
From Human Memory to AI Memory: A Survey on Memory Mechanisms in the Era of LLMs

Yaxiong Wu, Sheng Liang, Chen Zhang, Yichao Wang, Yongyue Zhang,
Huifeng Guo, Ruiming Tang, Yong Liu
Huawei Noah's Ark Lab
wu.yaxiong@huawei.com

Abstract

Memory is the process of encoding, storing, and retrieving information, allowing humans to retain experiences, knowledge, skills, and facts over time, and serving as the foundation for growth and effective interaction with the world. It plays a crucial role in shaping our identity, making decisions, learning from past experiences, building relationships, and adapting to changes. In the era of large language models (LLMs), memory refers to the ability of an AI system to retain, recall, and use information from past interactions to improve future responses and interactions. Although previous research and reviews have provided detailed descriptions of memory mechanisms, there is still a lack of a systematic review that summarizes and analyzes the relationship between the memory of LLM-driven AI systems and human memory, as well as how we can be inspired by human memory to construct more powerful memory systems. To achieve this, in this paper, we propose a comprehensive survey on the memory of LLM-driven AI systems. In particular, we first conduct a detailed analysis of the categories of human memory and relate them to the memory of AI systems. Second, we systematically organize existing memory-related work and propose a categorization method based on three dimensions (object, form, and time) and eight quadrants. Finally, we illustrate some open problems regarding the memory of current AI systems and outline possible future directions for memory in the era of large language models.

1 Introduction

Recently, large language models (LLMs) have become the core component of AI systems due to their powerful language understanding and generation capabilities, and are widely used in various applications such as intelligent customer service, automated writing, machine translation, information retrieval, and sentiment analysis [1–4]. Unlike traditional AI systems, which rely on predefined rules and manually labeled features, LLM-driven AI systems offer greater flexibility, handling a diverse range of tasks with enhanced adaptability and contextual awareness. Moreover, the introduction of memory enables LLMs to retain historical interactions with users and store contextual information, thereby providing more personalized, continuous, and context-aware responses in future interactions [2, 5, 6]. AI systems powered by LLMs with memory capabilities will not only elevate the user experience but also support more complex and dynamic use cases, steering AI technology toward greater intelligence and human-centric design [7, 8].

In neuroscience, human memory refers to the brain's ability to store, retain, and recall information [9, 10]. Human memory serves as the foundation for understanding the world, learning new knowledge, adapting to the environment, and making decisions, allowing us to preserve past experiences, skills, and knowledge, and helping us form our personal identity and behavior patterns [11]. Human memory can be broadly classified into *short-term memory* and *long-term memory* based

on the duration of new memory formation [12]. Short-term memory refers to the information we temporarily store and process, typically lasting from a few seconds to a few minutes, and includes sensory memory and working memory [11]. Long-term memory refers to the information we can store for extended periods, ranging from minutes to years, and includes declarative *explicit memory* (such as episodic and semantic memory) and non-declarative *implicit memory* (such as conditioned reflexes and procedural memory) [11]. Human memory is a complex and dynamic process that relies on different memory systems to process information for various purposes, influencing how we understand and respond to the world. The different types of human memory and their working mechanisms can greatly inspire us to develop more scientific and reasonable memory-enhanced AI systems [13–16].

In the era of large language models (LLMs), the most typical memory-enhanced AI system is the LLM-powered autonomous agent system [10]. Large language model (LLM) powered agents are AI systems that can perform complex tasks using natural language, incorporating capabilities like planning, tool use, memory, and multi-step reasoning to enhance interactions and problem-solving [1, 2, 10]. This memory-enhanced AI system is capable of autonomously decomposing complex tasks, remembering interaction history, and invoking and executing tools, thereby efficiently completing a series of intricate tasks. In particular, memory, as a key component of the LLM-powered agent, can be defined as the process of acquiring, storing, retaining, and subsequently retrieving information [10]. It enables the large language model to overcome the limitation of LLM’s context window, allowing the agent to recall interaction history and make more accurate and intelligent decisions. For instance, MemoryBank [17] proposed a long-term memory mechanism to allow LLMs for retrieving relevant memories, continuously evolving through continuous updates, and understanding and adapting to a user’s personality by integrating information from previous interactions. In addition, many commercial and open-source AI systems have also integrated memory systems to enhance the personalization capabilities of the system, such as OpenAI ChatGPT Memory [18], Apple Personal Context [19], mem0 [20], MemoryScope [21], etc.

