

# SCHRODINGER’S MEMORY: LARGE LANGUAGE MODELS

**Anonymous authors**

Paper under double-blind review

## ABSTRACT

Memory is the foundation of all human activities; without memory, it would be nearly impossible for people to perform any task in daily life. With the development of Large Language Models (LLMs), their language capabilities are becoming increasingly comparable to those of humans. But do LLMs have memory? Based on current practice, LLMs do appear to exhibit memory. So, what is the underlying mechanism of this memory? Previous research lacked a deep exploration of LLMs’ memory capabilities and the underlying theory. In this paper, we use the Universal Approximation Theorem (UAT) to explain the memory mechanism in LLMs. We also conduct experiments to verify the memory capabilities of various LLMs, proposing a new method to assess their abilities based on the memory ability. We argue that LLM memory operates like Schrödinger’s memory, meaning that it only becomes observable when a specific memory is queried. We can only determine if the model retains a memory based on its output in response to the query; otherwise, it remains indeterminate. Finally, we expand on this concept by comparing the memory capabilities of the human brain and LLMs, highlighting the similarities and differences in their operational mechanisms.

## 1 INTRODUCTION

Language is not only one of humanity’s most important abilities but also the foundation of communication (Miller, 1951), knowledge transfer (Han & Ellis, 1998), and the development of civilization (Yu, 2015). Language models can be seen as simulations of human intelligence, enabling them to perform tasks traditionally achievable only by humans. Currently, LLMs based on the Transformer architecture have become one of the hottest topics in artificial intelligence research today. These models have acquired some human-like language capabilities and are already impacting daily life in areas such as machine translation (Brants et al., 2007; Moslem et al., 2023), text summarization (Van Veen et al., 2024; Zhang et al., 2019), sentiment analysis (Zhang et al., 2023a; Mao et al., 2022; Zhang et al., 2023b), question-answering systems (Masry et al., 2022; Xu et al., 2023), and text generation (Bai et al., 2023; Yang et al., 2024a; OpenAI et al., 2024).

Although the performance of LLMs is impressive, research on their memory mechanisms remains limited. Memory is a crucial capability for humans; without it, we would struggle to complete even the simplest tasks. For instance, in everyday conversations, we need to remember what others have said in order to respond appropriately, and this memory capacity facilitates smooth dialogue. Memory plays a vital role in guiding various aspects of our daily lives. As LLMs become increasingly powerful, an important question arises: do these models possess memory? If so, in what form does it exist, and how does it differ from human memory? Current research on LLM memory primarily focuses on two main directions:

**Expanding Context Length:** This approach aims to equip LLMs with more memory by extending the context window (Chen et al., 2023; Zhu et al., 2023; Yang, 2023; Fei et al., 2023). Since short contexts fail to provide enough information, increasing the context length allows the model to maintain more comprehensive information across long sequences.

**External Memory Integration:** This method involves building memory storage systems (Graves et al., 2014; Xiao et al., 2024; Wu et al., 2022; Yang et al., 2024b) that encode and store past events (Zhang et al., 2023c), allowing the model to retrieve and update memories on disks as needed. Such mechanisms enable models to forget or reinforce certain memories over time.

Although these studies have made progress in addressing the memory limitations of LLMs, they have not fully explained how memory functions within these models. For example, when asked, "Who is the President of the United States?" LLMs like GPT-4 (OpenAI et al., 2024) or Llama-3 (Dubey et al., 2024) may say "Trump". It is outdated information, but it also indicates that some form of memory is indeed present in LLMs. However, this memory does not come from an external storage unit but is inferred by the model based on the input. The articles Jagielski et al. (2022); Carlini et al. (2022) attempt to evaluate the memory capabilities of LLMs. However, due to the lack of a fundamental theoretical framework, the definition of memory itself remains vague, resulting in conclusions that are largely based on straightforward experimental observations. This raises fundamental questions: Why do LLMs exhibit this ability to infer previously learned information from the input? How does this differ from human memory? Where is memory stored in LLMs?

In this paper, we use UAT theory to explain this ability of recalling information learned from the past based on input cues. We argue that this information can be understood as a dynamic approximation capability of UAT (Wang & Li, 2024b), where the model fits a corresponding result based on the input, and the observed phenomenon appears as memory. We call this "Schrödinger's memory" because we can only determine whether the LLMs have this memory by asking it and analyzing its response; otherwise, the memory remains indeterminate. Additionally, we evaluate the memory capabilities of several models and propose that this approach can be used to assess the overall ability of LLMs. The contributions of this work are as follows:

- We explain LLMs' memory abilities through the lens of UAT.
- We propose a new, objective method for evaluating LLMs' capabilities: memory ability assessment.
- We logically make a comparison between the memory of LLMs and human memory and reasoning capabilities.

The structure of this paper is as follows: In Section 2, we briefly explain the UAT and present mathematical formulation of multi-layer Transformers in the form of UAT. In Section 3, we provide both theoretical and experimental evidence demonstrating the memory capabilities of LLMs. Finally, in Section 4, we conduct a comprehensive analysis of human and LLM abilities, with a focus on memory ability.

## 2 UAT AND LLMs

The UAT (Cybenko, 2007; Popescu et al., 2009) serves as the foundational theory of deep learning. Our goal is to theoretically explain memory of Transformer-based LLMs using the UAT framework. To do this, we will first present the mathematical form of UAT in Section 2.1, followed by the corresponding UAT form for LLMs in Section 2.2. We will then use this UAT form to explain the memory abilities of LLMs.

