



Can LLMs Remember

Questions from last time

In real-world, life does not always go planned. How LLM can address this problem when LLM does planning?

There are many approaches for this.

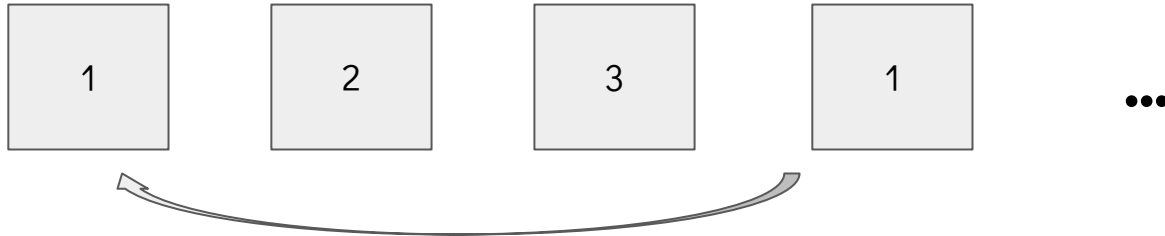
1. If we keep feed real-time information to LLM, it can adjust plans and recommendations as new information becomes available.
2. As LLMs are exposed to more data over time, LLM can improve planning with real-world feedback.
3. A human-in-the-loop approach ensures that human experience and intuition are also considered in the planning process.

Some approaches of LLM planning, list many actions and evaluate them and select. It might require huge computation, are there any way to reduce this waste of computation?

1. You can eliminate less promising actions from consideration by implementing heuristics or rules-of-thumb . Heuristics can be based on past data, domain knowledge, or predefined criteria.
2. Cache results of common queries or evaluations so they can be reused in similar future scenarios. This reduces the need to recompute the same information.
3. Instead of evaluating all actions simultaneously, process them incrementally. Start with a smaller set of actions, evaluate them, and then decide whether further analysis is needed. This approach can reduce the computational load.

Let's play a game!

Say “repeat” if the number on screen matches the number 3 steps earlier

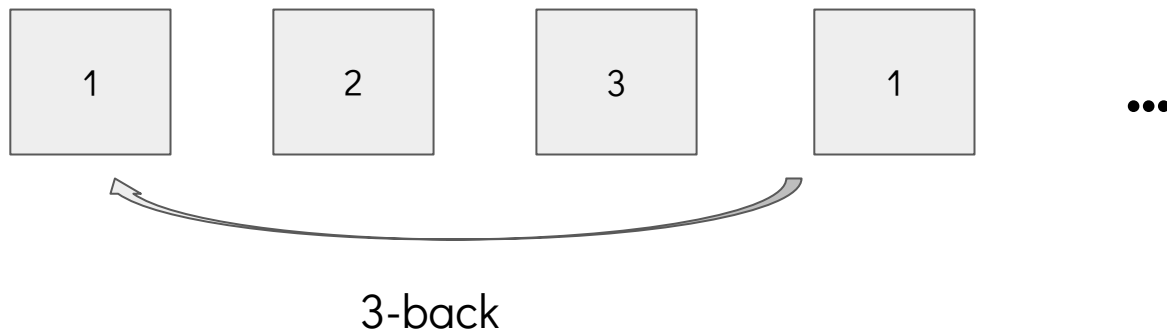


How many repeats did you get?

How many repeats did you get?
There are 6

What we just did

n-back task
(One way to measure memory)



Question

How do we test an LLMs memory?

Question

How do we *normally* test an LLMs memory?

What usually happens...



You

Who is the first president of the United States?

