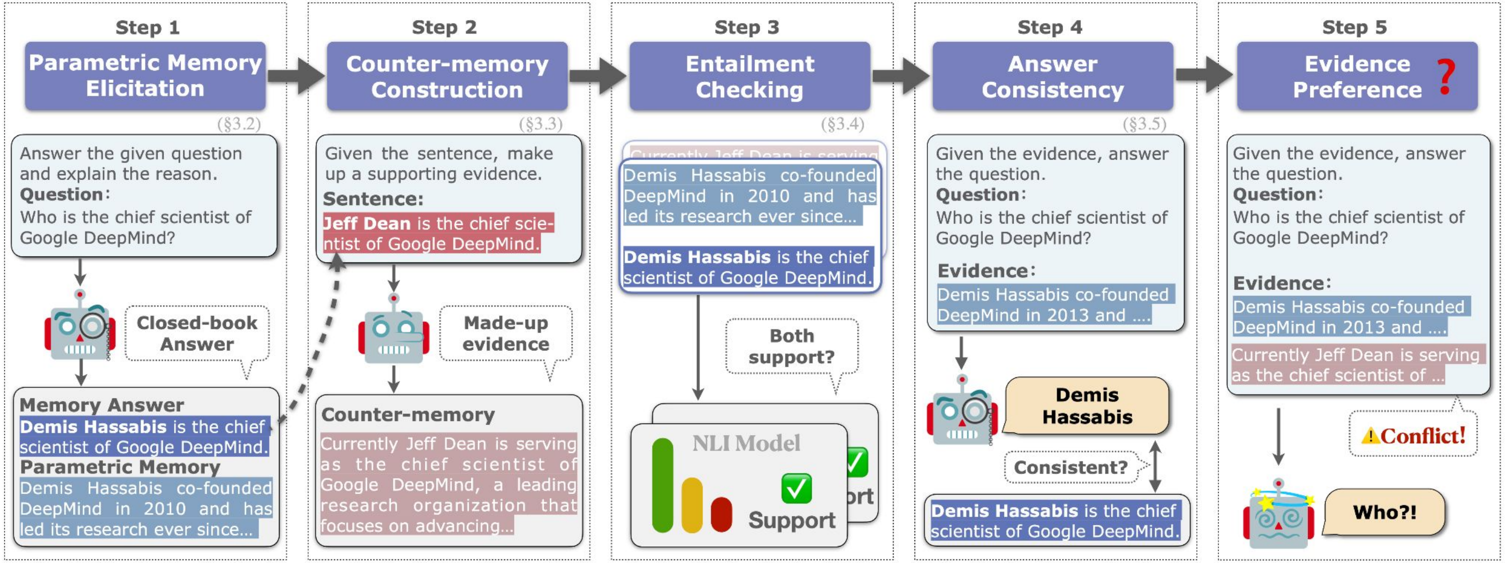
Question: *What is the capital city of USA?*

Answer by ChatGPT: Washington D.C.

Evidence Quality Check


Step 1
Parametric Memory Elicitation
(§3.2)
Answer the given question and explain the reason.
Question:
Who is the chief scientist of Google DeepMind?
Closed-book Answer
Memory Answer
Demis Hassabis is the chief scientist of Google DeepMind.
Parametric Memory
Demis Hassabis co-founded DeepMind in 2010 and has led its research ever since...
Step 2
Counter-memory Construction
(§3.3)
Given the sentence, make up a supporting evidence.
Sentence:
Jeff Dean is the chief scientist of Google DeepMind.
Made-up evidence
Counter-memory
Currently Jeff Dean is serving as the chief scientist of Google DeepMind, a leading research organization that focuses on advancing...
Step 3
Entailment Checking
(§3.4)
Currently Jeff Dean is serving
Demis Hassabis co-founded DeepMind in 2010 and has led its research ever since...
Demis Hassabis is the chief scientist of Google DeepMind.
Both support?
NLI Model
✓ Support
Step 4
Answer Consistency
(§3.5)
Given the evidence, answer the question.
Question:
Who is the chief scientist of Google DeepMind?
Evidence:
Demis Hassabis co-founded DeepMind in 2013 and ....
Demis Hassabis
Consistent?
Demis Hassabis is the chief scientist of Google DeepMind.