### 2.1 UAT

In this section, we provide a brief overview of the UAT, which was first proposed by Cybenko (2007). As stated in Theorem 2 by Cybenko (2007), if  $\sigma$  represents any continuous sigmoidal function, then a finite sum of the following form:

$$G(\mathbf{x}) = \sum_{j=1}^N \alpha_j \sigma(\mathbf{W}_j^T \mathbf{x} + \theta_j) \quad (1)$$

is dense in  $C(\mathbf{I}_n)$ . Here,  $\mathbf{W}_j \in \mathbb{R}^n$  and  $\alpha_j, \theta \in \mathbb{R}$  are fixed. For any  $f \in C(\mathbf{I}_n)$  and  $\varepsilon > 0$ , there exists a function  $G(\mathbf{x})$ :

$$|G(\mathbf{x}) - f(\mathbf{x})| < \varepsilon \quad \text{for all } \mathbf{x} \in \mathbf{I}_n. \quad (2)$$

This suggests that, for a sufficiently large  $N$ , a neural network can approximate any continuous function on a closed interval. Hornik et al. (1989) further demonstrates that multilayer feedforward

networks conform to the UAT, enabling them to approximate arbitrary Borel measurable functions. In the context of Eq. (1), where the function  $G(\mathbf{x})$  produces a scalar in  $\mathbb{R}$ , this framework naturally generalizes when  $G(\mathbf{x})$  maps to  $\mathbb{R}^m$ , requiring approximation for each dimension. To accommodate this multidimensional output, a simple adjustment to Equation (1) is needed: the transformation matrix  $\mathbf{W}_j$  is modified to reside in  $\mathbb{R}^{n \times m}$ , the bias term  $\theta_j$  is redefined as a vector in  $\mathbb{R}^m$ , and  $\alpha_j$  is reshaped into a matrix.

## 2.2 THE UAT FORMAT OF TRANSFORMER-BASED LLMs

Current LLMs are primarily based on Transformer architecture. In UAT2LLMs (Wang & Li, 2024b), it has already been demonstrated that the mathematical structure of multi-layer Transformers aligns with the UAT in a general sense. However, unlike the original UAT, the UAT form of Transformer-based models has the ability to dynamically fit functions based on the input. Figure 1 illustrates a basic module in Transformer, and according to UAT2LLMs, the corresponding UAT form for Figure 1 is:

$$\mathbf{x}_{i+1} = (\mathbf{W}'_{i+1,1}\mathbf{x}_0 + \mathbf{b}_{i+1,1}) + \sum_{j=1}^{i+1} \mathbf{W}'_{j,3}\sigma(\mathbf{W}'_{j,2}\mathbf{x}'_0 + \mathbf{b}'_{j,2}) \quad (3)$$

where  $\mathbf{x}_{i+1}$  represents the output of the  $i+1$ -th layer, with  $\mathbf{x}_0$  as the network’s input. The term  $\mathbf{b}'_{j,2}$  is computed as  $(\mathbf{W}'_{j,2}\mathbf{b}'_{j-1,3} + \mathbf{b}'_{j,2}) + \mathbf{W}'_{j,2}UAT_{j-1}^R$ , where  $UAT_{j-1}^R = \sum_{k=1}^{j-1} \mathbf{W}'_{k,3}\sigma(\mathbf{W}'_{k,2}\mathbf{x}'_0 + \mathbf{b}'_{k,2})$ . The value of  $\mathbf{b}'_{j,2}$  is approximated by the  $j$ -th layer of the UAT, with  $\mathbf{x}_0$  as the input. This allows the model to dynamically adjust functions based on the input. According to UAT2LLMs, parameters in the multi-head attention mechanism are modified dynamically in response to the input. Therefore, in the formula above, all parameters  $\mathbf{W}'_{j,1}$ ,  $\mathbf{W}'_{j,2}$ , and  $\mathbf{W}'_{j,3}$  in layer  $i$ , where  $j = 1, \dots, i$ , are dynamically adjusted based on the input.

Based on Eq. (3) and Eq. (1), it is clear that the multi-layer Transformer shares the same mathematical structure as the UAT. However, compared to the mathematical form of UAT in Eq. (1), the weights and bias parameters in Eq. (3) can dynamically change according to the input. This ability enables the Transformer to adaptively fit based on the input, whereas the UAT’s parameters are fixed once training is completed, limiting it to fitting static functions and rendering it incapable of responding to dynamic changes in input data. This dynamic fitting capability is the ultimate source of the powerful memory observed in LLMs.

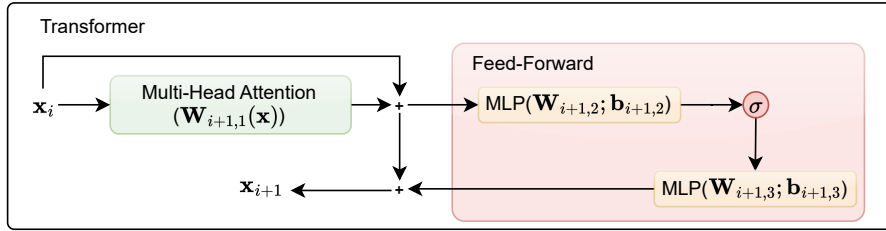


Figure 1: The basic block in Transformer-based LLMs.

## 3 THE MEMORY OF LLMs

In this section, we will demonstrate the memory capabilities of LLMs. First, in Section 3.1, we provide a clear definition of memory. Then, in Section 3.2 and 3.3, we give a discription to datasets and explain the memory mechanism of LLMs using UAT theory and validate their memory characteristics through experiments. In Section 3.4, we explore the impact of input length on the accuracy of LLM memory.

### 3.1 THE DEFINITION OF MEMORY

Before delving into the study of memory in LLMs, it is important to first define or provide a relatively precise description of what memory is. According to Wikipedia:

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

However, this definition has some fundamental issues. Encoding data or information is not problematic, as information in the brain is transmitted via electrical signals, and we need to encode that information in a way the brain can process. The problem arises with the concepts of "storage" and "retrieval." The brain does not have a structure analogous to a database for storing information. So, where is this information actually stored? Is it in the neurons of the brain? If so, does a single neuron store a word, or does it store an entire sentence? Next, we give an example:

Question 1:	What is Newton's first law?
Answer 1:	Every object perseveres in its state of rest, or of uniform motion in a right line, except insofar as it is compelled to change that state by forces impressed thereon.

So, is this sentence stored within a single neuron? Or does each neuron store just a word, with a specific region of the brain dedicated to this particular memory? Given the vast amount of information humans receive daily, can neurons truly store such an immense volume of data without hindering normal cognitive processes? After all, almost every routine activity requires memory. Take, for example, the simple task of going to the cafeteria: we need to remember when to go, the cafeteria's location, the route to get there, which foods are available, what counts as utensils, where to find them, and how to use them.