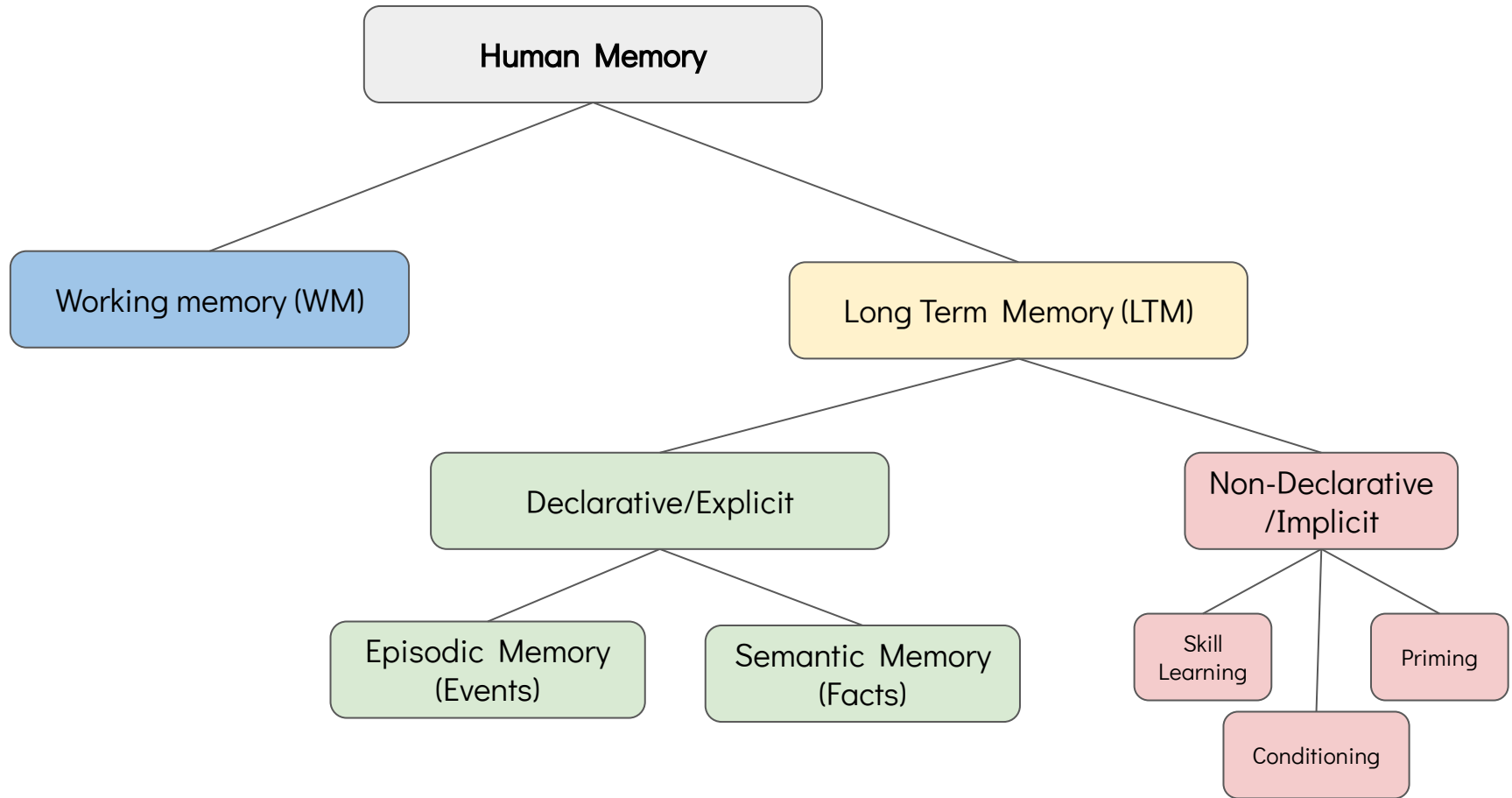


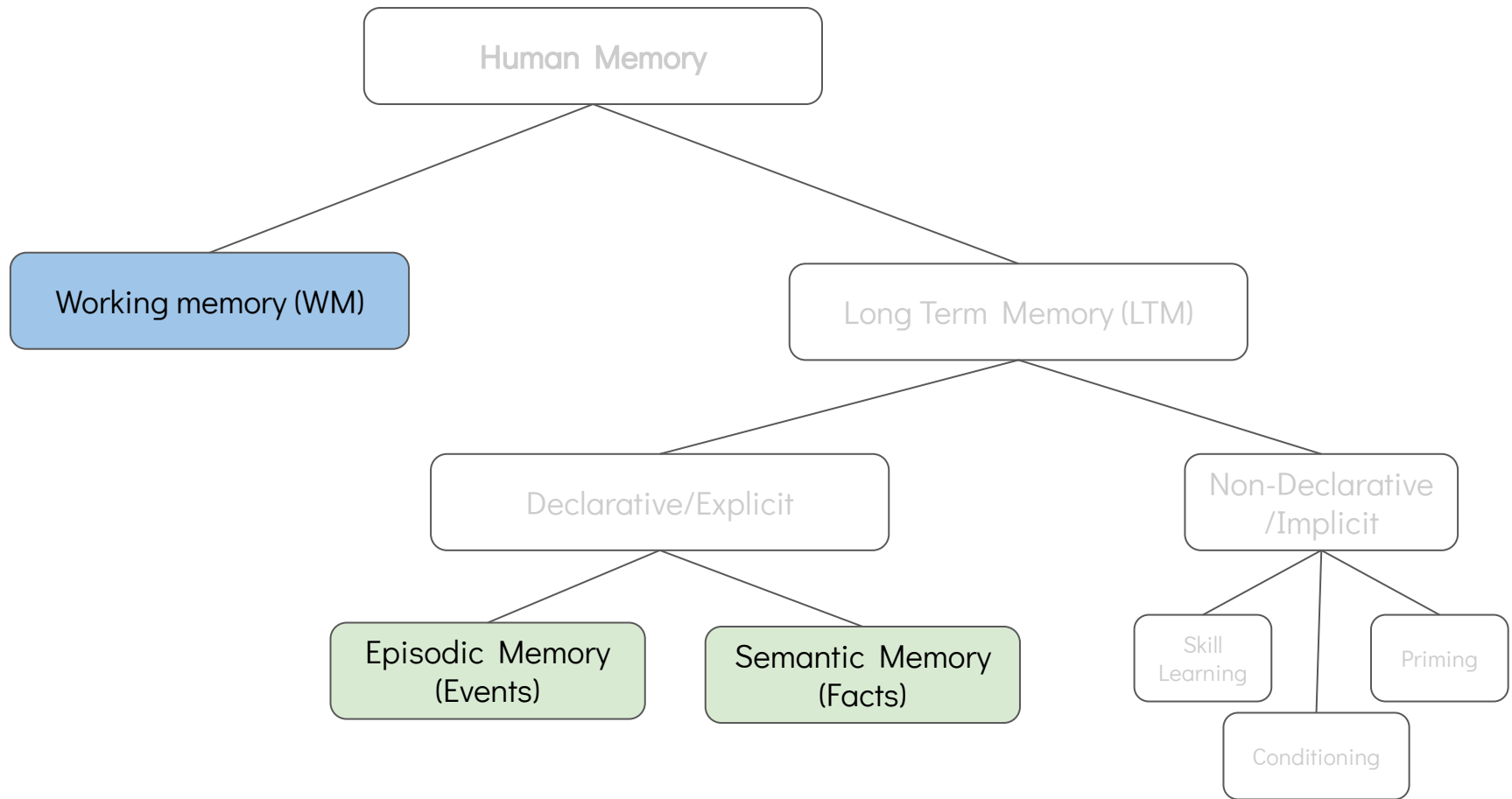
ChatGPT

The first president of the United States was George Washington.



