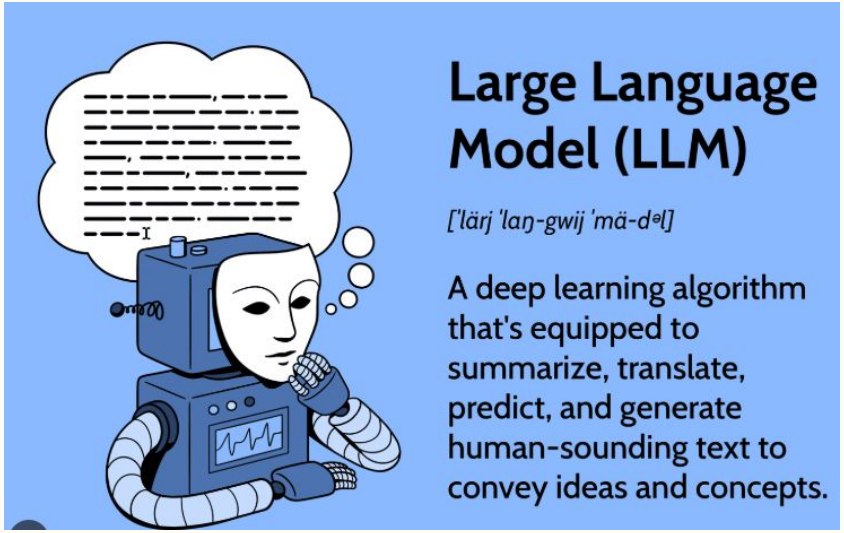
SCHRODINGER'S MEMORY: LLM

Chint Patel CMPE 255

Introduction



Ever since Open AI released the first LLM to the public, GPT-3, the world has been completely changed. You can even say it was a timeline shift for the course of civilization. We are constantly using new models everyday with many competitors arising from the ground to gain market share in this fierce competition. Models like GPT-4o and Llama are immensely useful with their ability to not only converse but also recall/remember details from previous inputs. This feature of memory is essential for good models and without which they lose their usefulness. This naturally raises the question, do these models possess memory? If so, is it similar to humans and how do they store this memory?

What is Memory?

First we need to start off with defining memory and for this an old definition that doesn't quite work in this case:

"Memory is the faculty of the mind by which data or information is encoded, stored, and retrieved when needed."

What is Memory

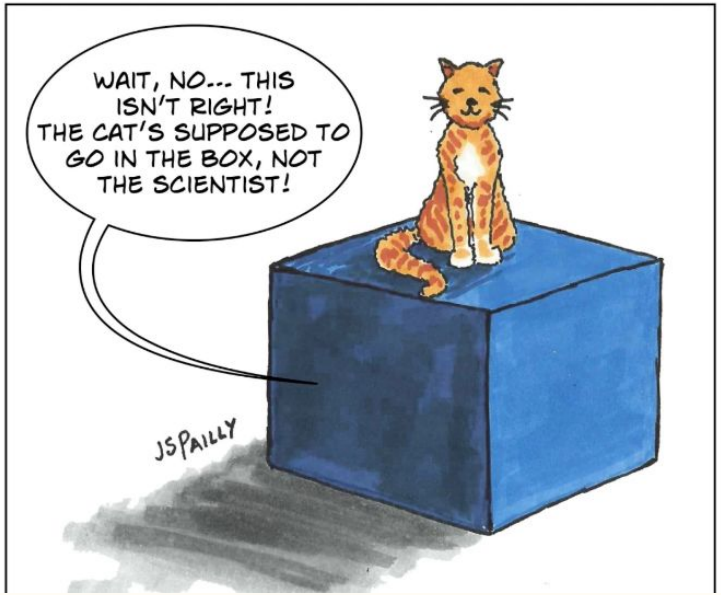
The issue with this is that where is the memory actually stored. In human brains we only really know it is some neural activity and with LLMs we know it is through weights. The storage of memory is not rigorous like a database and so the authors proposed a better definition:

* Input: To trigger a memory, the input must be the same or similar to information that the brain (or LLM) has previously encountered.