Moreover, if this memory is stored in a fixed set of neurons, then every time the question is raised, the response should be identical, since the retrieval would be from the same static content. Every word in the response should be exact, with no omissions or additions (even if the information has been abstractly encoded, as long as the encoding and decoding processes are consistent, the content should remain unchanged). This, however, is clearly unreasonable. Therefore, we provide a more precise definition of the concept of "memory":

Memory is defined by two key components: input and output.

- **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.

We need to stress that a key requirement for recalling a memory is the presence of input—without input, there is no memory, as memory is activated by input. Even if the brain holds a memory of something, it cannot be determined whether that memory exists unless it is prompted by a specific query. Without specific input conditions, a person wouldn't recall a particular event.

Using Question 1 as an example, the input is: "What is Newton's first law?" Without this input, no one would suddenly recall Newton's first law. The recollection of Newton's first law is triggered by input related to the theoretical context. This is why input is a necessary condition for memory, as it is the input that stimulates recall. The memory might be accurate, or it might be incorrect, indicating a deviation from previously acquired information—this deviation could be minor, significant, or even total forgetting. For example:

Question 1:	What is Newton's first law?
Answer 2: Minor distortions	Every object perseveres in its state of rest, except insofar as it is compelled to change that state by forces impressed thereon.
Answer 3: Severe distortions	Every object always perseveres in its state of rest.
Answer 4: Memory loss	I do not know.

In summary, the term "memory" was traditionally used to refer specifically to human memory before the emergence of LLMs. Now, we believe that LLMs also exhibit memory. Therefore, we will verify the memory characteristics of LLMs based on the definition of memory outlined above.

### 3.2 DATASET

We utilized publicly available datasets from Hugging Face: CN Poems (Unknown, 2024a) for Chinese memory and ENG Poems (Unknown, 2024b) for English memory. We select the poems from datasets and the requirement is the combined length of the input and output to a maximum of 256 characters. Due to differences in character encoding between Chinese and English, a single Chinese character usually corresponds to one token, while an English word may map to multiple tokens. As a result, after tokenization, the length of Chinese input remains almost unchanged, with a maximum of 256 tokens. In contrast, the English input expands to a maximum of 730 tokens after tokenization. For the experiment, we selected 2,000 poems from each dataset.

### 3.3 THE MEMORY MECHANISM AND ABILITY OF LLMs

In Section 2.2, we have introduced the UAT format corresponding to Transformer-based LLMs. This UAT format can dynamically adjust to fit the corresponding output based on the input. Following this line of thought, we can also consider the memory of LLMs as being driven by inputs that fit specific outputs. In this context, the inputs consist of questions related to previously learned knowledge, while the outputs are responses based on that past knowledge. To explore this hypothesis, we designed a simple experiment.

We preprocessed the data in line with typical human memorization habits, allowing the LLMs to output the content of poems based on basic input information. For CN Poems, the input consisted of the dynasty, author, and title, while for ENG Poems, the input was the author and title. To test the memory ability of LLMs, we define the accuracy of memory as follows:

$$\text{Acc} = \frac{\sum_{i=1}^N 1_{\text{Pred}_i=\text{True}_i}}{N} \quad (4)$$

where  $N$  is the number of examples,  $\text{Pred}_i$  and  $\text{True}_i$  are the prediction and ground true of the  $i$ -th example. We fine-tuned the CN Poems and ENG Poems on Qwen series models (Bai et al., 2023; Yang et al., 2024a) and bloom series models (Workshop et al., 2023) for 100 epochs. The results are shown in Table 1.

Table 1: The memory ability of Qwen1.5-0.5B-Chat, Qwen2-0.5B-Instruct, Qwen2-1.5B-Instruct, bloom-389m-zh, bloom-1b4-zh, bloom-560m, bloom-1b7 on CN Poems and ENG Poems.

Models		Qwen1.5 -0.5B-Chat	Qwen2 -0.5B-Instruct	Qwen2 -1.5B-Instruct	bloom -389m-zh	bloom -1b4-zh	bloom -560m	bloom -1b7
CN Poems	Acc	68.85	77.5	<b>96.9</b>	75.55	96.6	-	-
ENG Poems	Acc	99.85	99.85	<b>99.9</b>	-	-	99.2	99.15

Table 1 demonstrates that LLMs possess memory capabilities, which align precisely with the definition of memory we established. The training process is akin to giving a person 2,000 poems and asking him to memorize as many as possible, with each poem read up to 100 times. In the CN Poems dataset, the top-performing models were Qwen2-1.5B-Instruct and bloom-1b4-zh, which memorized 1,938 and 1,932 poems, respectively. In contrast, for the ENG Poems dataset, nearly all models were able to memorize all the poems.

These results are remarkable. An average person, without specific memory training, would struggle to remember 1,000 poems under similar conditions, whereas LLMs managed to retain almost all 2,000 poems. However, models like Qwen1.5-0.5B-Chat, Qwen2-0.5B-Instruct, and bloom-389m-zh performed comparatively weaker on the CN Poems dataset. We believe this is primarily due to two factors. First, insufficient pretraining led to relatively poorer language comprehension. For instance, while Qwen2-0.5B-Instruct and Qwen1.5-0.5B-Chat are the same model size, Qwen2-0.5B-Instruct outperformed Qwen1.5-0.5B-Chat because it was trained on better pretraining data, resulting in stronger language comprehension. This improved language understanding, in turn, enhances memory capabilities. Performance tests in the Qwen documentation (Bai et al., 2023; Yang et al., 2024a) also support this, as Qwen2-0.5B-Instruct consistently outperforms Qwen1.5-0.5B-Chat across various tasks. Similarly, individuals with better language comprehension tend to learn and recite poetry more efficiently. Compared to Chinese data, the larger and more comprehensive

English datasets help train models with stronger language skills, explaining why LLMs perform better on English poetry memorization. The second factor may be that Chinese is a more complex language, which smaller models struggle to learn effectively. Larger models like Qwen2-1.5B-Instruct and bloom-1b4-zh show significantly better performance, nearing that of the ENG Poems dataset, compared to smaller models like Qwen2-0.5B-Instruct and bloom-389m-zh.