Although previous studies and reviews have provided detailed explanations of memory mechanisms, most of the existing work focuses on analyzing and explaining memory from the temporal (time) dimension, specifically in terms of short-term and long-term memory [8, 7, 17]. We believe that categorizing memory solely based on the *time* dimension is insufficient, as there are many other aspects (such as *object* and *form*) to memory classification in AI systems. For example, from the object dimension, since AI systems often interact with humans, they need to perceive, store, recall, and use memories related to individual users, thus generating personal memories. Meanwhile, when AI systems perform complex tasks, they generate intermediate results (such as reasoning and planning processes, internet search results, etc.), which form system memory. In addition, from the form dimension, since AI systems are powered by large language models (LLMs), they can store memories through the parametric memory encoded within the model parameters, as well as through non-parametric memory in the form of external memory documents that are stored and managed outside the model. Therefore, insights that consider memory from the perspectives of object (personal and system), form (parametric and non-parametric), and time (short-term and long-term) are still lacking in the current era of large language models. There is still no comprehensive review that systematically analyzes the relationship between memory in LLM-driven AI systems and human memory, and how insights from human memory can be leveraged to build more efficient and powerful memory systems.

To fill this gap, this paper presents a comprehensive review of the memory mechanisms in LLM-driven AI systems. First, we provide a detailed analysis of the categories of human memory and relate them to the memory systems in AI. In particular, we explore how human memory types — short-term memory (including sensory memory and working memory) and long-term memory (including explicit memory and implicit memory) — correspond to personal and system memory, parametric and non-parametric memory, and short-term and long-term memory in LLM-driven AI systems. Next, we systematically organize the existing work related to memory and propose a classification method based on three dimensions (*object*, *form*, and *time*) with eight quadrants. In the object dimension, memory can be divided into personal memory and system memory; in the form dimension, it can be classified into parametric memory and non-parametric memory; in the time dimension, memory can be categorized into short-term memory and long-term memory. Finally, based on the classification results from the three dimensions and eight quadrants mentioned above, we an-

alyze some open issues in the memory of current AI systems and outline potential future directions for memory development in the era of large language models.

The main contributions of this paper are summarized as follows: (1) We systematically and comprehensively define LLM-driven AI systems’ memory and establish corresponding relationships with human memory. (2) We propose a classification method for memory based on three dimensions (object, form, and time) and eight quadrants, which facilitates a more systematic exploration of memory in the era of large language models. (3) From the perspective of enhancing personalized capabilities, we analyze and summarize research related to personal memory. (4) From the perspective of AI system’s ability to perform complex tasks, we analyze and summarize research related to system memory. (5) We identify the existing issues and challenges in current memory research and point out potential future directions for development.

The remainder of the paper is organized as follows: In Section 2, we present a detailed description of human memory and AI systems’ memory, comparing their differences and relationships, and introduce the classification method for memory based on three dimensions (object, form, and time) and eight quadrants. In Section 3, we summarize research related to personal memory, aimed at enhancing the personalized response capabilities of AI systems. In Section 4, we summarize research related to system memory, aimed at improving AI systems’ ability to perform complex tasks. In Section 5, we analyze some open issues related to memory and point out potential future directions for development. Finally, in Section 6, we conclude the survey.

2 Overview

The human brain has evolved complex yet efficient memory mechanisms over a long period, enabling it to encode, store, and recall information effectively [9]. Accordingly, in the development of AI systems, we can draw insights from human memory to design effective & efficient memory mechanisms or systems. In this section, we will first describe in detail the complex memory mechanisms and related memory systems of the human brain from the perspective of memory neuroscience. Then, we will discuss the memory mechanisms and types specific to LLM-driven AI systems. Finally, based on the memory features of LLM-driven AI systems, we will systematically review and categorize existing work from different dimensions.

2.1 Human Memory

Human memory typically relies on different memory systems to process information for various purposes, such as working memory for temporarily storing and processing information to support ongoing cognitive activities, and episodic memory for recording personal experiences and events for a long time [11].