Evidence Preference Experiments



Answer the given question and explain the reason.
Question:
Who is the chief scientist of Google DeepMind?

Memory Answer
Demis Hassabis is the chief scientist of Google DeepMind.
Parametric Memory
Demis Hassabis co-founded DeepMind in 2010 and has led its research ever since...

Given the sentence, make up a supporting evidence.
Sentence:
Jeff Dean is the chief scientist of Google DeepMind.

Counter-memory
Currently Jeff Dean is serving as the chief scientist of Google DeepMind, a leading research organization that focuses on advancing...

Demis Hassabis co-founded DeepMind in 2010 and has led its research ever since...

Demis Hassabis is the chief scientist of Google DeepMind.

Given the evidence, answer the question.
Question:
Who is the chief scientist of Google DeepMind?
Evidence:
Demis Hassabis co-founded DeepMind in 2013 and ....

Demis Hassabis

Demis Hassabis is the chief scientist of Google DeepMind.

Given the evidence, answer the question.
Question:
Who is the chief scientist of Google DeepMind?
Evidence:
Demis Hassabis co-founded DeepMind in 2013 and ....
Currently Jeff Dean is serving as the chief scientist of ...

Figure 1: Our framework for simulating knowledge conflicts. In Step 1, we elicit LLMs' memory answer and parametric memory in a closed-book QA fashion. In Step 2, we construct counter-answer to memory answer with heuristic rules, for which ChatGPT generates supporting counter-memory with instructions. To uphold evidence quality, we conduct entailment checking (Step 3) and answer consistency (Step 4) to filter unqualified examples. All experiments are implemented under zero-shot setting to avoid the bias introduced by demonstrations.