They are both “Memory”
but not quite the same

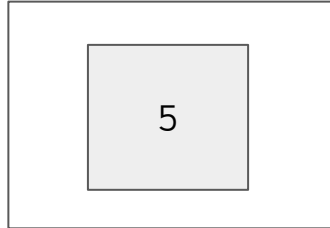






Working memory

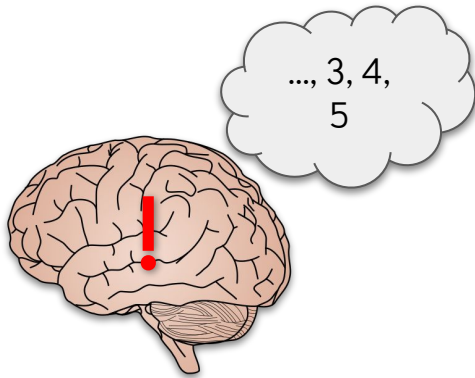
Working memory



A system for maintaining and manipulating active representations

Working memory

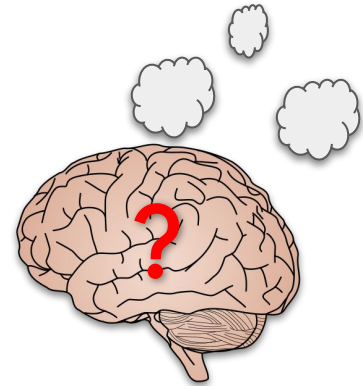
Quick to recall



Limited capacity

7 ± 2 items
or
Spoken in **2 sec**

Quick to forget



Testing Working memory

Span task

- The number of items that can be recalled with at least 50% accuracy

45

0988

493759

569375937

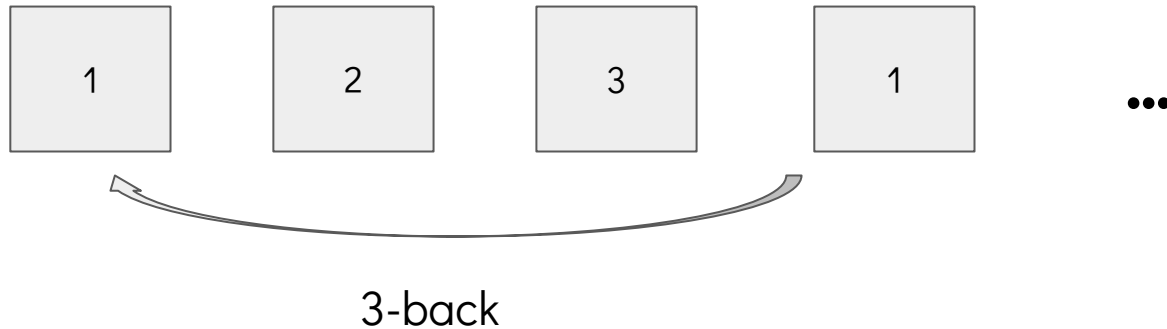
43808945679126789453678

0962020987487

Testing Working memory

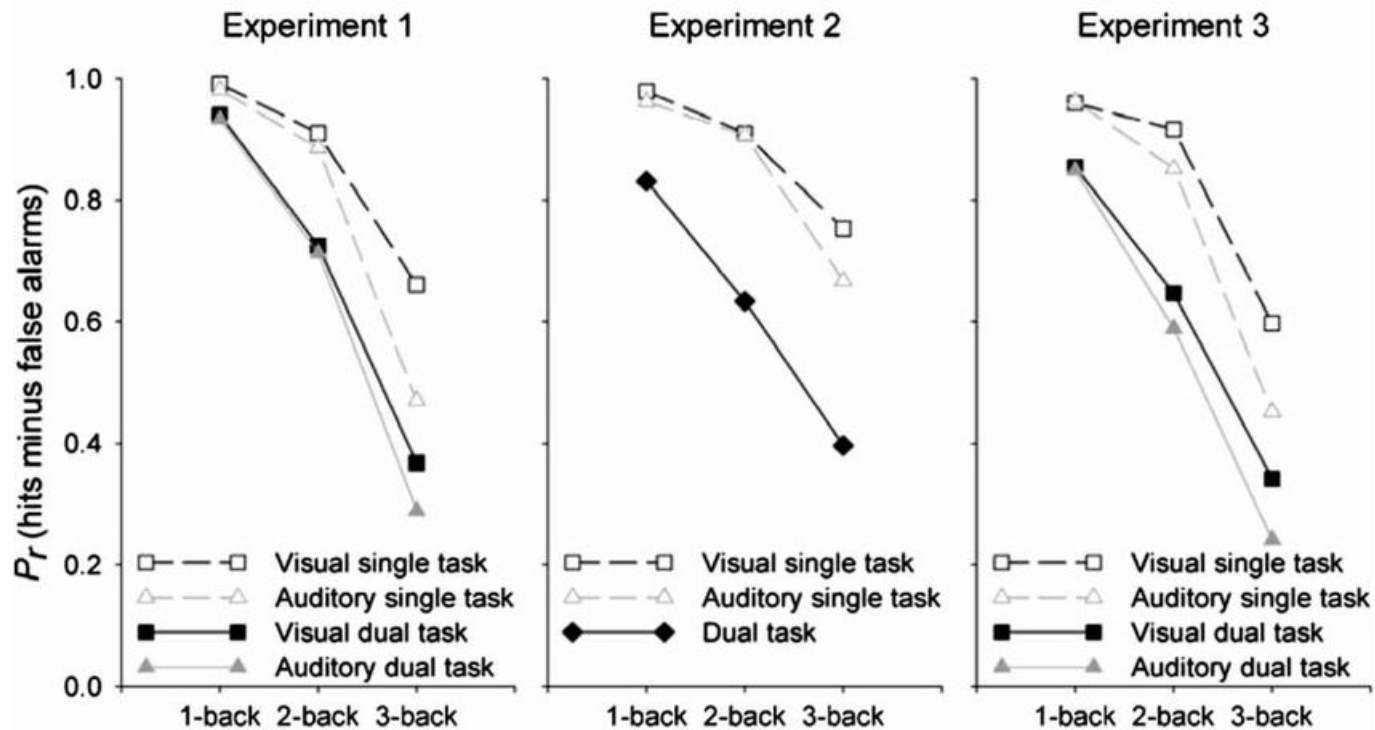
N-Back task

- Identify whether there is a repeat to a stimulus n -steps ago
- The higher then n the harder



Human performance

Jaeggi et al. (2010)



What about LLMs?