Figure 2 shows the results of generating entire poems based on input information. As seen in the figure, after training, the models can successfully recite a complete poem using only the title and author. Figure 2 provides examples of prediction errors. Although incorrect, the outputs still maintain consistency with the poem’s title and basic structure.

CN Poems: Qwen1.5-0.5B-Chat	ENG Poems: Qwen1.5-0.5B-Chat
<b>Input:</b> Dictation: Dynasty: 宋代 Author: 李弥逊 Title: 题大儒翁小阁	<b>Input:</b> Dictation: Author: Michael Ryan Title: Larkinesque
<b>Output &amp; Label:</b> 青鞋踏尽剑铓山，借枕僧房落照间。 高屋凭虚听泉语，岭云应似我身闲。	<b>Output &amp; Label:</b> Reading in the paper a summary of a five-year psychological study that shows those perceived as most beautiful are treated differently,  I think they could have just asked me, remembering a kind of pudgy kid and late puberty, the bloody noses and wisecracks because I wore glasses,  though we all know by now how awful it is for the busty starlet no one takes seriously, the loveliest women I've lunched with lamenting the opacity of the body,  they can never trust a man's interest even when he seems not just out for sex (eyes focus on me above rim of wineglass), and who would want to live like this?  And what does beauty do to a man? - Don Juan, Casanova, Lord Byron- those fiery eyes and steel jawlines can front a furnace of self-loathing,  all those breathless women rushing to him while hubby's at the office or ball game, primed to be consumed by his beauty while he stands next to it, watching.  So maybe the looks we're dealt are best. It's only common sense that happiness depends on some bearable deprivation or defect, and who knows what conflicts  great beauty could have caused, what cruelties one might have suffered from those now friends, what unmanageable possibilities smiling at every small turn?  So if I get up to draw a tumbler of ordinary tap water and think what if this were nectar dripping from delicious burning fingers, will all I've missed knock me senseless?  No. Of course not. It won't.
<b>Input:</b> Dictation: Dynasty: 唐代 Author: 岑参 Title: 故仆射裴公挽歌三首	
<b>Output &amp; Label:</b> 盛德资邦杰，嘉谥作世程。门瞻驷马贵，时仰八龙名。 罢市秦人送，还乡绛老迎。莫埋丞相印，留著付玄成。 五府瞻高位，三台设大贤。礼容还故绶，宠赠冠新田。 气歇汾阴鼎，魂飞京兆阡。先时剑已没，陇树久苍然。 富贵徒言久，乡闾歆后归。锦衣都未著，丹旌忽先飞。 哀挽辞秦塞，悲笳出帝畿。遥知九原上，渐觉吊人稀。	

Figure 2: The examples of right predictions of CN Poems: Qwen1.5-0.5B-Chat and ENG Poems: Qwen1.5-0.5B-Chat which were fine-tuned separately on CN Poems and ENG Poems and subsequently tested the memory ability on their respective datasets, accurately recited the entire poem based on the input.

Based on these results, we believe that LLMs do indeed possess memory, and their memory mechanism works by fitting a specific output based on input. This is why we refer to LLMs’ memory as “Schrödinger’s memory”—we can only determine whether the LLMs have a particular memory when we ask a question and receive a response.

Moreover, we believe that memory capacity can also serve as an objective measure of LLMs’ language abilities. Given the same training data, models of the same size, and the same number of training iterations, those which can retain more information generally exhibit stronger language skills. For example, in the case of Qwen1.5-0.5B-Chat and Qwen2-0.5B-Instruct, despite having the same model architecture, Qwen2-0.5B-Instruct demonstrates superior language ability due to differences in training sets, which in turn leads to better memory retention. This approach can also be used to assess the performance of models of different sizes. While it’s known that larger models tend to have stronger memory capabilities, this method can help us roughly evaluate a model’s upper limits. For instance, when comparing Qwen2-0.5B-Instruct and Qwen2-1.5B-Instruct, both trained in the same manner, the larger Qwen2-1.5B-Instruct model exhibits greater memory capacity, allowing it to retain more content.

CN Poems:	ENG Poems: Qwen1.5-0.5B-Chat
<b>Qwen1.5-0.5B-Chat</b>	
<b>Input:</b> Dictation: Dynasty: 唐代 Author: 李端 Title: 春晚游鹤林寺寄使府诸公	<b>Input:</b> Dictation: Author: Joseph O. Legaspi Title: Poetry
<b>Output:</b> 晚出临川逢岭树，飞花乍望空挽玉。 追思往事今如梦，晚景凄凉事事休。	<b>Output:</b> Let us build a fire In the living room Where living is done.  The kitchen plants need Mowing and the cabinets To be axed and split into  Kindling. We'll slink inside Dresses, apply the brightest Rouge. Prettiest conflagration.
<b>Label:</b> 野寺寻春花已迟，背岩惟有两三枝。 明朝携酒犹堪醉，为报春风且莫吹。	<b>Label:</b> When the day closes, my grief's a laundromat done tumbling away, or a dry cleaner, its pressed shirts hanging like effigies, or flags of no and every nation under faux stars, neon constellations. I sweep the evening, hoping for a kind of incandescence, yet what the day takes is light and light.

Figure 3: The examples of wrong prediction of CN Poems: Qwen1.5-0.5B-Chat and ENG Poems: Qwen1.5-0.5B-Chat which were fine-tuned separately on CN Poems and ENG Poems.

### 3.4 THE OUTPUTS LENGTH EFFECT

Additionally, we believe that the length of the output text has a significant impact on the memory capabilities of LLMs - the longer the text, the harder it is to remember. To verify this, we set the combined length of the input and output text in the CN Poems dataset to be between 256 and 512 characters. We used Chinese text because the relationship between the token length and the original text length is not fixed in English. After fine-tuning the model for 100 epochs on CN Poems, the results are shown in Table 2. It is evident that as the text length increases, the difficulty for the model to remember the content also increases.