2.1.1 Short-Term and Long-Term Memory

Based on the time range, human memory can be roughly divided into *short-term memory* and *long-term memory* according to the well-known Multi-Store Model (or Atkinson-Shiffrin Memory Model) [22].

Short-Term Memory Short-term memory is a temporary storage system that holds small amounts of information for brief periods, typically ranging from seconds to minutes. It includes *sensory memory*, which briefly captures raw sensory information from the environment (like sights or sounds), and *working memory*, which actively processes and manipulates information to complete tasks such as problem-solving or learning. Together, these components allow humans to temporarily hold and work with information before either discarding it or transferring it to long-term memory.

- **Sensory memory:** Sensory memory is the brief storage of sensory information we acquire from the external world, including iconic memory (visual), echoic memory (auditory), haptic memory (touch), and other sensory data. It typically lasts only a few milliseconds to a few seconds. Some sensory memories are transferred to working memory, while others are eventually stored in long-term memory (such as episodic memory).
- **Working memory:** Working memory is the system we use to temporarily store and process information. It not only helps us maintain current thoughts but also plays a role in decision-

making and problem-solving. For example, when solving a math problem, it allows us to keep track of both the problem and the steps involved in finding the solution.

Long-Term Memory Long-term memory is a storage system that holds information for extended periods, ranging from minutes to a lifetime. It includes *explicit memory*, which involves conscious recall of facts and events, and *implicit memory*, which involves unconscious skills and habits, like riding a bike. These two types work together to help humans retain knowledge, experiences, and learned abilities over time.

- **Explicit memory:** Explicit memory, also known as *declarative memory*, refers to memories that we can easily verbalize or declare. It can be further divided into episodic memory and semantic memory. *Episodic memory* refers to memories related to personal experiences and events, such as what you had for lunch. This type of memory is typically broken down into stages like encoding, storage, and retrieval. *Semantic memory*, on the other hand, refers to memories related to facts and knowledge, such as knowing that the Earth is round or that the Earth orbits the Sun.
- **Implicit memory:** Implicit memory, also known as *non-declarative memory*, refers to memories that are difficult to describe in words. It is associated with habits, skills, and procedures, and does not require conscious recall. *Procedural memory* (or "muscle memory") is a typical form of implicit memory. It refers to memories gained through actions, such as riding a bicycle or playing the piano. The planning and coordination of movements are key components of procedural memory.

Multiple memory systems typically operate simultaneously, storing information in various ways across different brain regions. These memory systems are not completely independent; they interact with each other and, in many cases, depend on one another. For example, when you hear a new song, the sensory memory in your ears and the brain regions responsible for processing sound will become active, storing the sound of the song for a few seconds. This sound is then transferred to your working memory system. As you use your working memory and consciously think about the song, your episodic memory will automatically activate, recalling where you heard the song and what you were doing at the time. As you hear the song in different places and at different times, a new semantic memory gradually forms, linking the melody of the song with its title. So, when you hear the song again, you'll remember the song's title, rather than a specific instance from your multiple listening experiences. When you practice playing the song on the guitar, your procedural memory will remember the finger movements involved in playing the song.

2.1.2 Memory Mechanisms

Memory is the ability to encode, store and recall information. The three main processes involved in human memory are therefore *encoding* (the process of acquiring and processing information into a form that can be stored), *storage* (the retention of encoded information over time in short-term or long-term memory), and *retrieval* (*recall*, the process of accessing and bringing stored information back into conscious awareness when needed).

- **Encoding** Memory encoding is the process of changing sensory information into a form that our brain can cope with and store effectively. In particular, there are different types of encoding in terms of how information is processed, such as *visual encoding*, which involves processing information based on its visual features like color, shape, or texture; *acoustic encoding*, which focuses on the auditory characteristics of information, such as pitch, tone, or rhythm; and *semantic encoding*, which is based on the meaning of the information, making it easier to structure and remember. In addition, there are many approaches to make our brain better at encoding memory, such as *mnemonics*, which involve using acronyms or peg-word systems to aid recall, *chunking*, where information is broken down into smaller, meaningful units to enhance retention, *imagination*, which strengthens encoding by linking images to words, and *association*, where new information is connected to prior knowledge to improve understanding and long-term memory storage.
- **Storage** The storage of memory involves the coordinated activity of multiple brain regions, with key areas including: the *prefrontal cortex*, which is associated with working memory and decision-making, helping us maintain and process information in the short term; the

hippocampus, which helps organize and consolidate information to form new explicit memories (such as episodic memory); the *cerebral cortex*, which is involved in the storage and retrieval of semantic memory, allowing us to retain facts, concepts, and general knowledge over time; and the *cerebellum*, which is primarily responsible for procedural memory formed through repetition.