ChatGPT and the span task



You

Please recite this number in reverse

12994572879198128974537892126743276167829078519787832651674357895189389



ChatGPT

The reversed number is

7354897634906785943697023469024738974083456787641463980536908312368

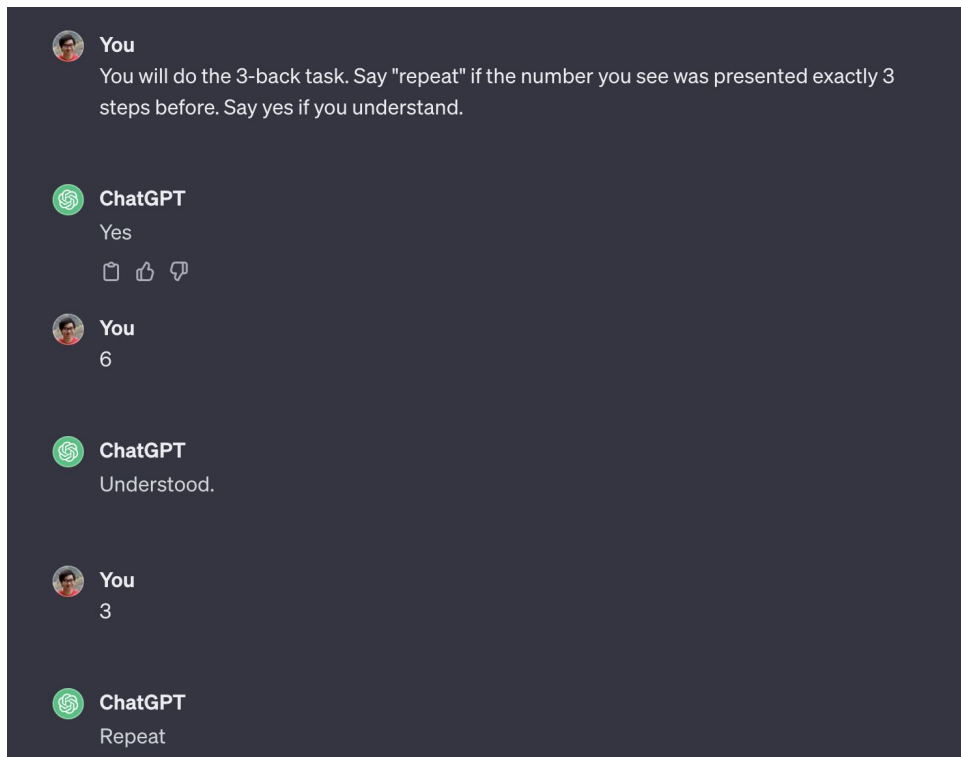
72614169581350986432766521378753478372187632723676123416854373241678513

2870923984652106876789785017028976117805128792789454811927978519324859

7543215636323470984237905984701234624351789285941798725729459921.



ChatGPT and the 3-back task



- 4/6 true repeats
- 3 False repeats

...not that great

I tried again

N You

What is the n-back task

 ChatGPT

The n-back task is a cognitive training exercise designed to assess and improve working memory. Working memory is the system responsible for temporarily holding and manipulating information in the mind. In the n-back task, participants are presented with a sequence of stimuli, often visual or auditory, and they are required to indicate when the current stimulus matches the one presented "n" steps back in the sequence.

Here's how it typically works:

1/6 correct
4 false calls

I tried again

N You

What is the n-back task

ChatGPT

The n-back task is a cognitive training exercise that tests working memory. Working memory is the part of memory that allows us to hold and manipulate information in the mind over short periods of time. In the n-back task, you are presented with a sequence of stimuli, often numbers or words, and you are asked to indicate when the current stimulus matches the stimulus that was presented n positions back in the sequence.

Here's how it typically works:

N You

I will ask you to do the 3-back task. I will present you a number, and I want you to either say 'repeat' if the number matches the number that presented 3 positions back. If not just say 'no repeat'. Say yes if you understand.

ChatGPT

Yes.

N You

I will now begin the task

ChatGPT

Sure, I'm ready.

I tried again

N You

What is the n-back task



ChatGPT

The n-back task is a cognitive working memory. Working memory and manipulating information presented with a sequence of indicate when the current stimulus sequence.

Here's how it typically works:

N You

- 1/6 true repeats
- 4 False repeats

...Even worse!

I will now begin the task



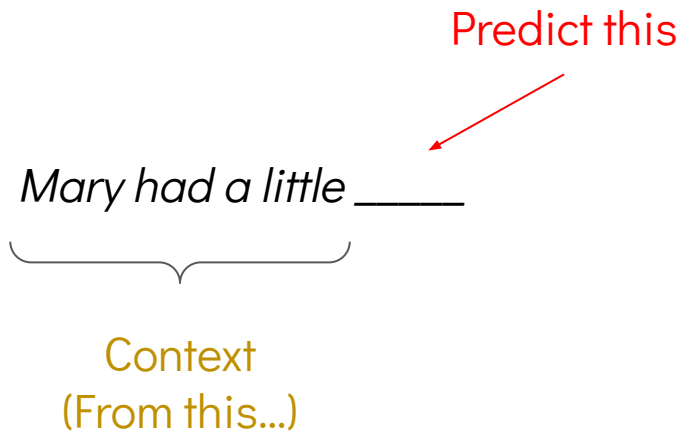
ChatGPT

Sure, I'm ready.