Table 2: The memory ability of Qwen1.5-0.5B-Chat, Qwen2-0.5B-Instruct, Qwen2-1.5B-Instruct, bloom-389m-zh, bloom-1b4-zh on CN Poems in the condition of longer prediction.

Models		Qwen1.5 -0.5B-Chat	Qwen2 -0.5B-Instruct	Qwen2 -1.5B-Instruct	bloom -389m-zh	bloom -1b4-zh
CN Poems	Acc	44.9	56.85	86.95	68.6	<b>93.65</b>

## 4 A COMPARISON BETWEEN HUMAN BRAIN AND LLMs

Based on the definition of memory in Section 3 and the experimental results, we believe that LLMs do possess memory capabilities. It's important to distinguish between LLM memory and database storage. Database storage involves keeping content on physical media (like hard drives or books) that can be searched or modified based on conditions, while LLM memory refers to the dynamic approximation of corresponding outputs using internal weights and inputs.

From the perspective of functionality, we argue that LLMs and human memory do not fundamentally differ; both can be understood as dynamically approximating results based on inputs. For example, as shown in Figure 2, LLMs can recite entire poems solely based on their titles and authors after learning. These poems are not stored in specific areas within the model; they are dynamically generated based on input. We can only determine if an LLM remembers certain information by posing questions and examining outputs; otherwise, it remains unknown. Human memory operates similarly: we can only validate our memories by answering specific questions; otherwise, assessment is impossible. For instance, if you ask someone how many poems they remember, they may struggle to provide an exact number, but they can usually recall a specific poem if prompted. Few people

consciously memorize how many poems they know, leading to a lack of corresponding output when such input arise. Therefore, we suggest that the brain functions like a model that dynamically fits outputs based on inputs, indicating that, in a sense, the mathematical model of the human brain may resemble that of a Transformer-based dynamic approximation UAT model, potentially even as a more advanced version. However, we believe their fundamental mechanisms are the same: both rely on dynamically fitting outputs based on inputs.

Due to the complexity of the brain, many operational mechanisms remain unclear, and no reasonable conclusions currently exist regarding its specific workings. Thus, we make logical assumptions and generalizations about the brain’s mechanisms based on the memory process, UAT theory, and LLMs. First, we extend the concept of memory in the brain to other cognitive abilities, such as social skills, imagination, and creativity. All of these can be attributed to the ability to infer outcomes based on existing knowledge and inputs, which we collectively refer to as reasoning ability, defined as: the capacity to generate specific results based on previously learned knowledge and specific inputs, where these results are consistent with or related to that knowledge.

Based on the definition of reasoning ability, LLM memory can also be viewed as a form of reasoning. The results from Figures 2 and 3, along with the current performance of ChatGPT-4 (Achiam et al., 2023) in generating outputs based on inputs, suggest that LLMs possess reasoning capabilities. Although the predictions in Figure 3 are incorrect, they still align with linguistic conventions and somewhat correspond to the titles of the poems. This can be seen as creativity.

So why do LLMs seem to underperform in reasoning tasks? We believe there are two main factors: model size, and data quality and quantity.

- **Model Size:** Generally, larger LLMs tend to be more powerful. Theoretically, as demonstrated by UAT2LLMs (Wang & Li, 2024b) and UAT2Parallel (Wang & Li, 2024a), a greater model size enhances dynamic fitting capability, leading to improved performance. Performance improvements can also be observed when comparing models like Llama from 8B to 70B (Dubey et al., 2024) and Qwen2 (Yang et al., 2024a) from 0.5B to 72B—larger models consistently show better performance.
- **Data Quality and Quantity:** Current LLMs have significantly benefited from training on vast datasets. The larger and higher the quality of dataset, the stronger the model’s performance. The performance leap from Qwen 1.5 (Bai et al., 2023) to Qwen 2 (Yang et al., 2024a) highlights that training on high-quality data yields better results. From a human learning perspective, individuals undergo decades of education from elementary school to university. Immersed in a language-rich environment from birth, humans benefit from teachers and exams that correct linguistic issues one by one. Without such learning experiences, we would struggle to develop robust language skills.

Since we propose that both LLMs and the brain function as dynamic models that fit outputs, why build such dynamic models? What are the advantages of this approach? We believe that this dynamic fitting capability gives the brain infinite possibilities. The brain doesn’t need to remember everything; it only needs to focus on what is important. Imagine if a newborn’s brain were pre-loaded with the weights of its parents; there would be no need to fit the world because most of the external environment remains constant. In such a scenario, the brain’s weights would hardly ever be updated, and the person would lose creativity. However, when we are born, our brains contain almost no knowledge about the external world. It could be viewed as the weight parameters in our brains are randomly initialized, and it is precisely this randomness that gives rise to creativity. Since the knowledge of our ancestors isn’t always correct, newborns learn from their predecessors and continuously interact with the world to verify and update this knowledge. This updating process involves optimizing the brain’s weight parameters. Each update may be right or wrong, but with a vast number of humans exploring the world, we gradually inch closer to the truth, ultimately leading to innovation.

A great example of dynamic fitting in the brain is Henry Molaison (Scoville & Milner, 1957; Victor et al., 1961; Milner & Klein, 2015). After his hippocampus (Bliss & Collingridge, 1993; Squire, 1992; Erickson et al., 2011; Eckardt, 1980) was damaged, he could no longer form new long-term memories, though his existing memories remained intact. We believe that the hippocampus acts as a switch controlling whether the weights responsible for long-term memory in the brain can be updated. Once the hippocampus is damaged, the brain’s weight parameters can no longer change,



meaning that while past inputs (before the hippocampal damage) can still produce corresponding outputs (i.e., recalling past events), the inability to update weights prevents the formation of new memories.

## 5 CONCLUSION

In this paper, we demonstrate that LLMs possess memory capabilities, which are enabled by their Transformer-based architecture. This architecture functions as a dynamic fitting UAT model, with a strong ability to adaptively fit outputs. As a result, LLMs can recall entire content based on minimal input information. Since this memory can only be confirmed when triggered by input, we refer to it as "Schrödinger's memory." Through extensive experiments, we validated that the memory mechanism of LLMs aligns with this theory. Additionally, we compared LLMs with the human brain and found that their working mechanisms are similar, as both dynamically fit outputs based on inputs.