- **Retrieval** Memory retrieval is the ability to access information and get it out of the memory storage. When we recall something, the brain reactivates neural pathways (also called synapses) linked to that memory. The prefrontal cortex helps in bringing memories back to awareness. Similarly, there are different types of memory retrieval, including *recognition*, where we identify previously encountered information or stimuli, such as recognizing a familiar face or a fact we have learned before; *recall*, which is the ability to retrieve information from memory without external cues, like remembering a phone number or address from memory; and *relearning*, a process in which we reacquire previously learned but forgotten information, often at a faster pace than initial learning due to the residual memory traces that still exist.

In addition to the fundamental memory processing stages of encoding, storage, and retrieval, human memory also includes *consolidation* (the process of stabilizing and strengthening memories to facilitate long-term storage), *reconsolidation* (the modification or updating of previously stored memories when they are reactivated, allowing them to adapt to new information or contexts), *reflection* (the active review and evaluation of one's memories to enhance self-awareness, improve learning strategies, and optimize decision-making), and *forgetting* (the process by which information becomes inaccessible).

- **Consolidation** Memory consolidation refers to the process of converting short-term memory into long-term memory, allowing information to be stably stored in the brain and reducing the likelihood of forgetting. It primarily involves the hippocampus and strengthens neural connections through *synaptic plasticity* (strengthening of connections between neurons) and *systems consolidation* (the gradual transfer and reorganization of memories from the hippocampus to the neocortex for long-term storage).
- **Reconsolidation** Memory reconsolidation refers to the process in which a previously stored memory is reactivated, entering an unstable state and requiring reconsolidation to maintain its storage. This process allows for the modification or updating of existing memories to adapt to new information or contexts, potentially leading to memory enhancement, weakening, or distortion. Once a memory is reactivated, it involves the hippocampus and amygdala and may be influenced by emotions, cognitive biases, or new information, resulting in memory adjustment or reshaping.
- **Reflection** Memory reflection refers to the process in which an individual actively reviews, evaluates, and examines their own memory content and processes to enhance self-awareness, adjust learning strategies, or optimize decision-making. It helps improve metacognitive ability, correct memory biases, facilitate deep learning, and regulate emotions. This process primarily relies on the brain's metacognitive ability (Metacognition) and involves the prefrontal cortex, which monitors and regulates memory functions.
- **Forgetting** Forgetting is a natural process that occurs when the brain fails to retrieve or retain information, which can result from *encoding failure* (when information is not properly encoded due to lack of attention or meaningful connection), *memory decay* (when memories fade over time without reinforcement as neural connections weaken), *interference* (when similar or new memories compete with or overwrite existing ones), *retrieval failure* (when information is inaccessible due to missing contextual cues despite being stored), or *motivated forgetting* (when individuals consciously suppress or unconsciously repress traumatic or distressing memories). However, forgetting is a natural and necessary process that enables our brains to filter out irrelevant and outdated information, allowing us to prioritize what is most important for our current needs.

2.2 Memory of LLM-driven AI Systems

Similar to humans, LLM-driven AI systems also rely on memory systems to encode, store and recall information for future use. A typical example is the LLM-driven agent system, which lever-

ages memory to enhance the agent system’s abilities in reasoning, planning, personalization, and more [10].

2.2.1 Fundamental Dimensions of AI Memory

The memory of an LLM-driven AI system is closely related to the features of the LLM, that define how information is processed, stored, and retrieved based on its architecture and capabilities. We primarily categorize and organize memory based on three dimensions: *object* (personal and system memory), *form* (non-parametric and parametric memory), and *time* (short-term and long-term memory). These three dimensions comprehensively capture what type of information is retained (object), how information is stored (form), and how long it is preserved (time), aligning with both the functional structure of LLMs and practical requirements for efficient recall and adaptability.