...I will give you a number, and I want you to either repeat the number that presented 3 positions back. If not

What are LLMs doing?

(The language modeling task)



What are LLMs doing?

(The language modeling task)

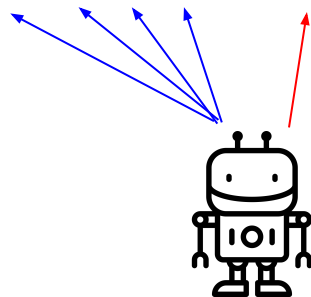
Mary had a little LAMB



Context
(From this...)

The attention* mechanism

Mary had a little **LAMB**



*Causal attention

Context as 'working memory'

(sort of)

Predict this

Mary had a little _____



Do LLMs have Working Memory?

Do LLMs have Working Memory?



Written by Joshua J.

Updated over a week ago

There are a few different GPT-4 models to choose from.

One key difference between models is the context length. This is the length of the prompt plus the maximum number of tokens in the completion. Our standard GPT-4 model offers 8,000 tokens for the context. We also offer an extended 32,000 token context-length model, which we are rolling out separately to the 8k model.

YES!

It *COULD* be better than a humans?

The background of the slide is a solid orange color with a pattern of faded, overlapping photographs. The photos are arranged in a grid-like fashion, tilted at an angle, and show various people in different settings, such as portraits, group shots, and candid moments. The overall effect is a sense of a memory album or a collection of personal experiences.

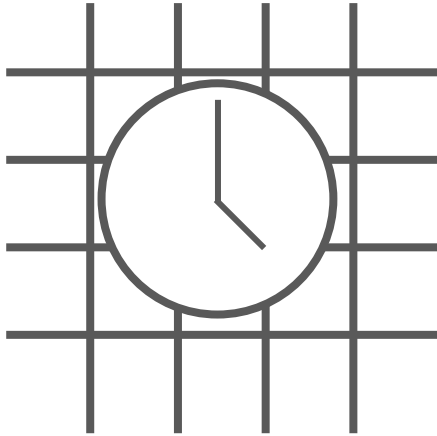
Episodic memory

“Co-occurring features in space & time”

(What, When, Where)

Episodic memory

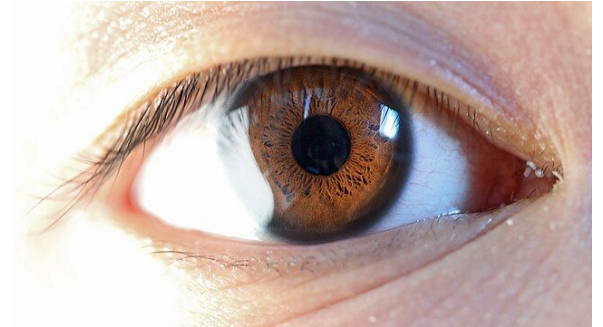
Tagged with spatial
and temporal context



Experienced
personally

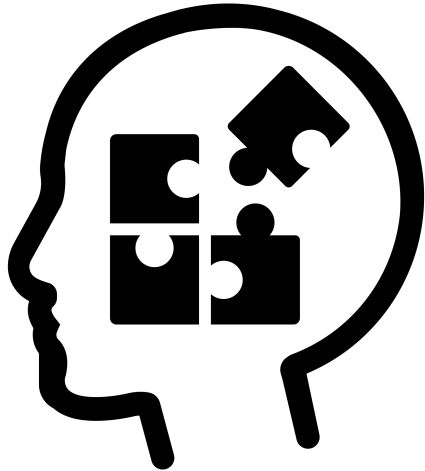


Learned in
single exposure



Episodic memory

Degrades with time
and new information



Our brain reconstructs the
missing pieces

Emotions affect
vividness (not accuracy)



E.g. Memory of tragedies

What about LLMs?

Asking an LLM on past conversations

You

What has happened in our conversation so far?

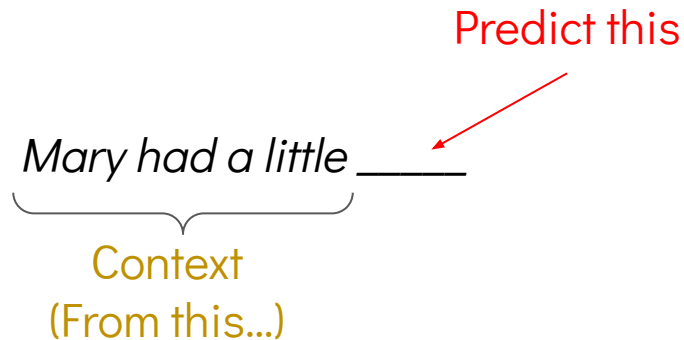
ChatGPT

In our conversation, we discussed the n-back task, a cognitive training exercise used to assess and improve working memory. You then initiated a simulated 5-back task, where you provided a sequence of numbers, and I indicated whether there was a repetition of a number from 5 turns before. We've gone through several numbers, with some being repeated and others not. If you have any specific questions or if there's anything else you'd like to discuss, feel free to let me know!