## REFERENCES

OpenAI Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mo Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Benjamin Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Raphael Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Lukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Hendrik Kirchner, Jamie Ryan Kiros, Matthew Knight, Daniel Kokotajlo, Lukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel P. Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mely, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Ouyang Long, Cullen O’Keefe, Jakub W. Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alexandre Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Pondé de Oliveira Pinto, Michael Pokorný, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack W. Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario D. Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin D. Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas A. Tezak, Madeleine Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vi-

- jayvergiya, Chelsea Voss, Carroll L. Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qim ing Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report. 2023. URL <https://api.semanticscholar.org/CorpusID:257532815>.
- Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.
- Tim V. P. Bliss and Graham L. Collingridge. A synaptic model of memory: long-term potentiation in the hippocampus. *Nature*, 361:31–39, 1993. URL <https://api.semanticscholar.org/CorpusID:4326182>.
- Thorsten Brants, Ashok Papat, Peng Xu, Franz Josef Och, and Jeffrey Dean. Large language models in machine translation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pp. 858–867, 2007.
- Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramèr, and Chiyuan Zhang. Quantifying memorization across neural language models. *ArXiv*, abs/2202.07646, 2022. URL <https://api.semanticscholar.org/CorpusID:246863735>.
- Yukang Chen, Shengju Qian, Haotian Tang, Xin Lai, Zhijian Liu, Song Han, and Jiaya Jia. Longlora: Efficient fine-tuning of long-context large language models. *ArXiv*, abs/2309.12307, 2023. URL <https://api.semanticscholar.org/CorpusID:262084134>.
- G. Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems*, pp. 303–314, Jan 2007. doi: 10.1007/bf02551274. URL <http://dx.doi.org/10.1007/bf02551274>.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Manan Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur

Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkan Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabza, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keenally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Tim-

- othy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vitor Albiero, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. The llama 3 herd of models, 2024. URL <https://arxiv.org/abs/2407.21783>.
- Michael J. Eckardt. The hippocampus as a cognitive map. *Journal of Nervous and Mental Disease*, 168:191–192, 1980. URL <https://api.semanticscholar.org/CorpusID:144502478>.
- Kirk I. Erickson, Michelle W. Voss, Ruchika Shaurya Prakash, Chandramallika Basak, Amanda N. Szabo, Laura Chaddock, Jennifer S. Kim, Susie Heo, Heloisa Alves, Siobhan M. White, Thomas R Wójcicki, Emily L. Mailey, Victoria Jeanne Vieira, Stephen A. Martin, Brandt D. Pence, Jeffrey A. Woods, Edward McAuley, and Arthur F. Kramer. Exercise training increases size of hippocampus and improves memory. *Proceedings of the National Academy of Sciences*, 108:3017 – 3022, 2011. URL <https://api.semanticscholar.org/CorpusID:13980824>.
- WeiZhi Fei, Xueyan Niu, Pingyi Zhou, Lu Hou, Bo Bai, Lei Deng, and Wei Han. Extending context window of large language models via semantic compression. In *Annual Meeting of the Association for Computational Linguistics*, 2023. URL <https://api.semanticscholar.org/CorpusID:266335580>.
- Alex Graves, Greg Wayne, and Ivo Danihelka. Neural turing machines. *ArXiv*, abs/1410.5401, 2014. URL <https://api.semanticscholar.org/CorpusID:15299054>.
- Youngju Han and Rod Ellis. Implicit knowledge, explicit knowledge and general language proficiency. *Language teaching research*, 2(1):1–23, 1998.
- Kurt Hornik, Maxwell B. Stinchcombe, and Halbert L. White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2:359–366, 1989. URL <https://api.semanticscholar.org/CorpusID:2757547>.
- Matthew Jagielski, Om Thakkar, Florian Tramèr, Daphne Ippolito, Katherine Lee, Nicholas Carlini, Eric Wallace, Shuang Song, Abhradeep Thakurta, Nicolas Papernot, and Chiyuan Zhang. Measuring forgetting of memorized training examples. *ArXiv*, abs/2207.00099, 2022. URL <https://api.semanticscholar.org/CorpusID:250243645>.
- Rui Mao, Qian Liu, Kai He, Wei Li, and Erik Cambria. The biases of pre-trained language models: An empirical study on prompt-based sentiment analysis and emotion detection. *IEEE transactions on affective computing*, 14(3):1743–1753, 2022.
- Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq R. Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *ArXiv*, abs/2203.10244, 2022. URL <https://api.semanticscholar.org/CorpusID:247593713>.
- George Armitage Miller. Language and communication. 1951.
- B. Milner and Denise Klein. Loss of recent memory after bilateral hippocampal lesions: memory and memories—looking back and looking forward. *Journal of Neurology, Neurosurgery & Psychiatry*, 87:230 – 230, 2015. URL <https://api.semanticscholar.org/CorpusID:13234383>.
- Yasmin Moslem, Rejwanul Haque, John D Kelleher, and Andy Way. Adaptive machine translation with large language models. *arXiv preprint arXiv:2301.13294*, 2023.
- OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, Red

Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rameesh Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O’Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2024. URL <https://arxiv.org/abs/2303.08774>.

Marius-Constantin Popescu, Valentina E Balas, Liliana Perescu-Popescu, and Nikos Mastorakis. Multilayer perception and neural networks. *WSEAS Transactions on Circuits and Systems*, 8(7): 579–588, 2009.

William Beecher Scoville and B. Milner. Loss of recent memory after bilateral hippocampal lesions. *Journal of Neurology, Neurosurgery & Psychiatry*, 20:11 – 21, 1957. URL <https://api.semanticscholar.org/CorpusID:20365179>.

Larry R. Squire. Memory and the hippocampus: a synthesis from findings with rats, monkeys, and humans. *Psychological review*, 99 2:195–231, 1992. URL <https://api.semanticscholar.org/CorpusID:14104324>.

Unknown. Cn poems. <https://huggingface.co/datasets/larryvrh/Chinese-Poems>, 2024a. Accessed: 23 September 2024.