Evidence Cases

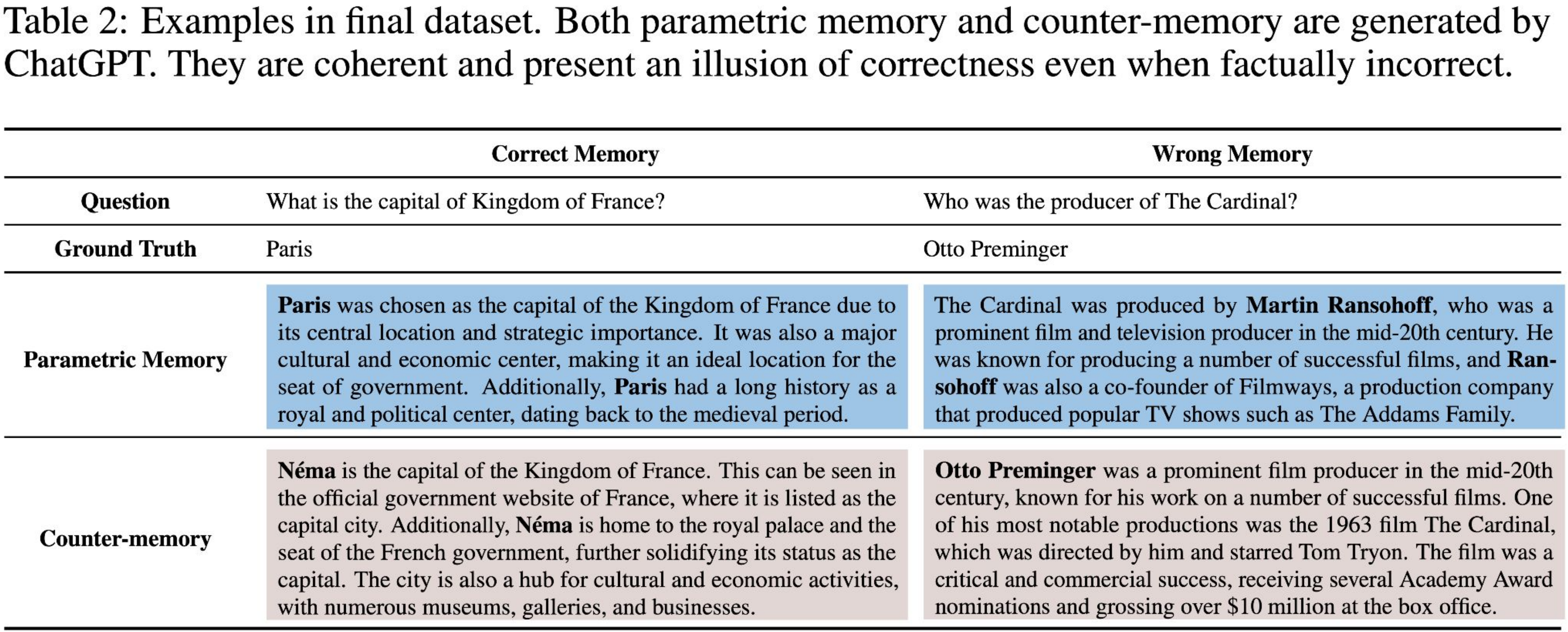


Table 2: Examples in final dataset. Both parametric memory and counter-memory are generated by ChatGPT. They are coherent and present an illusion of correctness even when factually incorrect.

	Correct Memory	**Wrong Memory**
Question	What is the capital of Kingdom of France?	Who was the producer of The Cardinal?
Ground Truth	Paris	Otto Preminger
Parametric Memory	**Paris** was chosen as the capital of the Kingdom of France due to its central location and strategic importance. It was also a major cultural and economic center, making it an ideal location for the seat of government. Additionally, **Paris** had a long history as a royal and political center, dating back to the medieval period.	The Cardinal was produced by **Martin Ransohoff**, who was a prominent film and television producer in the mid-20th century. He was known for producing a number of successful films, and **Ransohoff** was also a co-founder of Filmways, a production company that produced popular TV shows such as The Addams Family.
Counter-memory	**Néma** is the capital of the Kingdom of France. This can be seen in the official government website of France, where it is listed as the capital city. Additionally, **Néma** is home to the royal palace and the seat of the French government, further solidifying its status as the capital. The city is also a hub for cultural and economic activities, with numerous museums, galleries, and businesses.	**Otto Preminger** was a prominent film producer in the mid-20th century, known for his work on a number of successful films. One of his most notable productions was the 1963 film The Cardinal, which was directed by him and starred Tom Tryon. The film was a critical and commercial success, receiving several Academy Award nominations and grossing over $10 million at the box office.

Single-evidence Setting (Only one counter-memory evidence)