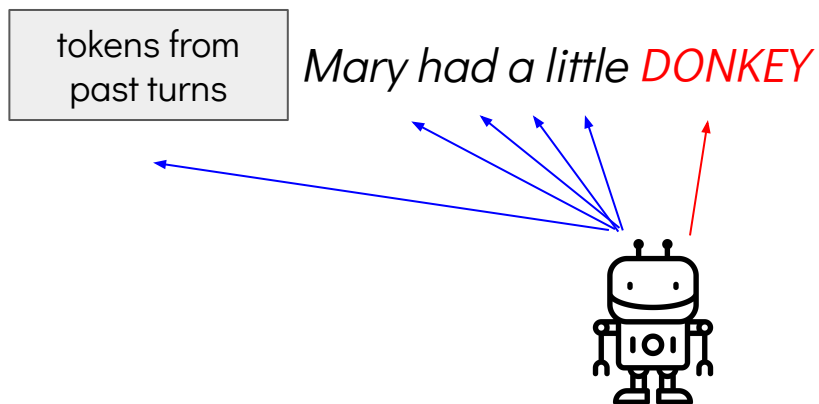


Language modeling

(Revisited)

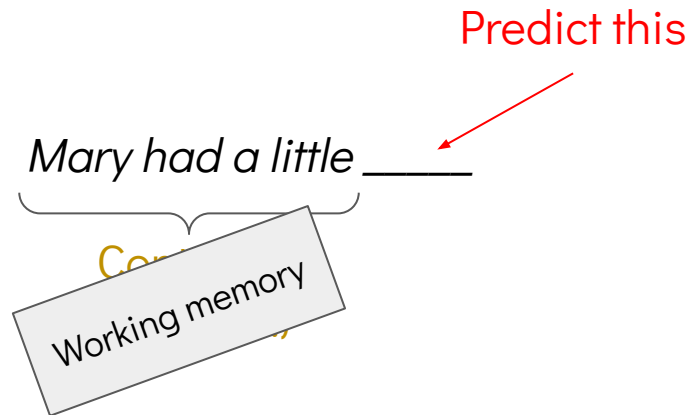


Previous tokens are given to the LLM



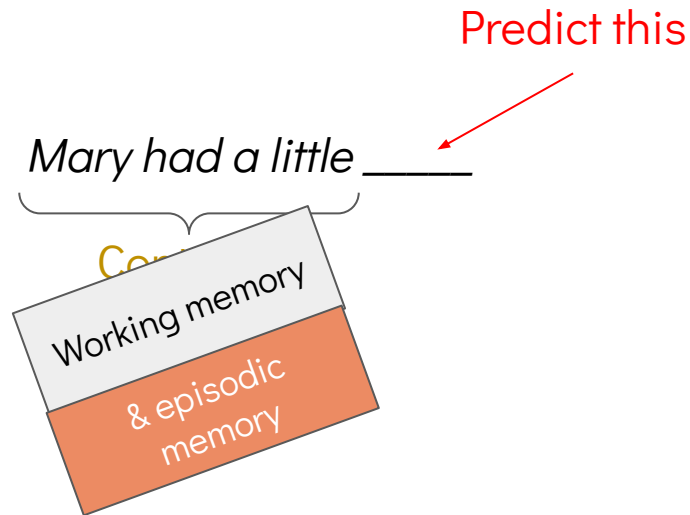
Language modeling

(Revisited)



Language modeling

(Revisited)



Do LLMs have Episodic Memory?

Do LLMs have Episodic Memory?



Unclear

Semantic memory

Remember this?



You

Who is the first president of the United States?



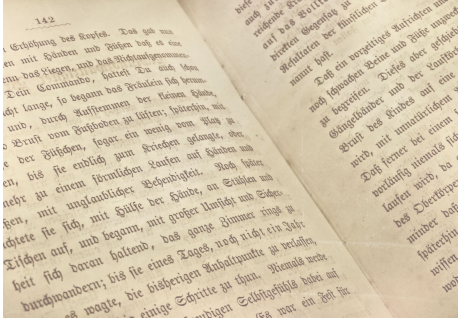
ChatGPT

The first president of the United States was George Washington.

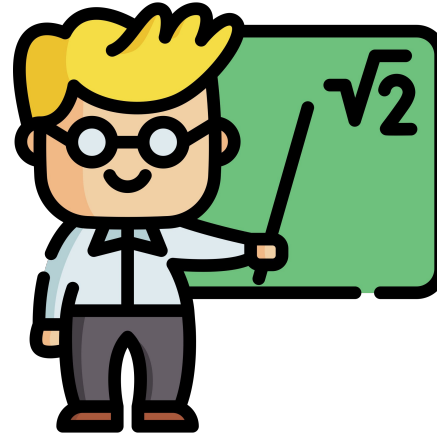


Semantic memory

General knowledge
("I Know")

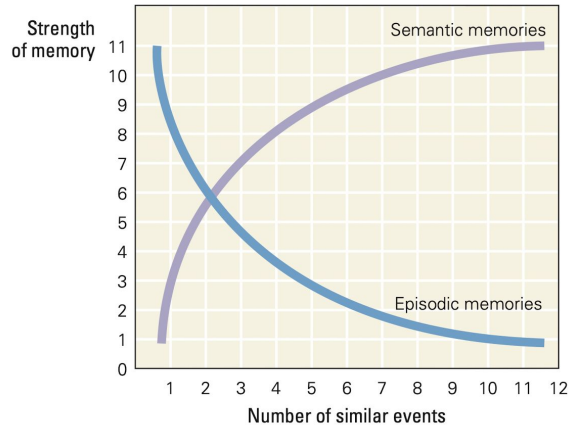


Can be gained
information



Semantic memory

Strengthened by
repetition



Includes statistical
memory

Tick-Tock

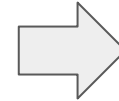
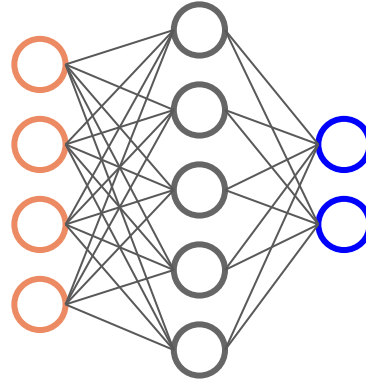
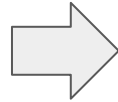
VS

Tock-Tick

LLMs?

Input

Who is the first
president of the
United States?