Object Dimension The object dimension is closely tied to the interaction between LLM-driven AI systems and humans, as it defines how information is categorized based on its source and purpose. On one hand, the system receives human input and feedback (i.e., personal memory); on the other hand, it generates a series of intermediate output results during task execution (i.e., system memory). Personal memory helps the system improve its understanding of user behavior and enhances its personalization capabilities, while system memory can strengthen the system’s reasoning ability, such as in approaches like CoT (Chain-of-Thought) [23] and ReAct [24].

Form Dimension The form dimension focuses on how memory is represented and stored in LLM-driven AI systems, shaping how information is encoded and retrieved. Some memory is embedded within the model’s parameters through training, forming parametric memory, while other memory exists externally in structured databases or retrieval mechanisms, constituting non-parametric memory. Non-parametric memory serves as a supplementary knowledge source that can be dynamically accessed by the large language model, enhancing its ability to retrieve relevant information in real-time, as seen in retrieval-augmented generation (RAG) [25].

Time Dimension The time dimension defines how long memory is retained and how it influences the LLM’s interactions over different timescales. Short-term memory refers to contextual information temporarily maintained within the current conversation, enabling coherence and continuity in multi-turn dialogues. In contrast, long-term memory consists of information from past interactions that is stored in an external database and retrieved when needed, allowing the model to retain user-specific knowledge and improve personalization over time. This distinction ensures that the system can balance real-time responsiveness with accumulated learning for enhanced adaptability.

In addition to the three primary dimensions discussed above, memory can also be classified based on other criteria, such as *modality*, which distinguishes between unimodal memory (single data type) and multimodal memory (integrating multiple data types, such as text, images, and audio), or *dynamics*, which differentiates between static memory (fixed and unchanging) and streaming memory (dynamically updated in real-time). However, these alternative classifications are not considered the primary criteria here, as our focus is on the core structural aspects that most directly influence memory organization and retrieval in LLM-driven AI systems.

2.2.2 Parallels Between Human and AI Memory

The memory of LLM-driven AI system exhibits similarities to human memory in terms of structure and function. Human memory is generally categorized into short-term memory and long-term memory, a distinction that also applies to AI memory systems. Below, we draw a direct comparison between these categories, mapping human cognitive memory processes to their counterparts in intelligent AI systems. Figure 1 illustrates the parallels between human and AI memory.

- **Sensory Memory:** When an LLM-driven AI system perceives external information, it converts inputs such as text, images, speech, and video into machine-processable signals. This initial stage of information processing is analogous to human sensory memory, where raw data is briefly held before further cognitive processing. If these signals undergo additional processing, they transition into working memory, facilitating reasoning and decision-making. However, if no further processing or storage occurs, the information is quickly discarded, mirroring the transient nature of human sensory memory.

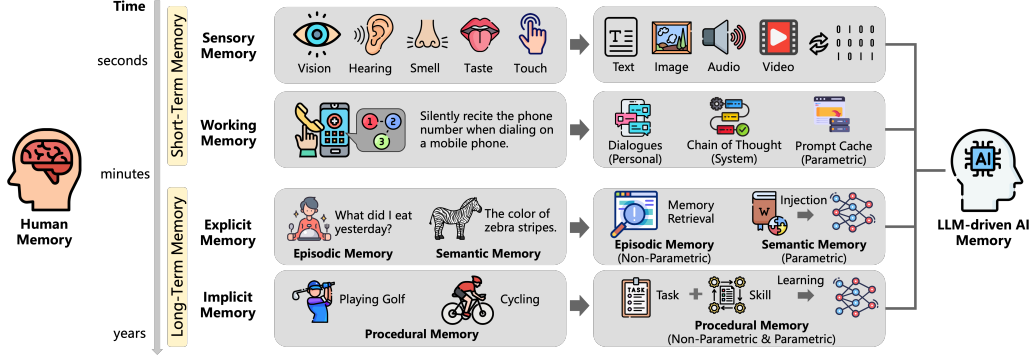


Figure 1: Illustrating the parallels between human and AI memory.