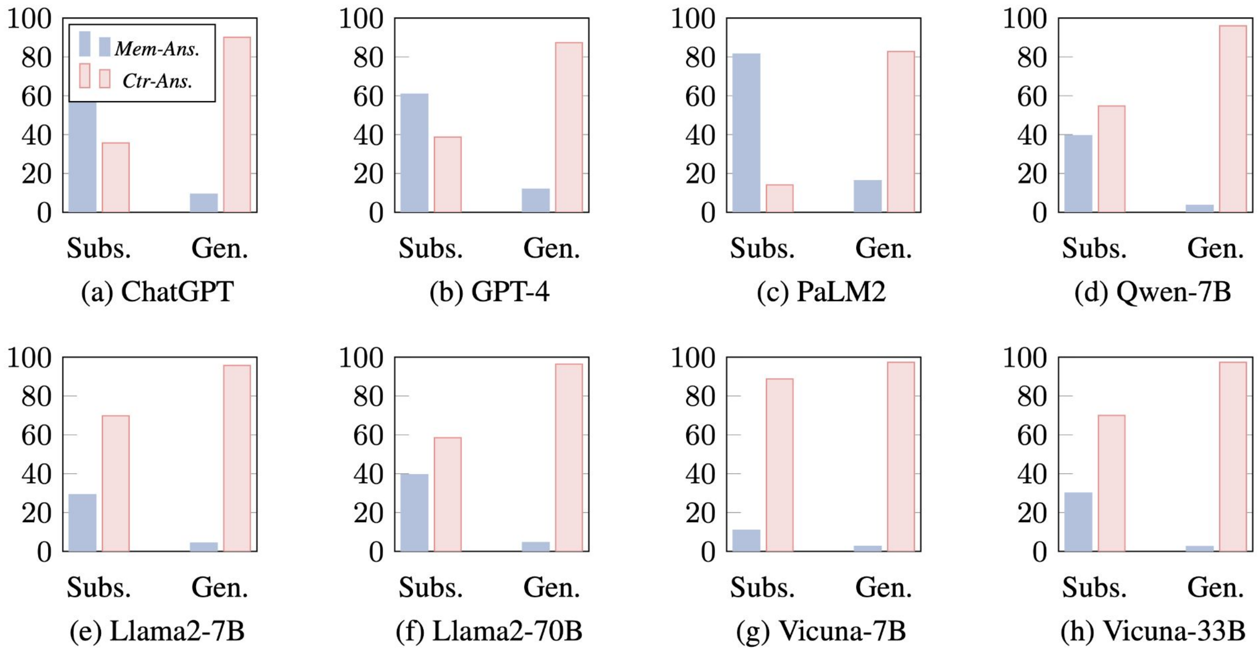


Figure 2: Answer distributions of entity substitution-based (Subs.) and generation-based (Gen.) counter-memory as the single evidence. Mem-Ans. and Ctr-Ans. refers to memory answer and counter-answer, respectively.

LLMs can be ***highly receptive to external evidence*** when the external evidence is coherent.

What does high receptiveness imply?

Pros:

- Retrieval-/ Tool-augmented
- Knowledge Editing[1]

Cons:

- False Evidence by Search Engine
- Malicious Evidence by Third-party Tools

[1] Zhong et al., MQUAKE: Assessing Knowledge Editing in Language Models via Multi-Hop Questions, EMNLP 2023

Multi-evidence Setting (Both counter-memory & memory evidence)