* Output: The result is based on the input, which could be correct, incorrect, or forgotten. If the result is correct, it means it aligns with information previously acquired.

What is Memory

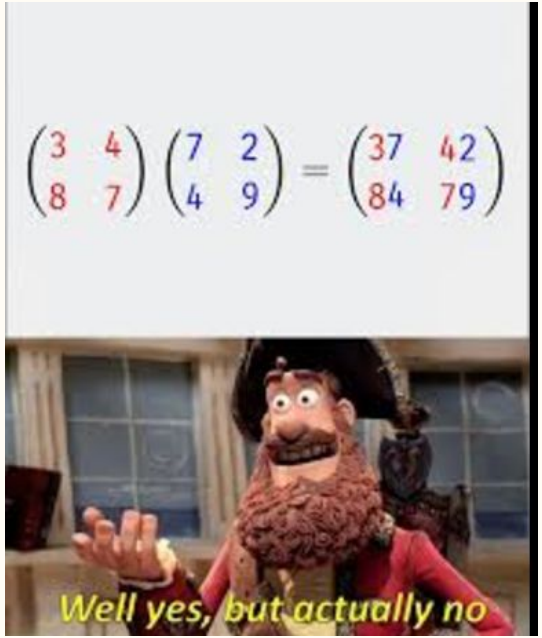
So Schrodinger’s memory is a concept where the memory can only be accessed by providing a corresponding input similar to the famous Schrodinger's cat experiment. Without a specific input you cannot gauge whether the memory exists.



How Memory is Achieve Mathematically

Essentially, memory in LLMs does not have storage like databases but rather a dynamic computation of outputs based on inputs. This phenomenon is underpinned by Universal Approximation Theory (UAT).

What is UAT?
UAT states that a neural network with sufficient complexity can approximate any continuous function. This implies that neural networks, including those in LLMs, can learn to map inputs to desired outputs with remarkable precision.



Transformers and Dynamic Memory

Transformer-based LLMs extend the principles of UAT. Unlike traditional networks, which rely on fixed parameters post-training, Transformers:

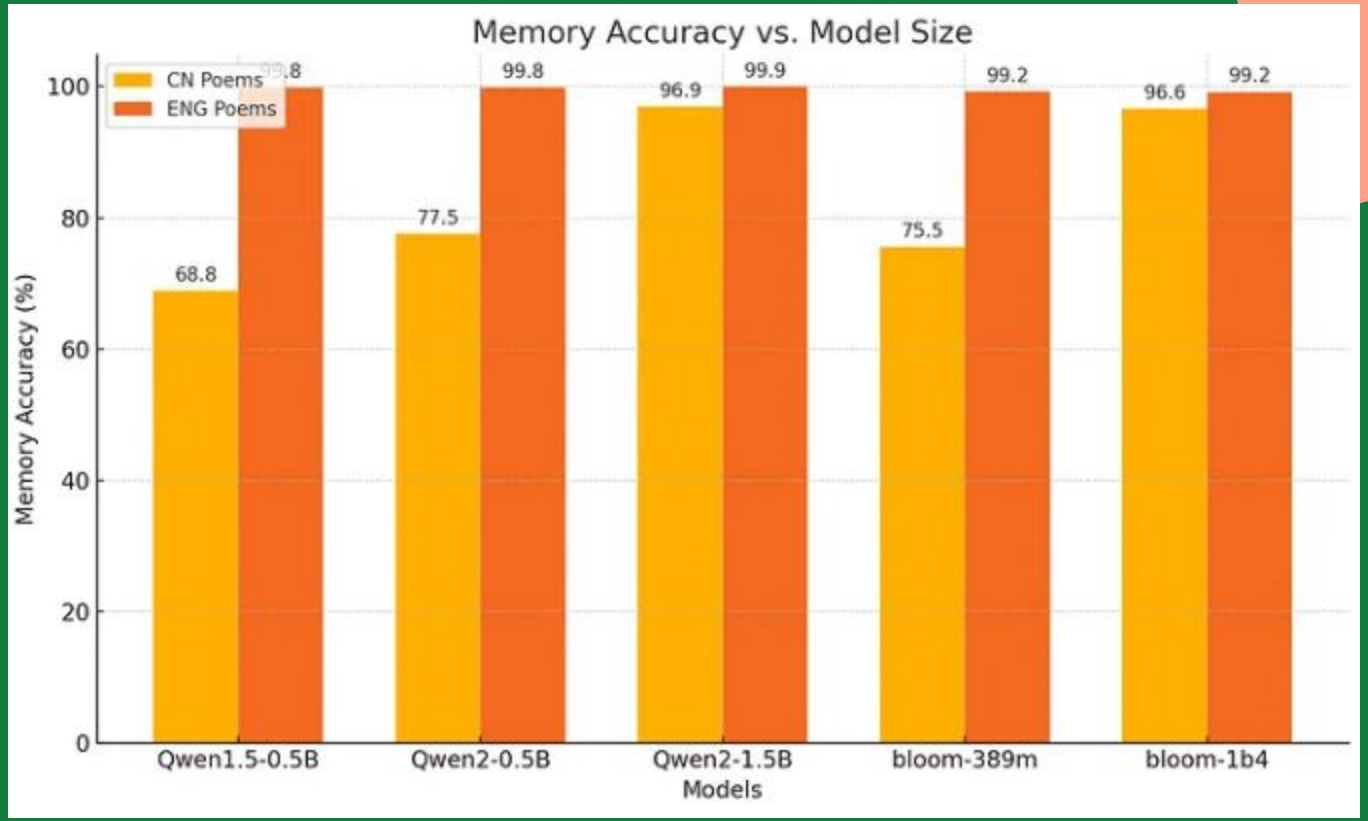
Adaptively fit outputs based on inputs: Every input triggers dynamic adjustments in intermediate layers, enabling nuanced responses.

Leverage multi-head attention: This mechanism extracts relationships between different parts of the input, reinforcing relevant patterns dynamically.

Experimentation with Memory

To analyze the memory abilities of LLMs, the researchers used a dataset of Chinese and English poems and tested recall and recreation of new poems based on existing inputs. They used different models like Qwen and bloom. Key findings:

- Larger models with richer training data excel at memory tasks.
- Memory accuracy decreases with longer outputs.
- Even with errors, LLM outputs remain coherent and aligned with the input's context.