- **Working Memory:** The working memory of an AI system serves as a temporary storage and processing mechanism, enabling real-time reasoning and decision-making. It encompasses personal memory, such as contextual information retained during multi-turn dialogues, and system memory, including the chain of thoughts generated during task execution. As a form of short-term memory, working memory can undergo further processing and consolidation, eventually transitioning into long-term memory (e.g., episodic memory) that can be retrieved for future use. Additionally, during inference, large language models generate intermediate computational results, such as KV-Caches, which act as a form of parametric short-term memory that enhances efficiency by accelerating the inference process.
- **Explicit Memory:** The explicit memory of an AI system can be categorized into two distinct components. The first is non-parametric long-term memory, which involves the storage and retrieval of user-specific information, allowing the system to retain and utilize personalized data—analogueous to episodic memory in humans. The second is parametric long-term memory, where factual knowledge and learned information are embedded within the model’s parameters, forming an internalized knowledge base—corresponding to semantic memory in human cognition. Together, these components enable the system to recall past interactions and apply acquired knowledge effectively.
- **Implicit Memory:** The implicit memory of an AI system encompasses the learned processes and patterns involved in task execution, enabling the development of specialized skills for specific tasks—analogueous to procedural memory in humans. This form of memory can parallel the human process of learning from both successes and failures in a non-parameterized manner, involving the reflection and refinement of accumulated traces, which allows the retention and replication of effective strategies from past experiences. Additionally, it can be encoded within the model’s parameters, enabling the system to internalize task-related knowledge and perform operations efficiently without the need for explicit recall.

Beyond these parallels, insights from human memory can further guide the design of more effective and efficient AI memory systems, enhancing their ability to process, store, and retrieve information in a more structured and adaptive manner.

2.2.3 3D-8Q Memory Taxonomy

Building upon the three fundamental memory dimensions—object (personal & system), form (non-parametric & parametric), and time (short-term & long-term)—as well as the established parallels between human and AI memory, we propose a *three-dimensional, eight-quadrant (3D-8Q) memory taxonomy* for AI memory. This memory taxonomy systematically categorizes AI memory based on its function, storage mechanism, and retention duration, providing a structured approach to understanding and optimizing AI memory systems. Table 1 presents the eight quadrants and their respective roles and functions.

Object	Form	Time	Quadrant	Role	Function
Personal	Non-Parametric	Short-Term	I	Working Memory	Supports real-time context supplementation, enhancing the AI’s ability to maintain coherent interactions within a session.
		Long-Term	II	Episodic Memory	Enables memory retention beyond session limits, allowing the system to recall and retrieve past user interactions for personalization.
	Parametric	Short-Term	III	Working Memory	Temporarily enhances contextual understanding in ongoing interactions, improving response relevance and coherence.
		Long-Term	IV	Semantic Memory	Facilitates the continuous integration of newly acquired knowledge into the model, improving adaptability and personalization
System	Non-Parametric	Short-Term	V	Working Memory	Assists in complex reasoning and decision-making by storing intermediate outputs such as chain-of-thought prompts.
		Long-Term	VI	Procedural Memory	Captures historical experiences and self-reflection insights, enabling the AI to refine its reasoning and problem-solving skills over time.
	Parametric	Short-Term	VII	Working Memory	Enhances computational efficiency through temporary parametric storage mechanisms such as KV-Caches, optimizing inference speed and reducing resource consumption.
		Long-Term	VIII	Semantic Memory Procedural Memory	Forms a foundational knowledge base encoded in the model’s parameters, serving as a long-term repository of factual & conceptual knowledge and task-related knowledge.

Table 1: Three-dimensional, eight-quadrant (3D-8Q) memory taxonomy for LLM-driven AI systems.

Next, we will provide insights and descriptions of existing works from the perspectives of personal memory (in Section 3) and system memory (in Section 4). In particular, personal memory focuses more on the individual data perceived and observed by the model from the environment, while system memory emphasizes the system’s internal or endogenous memory, such as the intermediate memory generated during task execution.