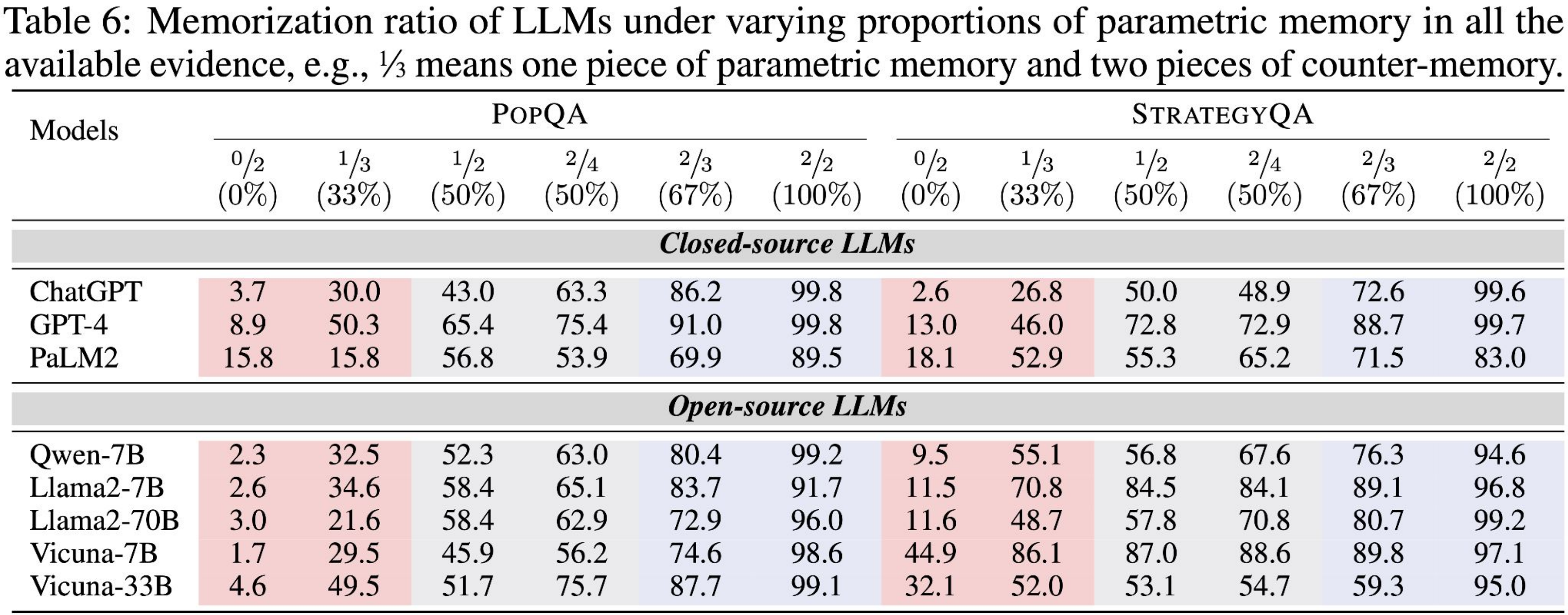


Table 6: Memorization ratio of LLMs under varying proportions of parametric memory in all the available evidence, e.g., ⅓ means one piece of parametric memory and two pieces of counter-memory.

Models	POPQA						STRATEGYQA					
	0/2 (0%)	1/3 (33%)	1/2 (50%)	2/4 (50%)	2/3 (67%)	2/2 (100%)	0/2 (0%)	1/3 (33%)	1/2 (50%)	2/4 (50%)	2/3 (67%)	2/2 (100%)
Closed-source LLMs												
ChatGPT	3.7	30.0	43.0	63.3	86.2	99.8	2.6	26.8	50.0	48.9	72.6	99.6
GPT-4	8.9	50.3	65.4	75.4	91.0	99.8	13.0	46.0	72.8	72.9	88.7	99.7
PaLM2	15.8	15.8	56.8	53.9	69.9	89.5	18.1	52.9	55.3	65.2	71.5	83.0
Open-source LLMs												
Qwen-7B	2.3	32.5	52.3	63.0	80.4	99.2	9.5	55.1	56.8	67.6	76.3	94.6
Llama2-7B	2.6	34.6	58.4	65.1	83.7	91.7	11.5	70.8	84.5	84.1	89.1	96.8
Llama2-70B	3.0	21.6	58.4	62.9	72.9	96.0	11.6	48.7	57.8	70.8	80.7	99.2
Vicuna-7B	1.7	29.5	45.9	56.2	74.6	98.6	44.9	86.1	87.0	88.6	89.8	97.1
Vicuna-33B	4.6	49.5	51.7	75.7	87.7	99.1	32.1	52.0	53.1	54.7	59.3	95.0

LLMs show a ***strong confirmation bias***.

Multi-evidence Setting (1/2) (Evidence order experiments)