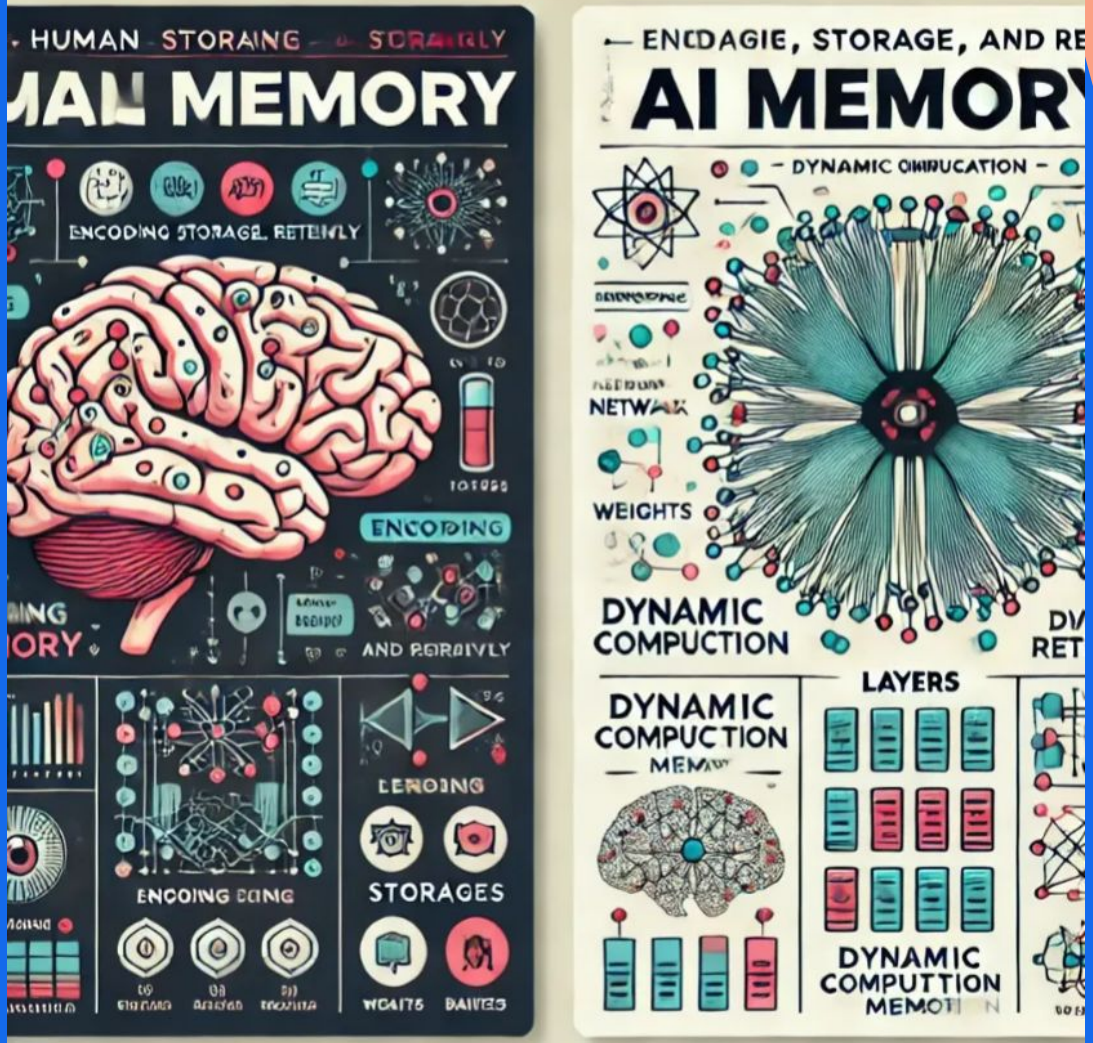

Memory Accuracy vs. Model Size
CN Poems
ENG Poems
Memory Accuracy (%)
0
20
40
60
80
100
68.8
99.8
77.5
99.8
96.9
99.9
75.5
99.2
96.6
99.2
Qwen1.5-0.5B
Qwen2-0.5B
Qwen2-1.5B
bloom-389m
bloom-1b4
Models

The Brain vs AI

Surprisingly, LLMs are quite similar to human brains in their dynamic response to stimuli. However, the underlying methods differ substantially:

- Humans: Use neurons and synapses for memory
- LLMs: Store knowledge in trained parameters

Comparing the human hippocampus to the iterative training of LLMs is quite profound as they server to update the memory in both cases.


HUMAN STORAING
MEMORY
ENCODING STORAGE RETRIVLY
ENCODING
AND RETRIVLY
ENCODING
STORAGES
AI MEMORY
DYNAMIC
COMPUCTION
LAYERS
DYNAMIC
COMPUCTION
DYNAMIC
COMPUTTION

Why Does this matter

Understanding memory in LLMs could lead to insights into making LLMs better and faster. Imagine Scaling LLMs with faster and better memory structures. This could make everything better. We could even understand the memory structure of our own brains leading to better memory training for ourselves. Maybe in a distant future we could just upload memories into our brain and not need to learn in the same way we do today. I do not know but I know we must push forward and accelerate since we owe it to the future generations that have yet to come.