3 Personal Memory

Personal memory refers to the process of storing and utilizing human input and response data during interactions with an LLM-driven AI system. The development and application of personal memory play a crucial role in enhancing AI systems’ personalization capabilities and improving user experience. In this section, we explore the concept of personal memory and relevant research, examining both non-parametric and parametric approaches to its construction and implementation. Table 2 shows the categories, features, and related research work of personal memory.

3.1 Contextual Personal Memory

In personal memory, the non-parametric contextual memory that can be loaded is generally divided into two categories: the short-term memory of the current session’s multi-turn dialogue and the long-term memory of historical dialogues across sessions. The former can effectively supplement contextual information, while the latter can effectively fill in missing information and overcome the limitations of context length.

3.1.1 Loading Multi-Turn Dialogue (Quadrant-I)

In multi-turn dialogue scenarios, the conversation history of the current session can significantly enhance the LLM-driven AI system’s understanding of the user’s real-time intent, leading to more relevant and contextually appropriate responses. Many modern dialogue systems are capable of

Quadrant	Dimension	Feature	Models
I	Personal Non-Parametric Short-Term	Multi-Turn Dialogue	ChatGPT [26], DeepSeek-Chat [27], Claude [28], QWEN-CHAT [29], Llama 2-Chat [30], Gemini [31], PANGU-BOT [32], ChatGLM [33], OpenAssistant [34]
II	Personal Non-Parametric Long-Term	Personal Assistant	ChatGPT Memory [18], Apple Intelligence [19], Microsoft Recall [35], Me.bot [36]
		Open-Source Framework	MemoryScope [21], mem0 [20], Memory [37], LangGraph Memory [38], Charlie Mnemonic [39], Memobase [40], Letta [41], Cognee [42]
		Construction	MPC [43], RET-LLM [44], MemoryBank [17], MemGPT [45], KGT [46], Evolving Conditional Memory [47], SECOM [48], Memory ³ [49], MemInsight [50]
		Management	MemoChat [51], MemoryBank [17], RMM [52], LD-Agent [53], A-MEM [54], Generative Agents [55], EMG-RAG [56], KGT [46], LLM-Rsum [57], COMEDY [58]
		Retrieval	RET-LLM [44], ChatDB [59], Human-like Memory [60], HippoRAG [13], HippoRAG 2 [61], EgoRAG [62], MemInsight [50]
		Usage	MemoCRS [63], RecMind [64], RecAgent [65], InteRecAgent [66], SCM [67], ChatDev [68], MetaAgents [69], S ³ [70], TradingGPT [71], Memolet [72], Synaptic Resonance [14], MemReasoner [73]
		Benchmark	MADial-Bench [74], LOCOMO [75], MemDaily [76], ChMapData [77], MSC [78], MMRC [79], Ego4D [80], EgoLife [62], BABILong [81, 82]
III	Personal Parametric Short-Term	Caching for Acceleration	Prompt Cache [83], Contextual Retrieval [84]
IV	Personal Parametric Long-Term	Knowledge Editing	Character-LLM [85], AI-Native Memory [36], MemoRAG [86], Echo [87]

Table 2: Personal Memory

handling multi-turn conversations and fully consider the current dialogue context in their responses. Notable examples include ChatGPT [26], DeepSeek-Chat [27], and Claude [28], which excel at maintaining coherence and relevance over extended interactions.

For instance, ChatGPT [26] is a prime example of a multi-turn dialogue system where the conversation history of the current session serves as short-term memory, helping to supplement the contextual information of the dialogue. In ChatGPT, the dialogue memory is encoded in a role-content format, with distinct roles such as “User” and “Assistant”. This encoding allows the system to maintain clarity regarding the speaker and the flow of the conversation.

Through effective dialogue management at different levels, including “Assistant”, “Threads”, “Messages”, and “Runs”, the system can precisely track the state of each turn and each step of the conversation, ensuring continuity and consistency in interactions. Additionally, when the conversation length becomes too extensive, the dialogue system manages the conversation’s input by truncating the number of turns, thereby preventing the input from exceeding the model’s length limitations. This ensures that the system can continue processing the dialogue without losing track of essential context, maintaining the effectiveness of multi-turn interactions.