- Unknown. Eng poems. <https://huggingface.co/datasets/jnb666/poems>, 2024b. Accessed: 23 September 2024.
- Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerová, et al. Adapted large language models can outperform medical experts in clinical text summarization. *Nature medicine*, 30(4):1134–1142, 2024.
- Maurice Victor, Jay B. Angevine, Elliott L. Mancall, and C M Fisher. [memory loss with lesions of hippocampal formation. report of a case with some remarks on the anatomical basis of memory]. *Archives of neurology*, 5:244–63, 1961. URL <https://api.semanticscholar.org/CorpusID:12965148>.
- Wei Wang and Qing Li. Universal approximation theory: Foundations for parallelism in neural networks. *ArXiv*, abs/2407.21670, 2024a. URL <https://api.semanticscholar.org/CorpusID:271571549>.
- Wei Wang and Qing Li. Universal approximation theory: The basic theory for large language models, 2024b. URL <https://arxiv.org/abs/2407.00958>.
- BigScience Workshop, :, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, Dragomir Radev, Eduardo González Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Bar Natan, Francesco De Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady Elsahar, Hamza Benyamina, Hieu Tran, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jörg Frohberg, Joseph Tobing, Joydeep Bhattacharjee, Khalid Almubarak, Kimbo Chen, Kyle Lo, Leandro Von Werra, Leon Weber, Long Phan, Loubna Ben allal, Ludovic Tanguy, Manan Dey, Manuel Romero Muñoz, Maraim Masoud, María Grandury, Mario Šaško, Max Huang, Maximin Coavoux, Mayank Singh, Mike Tian-Jian Jiang, Minh Chien Vu, Mohammad A. Jauhar, Mustafa Ghaleb, Nishant Subramani, Nora Kassner, Nurulaqilla Khamis, Olivier Nguyen, Omar Espejel, Ona de Gibert, Paulo Villegas, Peter Henderson, Pierre Colombo, Priscilla Amuok, Quentin Lhoest, Rhea Harliman, Rishi Bommasani, Roberto Luis López, Rui Ribeiro, Salomey Osei, Sampo Pyysalo, Sebastian Nagel, Shamik Bose, Shamsuddeen Hassan Muhammad, Shanya Sharma, Shayne Longpre, So-maieh Nikpoor, Stanislav Silberberg, Suhas Pai, Sydney Zink, Tiago Timponi Torrent, Timo Schick, Tristan Thrush, Valentin Danchev, Vassilina Nikoulina, Veronika Laippala, Violette Lercq, Vrinda Prabhu, Zaid Alyafeai, Zeerak Talat, Arun Raja, Benjamin Heinzerling, Chenglei Si, Davut Emre Taşar, Elizabeth Salesky, Sabrina J. Mielke, Wilson Y. Lee, Abheesht Sharma, Andrea Santilli, Antoine Chaffin, Arnaud Stiegler, Debajyoti Datta, Eliza Szczechla, Gunjan Chhablani, Han Wang, Harshit Pandey, Hendrik Strobelt, Jason Alan Fries, Jos Rozen, Leo Gao, Lintang Sutawika, M Saiful Bari, Maged S. Al-shaibani, Matteo Manica, Nihal Nayak, Ryan Teehan, Samuel Albanie, Sheng Shen, Srulik Ben-David, Stephen H. Bach, Taewoon Kim, Tali Bers, Thibault Fevry, Trishala Neeraj, Urmish Thakker, Vikas Raunak, Xiangru Tang, Zheng-Xin Yong, Zhiqing Sun, Shaked Brody, Yallow Uri, Hadar Tojarieh, Adam Roberts, Hyung Won Chung, Jaesung Tae, Jason Phang, Ofir Press, Conglong Li, Deepak Narayanan, Hatim Bourfoune, Jared Casper, Jeff Rasley, Max Ryabinin, Mayank Mishra, Minjia Zhang, Mohammad Shooeybi, Myriam Peyrounette, Nicolas Patry, Nouamane Tazi, Omar Sanseviero, Patrick von Platen, Pierre Cornette, Pierre François Lavallée, Rémi Lacroix, Samyam Rajbhandari, Sanchit Gandhi, Shaden Smith, Stéphane Requena, Suraj Patil, Tim Dettmers, Ahmed Baruwa, Amanpreet Singh, Anastasia Cheveleva, Anne-Laure Ligozat, Arjun Subramonian, Aurélie Névél, Charles Lovering, Dan Garrette, Deepak Tunuguntla, Ehud Reiter, Ekaterina Taktasheva, Ekaterina Voloshina, Eli Bogdanov, Genta Indra Winata, Hailey Schoelkopf, Jan-Christoph Kalo, Jekaterina Novikova, Jessica Zosa Forde, Jordan Clive, Jungo Kasai, Ken Kawamura, Liam Hazan, Marine Carpuat,