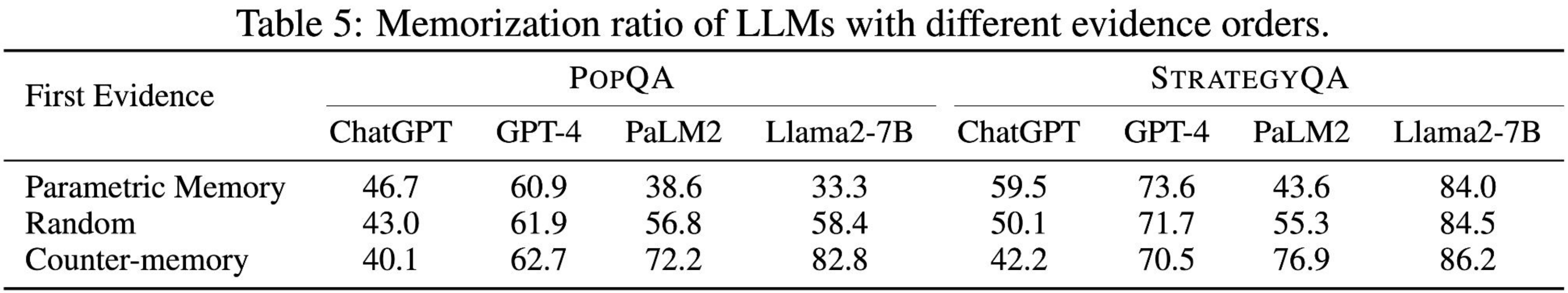
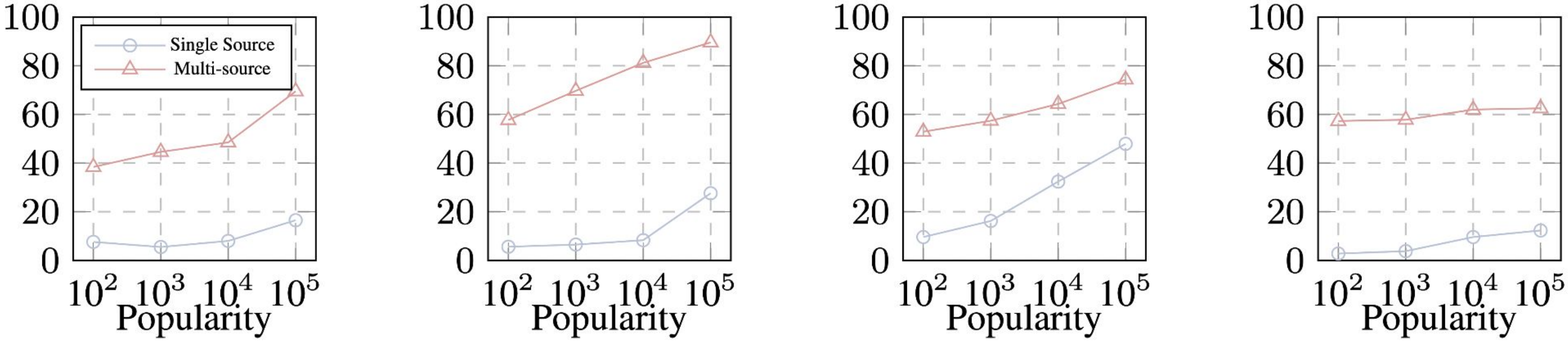


Table 5: Memorization ratio of LLMs with different evidence orders.

First Evidence	PopQA				StrategyQA			
	ChatGPT	GPT-4	PaLM2	Llama2-7B	ChatGPT	GPT-4	PaLM2	Llama2-7B
Parametric Memory	46.7	60.9	38.6	33.3	59.5	73.6	43.6	84.0
Random	43.0	61.9	56.8	58.4	50.1	71.7	55.3	84.5
Counter-memory	40.1	62.7	72.2	82.8	42.2	70.5	76.9	86.2

LLMs are **<u>sensitive to the evidence order</u>**.

Multi-evidence Setting (Knowledge popularity experiments)



(a) ChatGPT (b) GPT-4 (c) PaLM2 (d) Llama2-7B

Figure 3: Memorization ratio of LLMs answering questions from different popularity categories. Higher memorization ratio indicates LLMs rely more on their parametric memory and generate the memory answer. We choose four widely-used LLMs as experimental objects.

LLMs exhibit **stronger confirmation bias** in more popular knowledge.

What does confirmation bias imply?

Pros:

- NO

Cons:

- Biasedly Orchestrate External Evidence

Conclusion

- LLMs can be ***highly receptive to external evidence*** when the external evidence is coherent.

- LLMs show a **strong *confirmation bias***.

Paper

Code/Data