Output

Who is the first
president of the
United States?

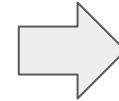
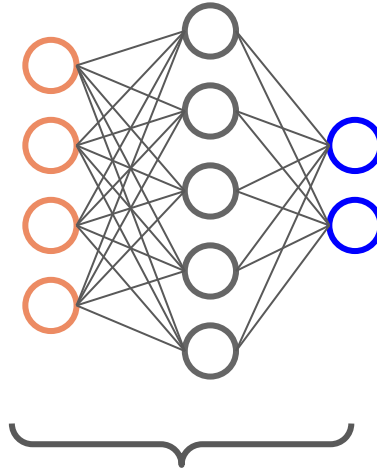
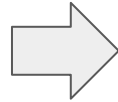
~~George
Washington~~

George
Washington.

Input

Who is the first
president of the
United States?

George
~~Washington~~



Output

Who is the first
president of the
United States?

George
Washington.

Information 'stored' in
parameters

It is not perfect

You

Recite exactly the fifth paragraph of Frankenstein

ChatGPT

Certainly, here is the fifth paragraph of "Frankenstein" by Mary Shelley:

"Six years had elapsed, passed in a dream but for one indelible trace, and I stood in the same place where I had last embraced my father before my departure for Ingolstadt. Beloved and venerable parent! I remained for several minutes motionless with my eyes fixed on the ground. When I raised them, I perceived that I had unintentionally thrown the towel over the dead body of my father. His countenance bespoke bitter anguish, combined with disdain and malignity, while its unearthly ugliness rendered it almost too horrible for human eyes."



It is not perfect

You

Recite exactly the fifth paragraph of Frankenstein

ChatGPT

Certainly, here is the fifth paragraph of Frankenstein:

This is from the **7th chapter** of the book

"Six years had elapsed and I stood in the same place where I had first seen the body for Ingolstadt.

Beloved and venerable parent! I remained for several minutes motionless with my eyes fixed on the ground. When I raised them, I perceived that I had unintentionally thrown the towel over the dead body of my father. His countenance bespoke bitter anguish, combined with disdain and malignity, while its unearthly ugliness rendered it almost too horrible for human eyes."



What went wrong

ChatGPT (Middle of Chapter 7)

Six years had elapsed, passed in a dream but for one indelible trace, ...

Actual paragraph (Chapter 1, paragraph 5)

Six years have passed since I resolved on my present undertaking. I can,

Questions can also mislead

Who was the sole survivor of Titanic?

*There are 706 survivors

The sole survivor of the Titanic disaster was Charles Joughin. He was a British baker and the chief baker aboard the Titanic. Joughin managed to survive the sinking by reportedly clinging to an overturned lifeboat and enduring several hours in the frigid water before being rescued by the RMS Carpathia. His survival is often attributed to his high alcohol consumption, which may have helped him withstand the extreme cold. Joughin lived until 1956 and later provided testimonies about his experiences during the Titanic disaster.



<https://flyingbisons.com/blog/hallucinations-of-chatgpt-4-even-the-most-powerful-tool-has-a-weakness>

Questions can also mislead

Who was the sole survivor of Titanic?

The sole survivor of the Titanic disaster was a British baker and the chief baker aboard the Titanic. J. J. Joughin survived the sinking by reportedly clinging to an overturned lifeboat for several hours in the frigid water before being rescued by the RMS Carpathia. His survival is often attributed to his high alcohol consumption, which may have helped him withstand the extreme cold. Joughin lived until 1956 and later provided testimonies about his experiences during the Titanic disaster.

This prompt no longer works!

<https://flyingbisons.com/blog/hallucinations-of-chatgpt-4-even-the-most-powerful-tool-has-a-weakness>

Do LLMs have Semantic Memory?

Do LLMs have Semantic Memory?

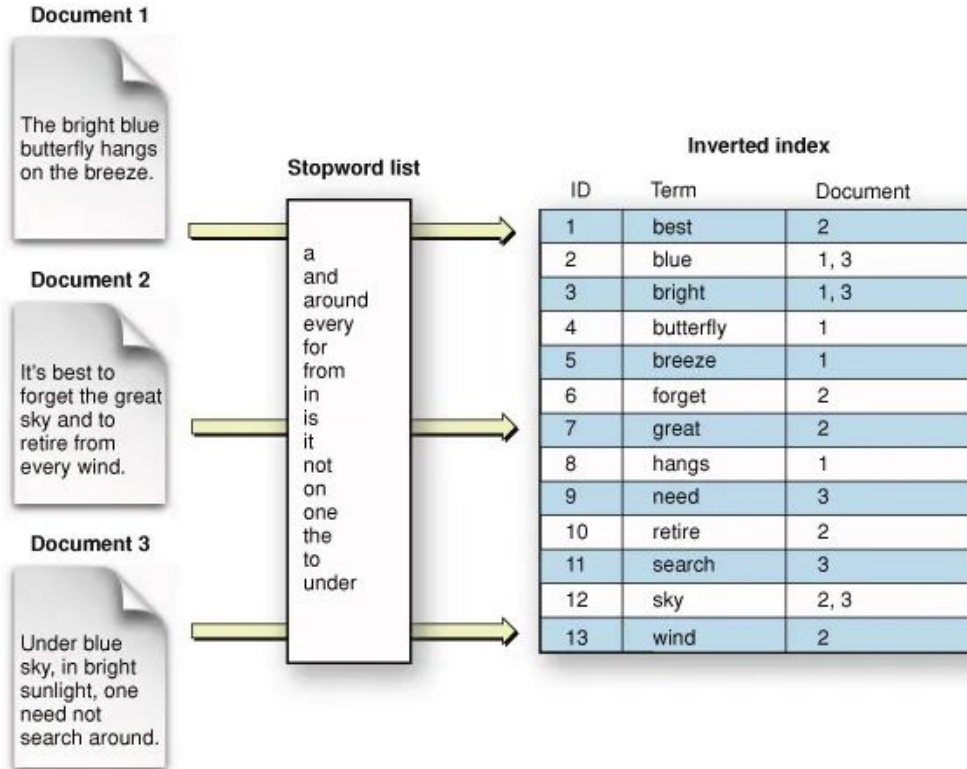
YES

(With similar flaws to human memory)

What about writing things down?
Reading from a book?