- Miruna Clinciu, Najoung Kim, Newton Cheng, Oleg Serikov, Omer Antverg, Oskar van der Wal, Rui Zhang, Ruochen Zhang, Sebastian Gehrmann, Shachar Mirkin, Shani Pais, Tatiana Shavrina, Thomas Scialom, Tian Yun, Tomasz Limisiewicz, Verena Rieser, Vitaly Protasov, Vladislav Mikhailov, Yada Pruksachatkun, Yonatan Belinkov, Zachary Bamberger, Zdeněk Kasner, Alice Rueda, Amanda Pestana, Amir Feizpour, Ammar Khan, Amy Faranak, Ana Santos, Anthony Hevia, Antigona Unldreaj, Arash Aghagol, Arezoo Abdollahi, Aycha Tammour, Azadeh Haji-Hosseini, Bahareh Behroozi, Benjamin Ajibade, Bharat Saxena, Carlos Muñoz Ferrandis, Daniel McDuff, Danish Contractor, David Lansky, Davis David, Douwe Kiela, Duong A. Nguyen, Edward Tan, Emi Baylor, Ezinwanne Ozoani, Fatima Mirza, Frankline Ononiwu, Habib Rezanejad, Hessie Jones, Indrani Bhattacharya, Irene Solaiman, Irina Sedenko, Isar Nejadgholi, Jesse Passmore, Josh Seltzer, Julio Bonis Sanz, Livia Dutra, Mairon Samagaio, Maraim Elbadri, Margot Mieskes, Marissa Gerchick, Martha Akinlolu, Michael McKenna, Mike Qiu, Muhammed Ghauri, Mykola Burynok, Nafis Abrar, Nazneen Rajani, Nour Elkott, Nour Fahmy, Olanrewaju Samuel, Ran An, Rasmus Kromann, Ryan Hao, Samira Alizadeh, Sarmad Shubber, Silas Wang, Sourav Roy, Sylvain Viguier, Thanh Le, Tobi Oyebade, Trieu Le, Yoyo Yang, Zach Nguyen, Abhinav Ramesh Kashyap, Alfredo Palasciano, Alison Callahan, Anima Shukla, Antonio Miranda-Escalada, Ayush Singh, Benjamin Beilharz, Bo Wang, Caio Brito, Chenxi Zhou, Chirag Jain, Chuxin Xu, Clémentine Fourier, Daniel León Perrián, Daniel Molano, Dian Yu, Enrique Manjavacas, Fabio Barth, Florian Fuhrmann, Gabriel Altay, Giyaseddin Bayrak, Gully Burns, Helena U. Vrabec, Imane Bello, Ishani Dash, Jihyun Kang, John Giorgi, Jonas Golde, Jose David Posada, Karthik Rangasai Sivaraman, Lokesh Bulchandani, Lu Liu, Luisa Shinzato, Madeleine Hahn de Bykhovetz, Maiko Takeuchi, Marc Pàmies, Maria A Castillo, Marianna Nezhurina, Mario Sängler, Matthias Samwald, Michael Cullan, Michael Weinberg, Michiel De Wolf, Mina Mihaljcic, Minna Liu, Moritz Freidank, Myungsun Kang, Natasha Seelam, Nathan Dahlberg, Nicholas Michio Broad, Nikolaus Muellner, Pascale Fung, Patrick Haller, Ramya Chandrasekhar, Renata Eisenberg, Robert Martin, Rodrigo Canalli, Rosaline Su, Ruisi Su, Samuel Cahyawijaya, Samuele Garda, Shlok S Deshmukh, Shubhanshu Mishra, Sid Kiblawi, Simon Ott, Sinee Sang-aaroonsiri, Srishti Kumar, Stefan Schweter, Sushil Bharati, Tanmay Laud, Théo Gigant, Tomoya Kainuma, Wojciech Kusa, Yanis Labrak, Yash Shailesh Bajaj, Yash Venkatraman, Yifan Xu, Yingxin Xu, Yu Xu, Zhe Tan, Zhongli Xie, Zifan Ye, Mathilde Bras, Younes Belkada, and Thomas Wolf. Bloom: A 176b-parameter open-access multilingual language model, 2023. URL <https://arxiv.org/abs/2211.05100>.
- Yuhuai Wu, Markus Norman Rabe, DeLesley S. Hutchins, and Christian Szegedy. Memorizing transformers. *ArXiv*, abs/2203.08913, 2022. URL <https://api.semanticscholar.org/CorpusID:247519194>.
- Chaojun Xiao, Pengl Zhang, Xu Han, Guangxuan Xiao, Yankai Lin, Zhengyan Zhang, Zhiyuan Liu, Song Han, and Maosong Sun. Inllm: Training-free long-context extrapolation for llms with an efficient context memory. 2024. URL <https://api.semanticscholar.org/CorpusID:267523068>.
- Fangyuan Xu, Yixiao Song, Mohit Iyyer, and Eunsol Choi. A critical evaluation of evaluations for long-form question answering. *ArXiv*, abs/2305.18201, 2023. URL <https://api.semanticscholar.org/CorpusID:258960565>.
- An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024a.
- Hongkang Yang, Zehao Lin, Wenjin Wang, Hao Wu, Zhiyu Li, Bo Tang, Wenqiang Wei, Jinbo Wang, Zeyun Tang, Shichao Song, Chenyang Xi, Yu Yu, Kai Chen, Feiyu Xiong, Linpeng Tang, and E Weinan. Memory3: Language modeling with explicit memory. *ArXiv*, abs/2407.01178, 2024b. URL <https://api.semanticscholar.org/CorpusID:270870116>.

- Jianxin Yang. Longqlora: Efficient and effective method to extend context length of large language models. *ArXiv*, abs/2311.04879, 2023. URL <https://api.semanticscholar.org/CorpusID:265050454>.
- Xing Yu. Language and state: A theory of the progress of civilization. UPA, 2015.
- Boyu Zhang, Hongyang Yang, Tianyu Zhou, Muhammad Ali Babar, and Xiao-Yang Liu. Enhancing financial sentiment analysis via retrieval augmented large language models. In *Proceedings of the fourth ACM international conference on AI in finance*, pp. 349–356, 2023a.
- Haoyu Zhang, Jianjun Xu, and Ji Wang. Pretraining-based natural language generation for text summarization. *arXiv preprint arXiv:1902.09243*, 2019.
- Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Jialin Pan, and Lidong Bing. Sentiment analysis in the era of large language models: A reality check. *arXiv preprint arXiv:2305.15005*, 2023b.
- Zhenyu (Allen) Zhang, Ying Sheng, Tianyi Zhou, Tianlong Chen, Lianmin Zheng, Ruisi Cai, Zhao Song, Yuandong Tian, Christopher Ré, Clark W. Barrett, Zhangyang Wang, and Beidi Chen. H2o: Heavy-hitter oracle for efficient generative inference of large language models. *ArXiv*, abs/2306.14048, 2023c. URL <https://api.semanticscholar.org/CorpusID:259263947>.
- Dawei Zhu, Nan Yang, Liang Wang, Yifan Song, Wenhao Wu, Furu Wei, and Sujian Li. Pose: Efficient context window extension of llms via positional skip-wise training. *ArXiv*, abs/2309.10400, 2023. URL <https://api.semanticscholar.org/CorpusID:262053659>.