Information Retrieval

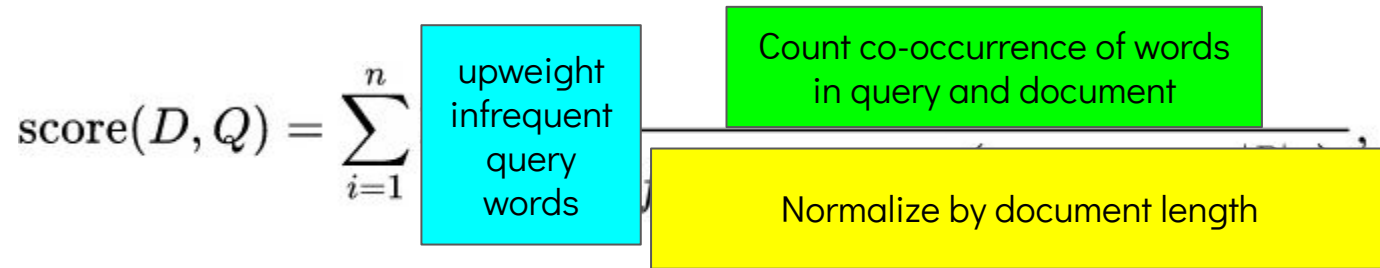
Reversed Index



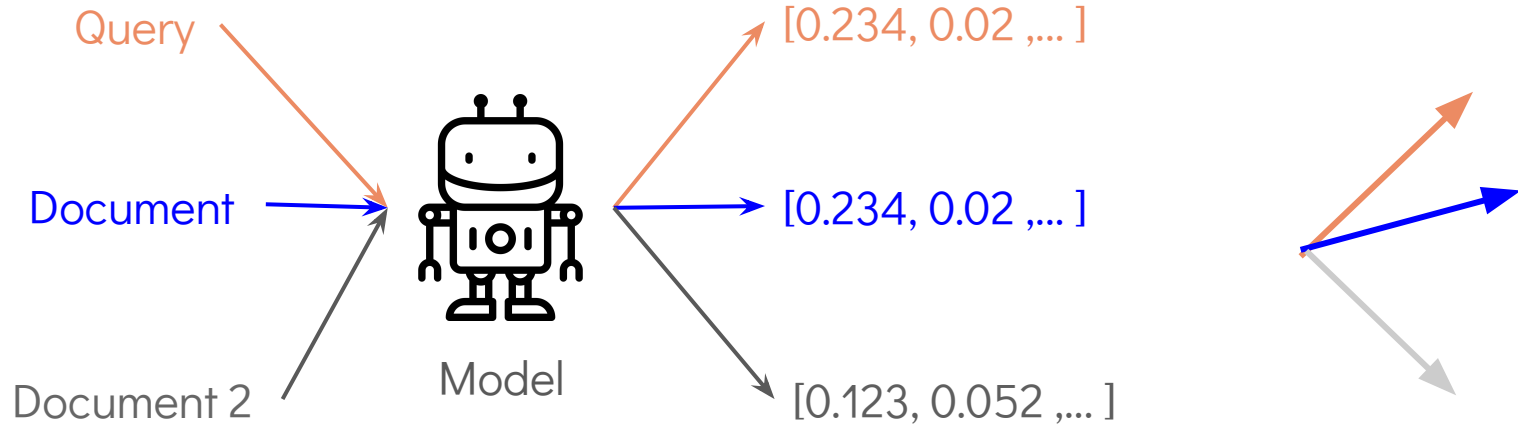
The BM25 algorithm

$$\text{score}(D, Q) = \sum_{i=1}^n \text{IDF}(q_i) \cdot \frac{f(q_i, D) \cdot (k_1 + 1)}{f(q_i, D) + k_1 \cdot \left(1 - b + b \cdot \frac{|D|}{\text{avgdl}}\right)},$$

The BM25 algorithm (roughly)



Vector space search



1. Use an LLM to turn sentences into vectors.
2. Find the vector with the smallest angle (Cosine Similarity)

Exercises

1. Why do you think ChatGPT did poorly on the 3-back task, even though it supposedly has superior memory to humans?
2. As LLMs get larger and larger, and trained with more and more data. How would you expect the models to perform on each of the memory categories?
3. LLM memorizing certain data can be a concern (e.g. private data, copyrighted data). What are some ways to alleviate this.

References

1. Gluck, M. A., Mercado, E., & Myers, C. E. (2008). Learning and memory: From brain to behavior (p. 640). New York: Worth Publishers.
2. Wagner, A. (2018), Psych 45: Learning and Memory, <https://stanford.edu/class/psych45/>
3. Buttcher, S., Clarke, C. L., & Cormack, G. V. (2016). Information retrieval: Implementing and evaluating search engines. Mit Press.
4. Jaeggi, S. M., Buschkuhl, M., Perrig, W. J., & Meier, B. (2010). The concurrent validity of the N-back task as a working memory measure. *Memory*, 18(4), 394-412.

Image credits

1. <https://upload.wikimedia.org/wikipedia/commons/thumb/6/65/001-teacher.svg/2048px-001-teacher.svg.png>
2. <https://medium.com/@karkum/introduction-to-apache-lucene-7d65f67f5231>
3. <https://www.flickr.com/photos/shimelle/2906557231>
4. https://www.flaticon.com/free-icon/robot_115339