3.1.2 Memory Retrieval-Augmented Generation (Quadrant-II)

In cross-session dialogue scenarios, retrieving relevant user long-term memories from historical conversations can effectively supplement missing information in the current session, such as personal preferences and character relationships. The advantage of memory retrieval-augmented generation is that large language models (LLMs) do not need to load all multi-session conversations. Given the limited length of LLMs’ context windows—even when extended to millions of tokens—retrieving relevant information from historical sessions is also more efficient and cost-effective in terms of computation. In addition to multi-session conversations, long-term personal memory also encom-

passes users’ behavioral history, preferences, and interaction records with AI agents over an extended period of time.

By leveraging retrieval-augmented generation from long-term memory, LLM-driven AI systems can better tailor their responses and behaviors, thereby improving user satisfaction and engagement. For instance, a personal assistant that remembers a user’s preferred news sources can prioritize those outlets in daily briefings, while a recommendation system that understands past viewing habits can suggest content more aligned with the user’s tastes. Currently, many commercial and open-source platforms are striving to construct and utilize long-term memory for personalized AI systems—for example, ChatGPT Memory [18] and Me.bot [36] for personal assistants, and MemoryScope [21] and mem0 [20] as open-source frameworks. Long-term personal memory typically follows four core processing stages: *construction*, *management*, *retrieval*, and *usage*. The second section of Table 2 (organized by rows) provides an overview of existing work on personal non-parametric long-term memory, classified based on their primary contributions.

Construction The construction of user memory requires extraction and refinement from raw memory data, such as multi-turn conversations. This process is analogous to human memory consolidation—the process of stabilizing and strengthening memories to facilitate their long-term storage. Well-organized long-term memory enhances both the efficiency of storage and the effectiveness of retrieval in user memory. For example, MemoryBank [17] leverages a memory module to store conversation histories and summaries of key events, enabling the construction of a long-term user profile. Similarly, RET-LLM [44] uses its memory module to retain essential factual knowledge about the external world, allowing the agent to monitor and update real-time environmental context relevant to the user. In addition, to accommodate different types of memory, a variety of storage formats have been developed, including *key-value*, *graph*, and *vector* representations. Specifically, *key-value* formats [44, 50, 63] enable efficient access to structured information such as user facts and preferences. *Graph*-based formats [46, 13, 61, 20] are designed to capture and represent relationships among entities, such as individuals and events. Meanwhile, *vector* formats [17, 48, 20], which are typically derived from textual, visual, or audio memory representations, are utilized to encode the semantic meaning and contextual information of conversations.

Management The management of user memory involves further processing and refinement of previously constructed memories, such as deduplication, merging, and conflict resolution. This process is analogous to human memory reconsolidation and reflection, where existing memories are reactivated, updated, and integrated to maintain coherence and relevance over time. For instance, Reflective Memory Management (RMM) [52] is a user long-term memory management framework that combines Prospective Reflection for dynamic summarization with Retrospective Reflection for retrieval optimization via reinforcement learning. This dual-process approach addresses limitations such as rigid memory granularity and fixed retrieval mechanisms, enhancing the accuracy and flexibility of long-term memory management. LD-Agent [53] enhances long-term dialogue personalization and consistency by constructing personalized persona information for both users and agents through a dynamic persona modeling module, while integrating retrieved memories to optimize response generation. A-MEM [54] introduces a self-organizing memory system inspired by the Zettelkasten method [88], which constructs interconnected knowledge networks through dynamic indexing, linking, and memory evolution, enabling LLM agents to more flexibly organize, update, and retrieve long-term memories, thereby enhancing task adaptability and contextual awareness. In addition, MemoryBank [17] incorporates a memory updating mechanism inspired by the Ebbinghaus Forgetting Curve [89], allowing the AI to forget or reinforce memories based on the time elapsed and their relative importance, thereby enabling a more human-like memory system and enhancing the user experience.

Retrieval Retrieving personal memory involves identifying memory entries relevant to the user’s current request, and the retrieval method is closely tied to how the memory is stored. For key-value memory, ChatDB [59] performs retrieval using SQL queries over structured databases. RET-LLM [44], on the other hand, employs a fuzzy search to retrieve triplet-structured memories, where information is stored as relationships between two entities connected by a predefined relation. For graph-based memory, HippoRAG [13] constructs knowledge graphs over entities, phrases, and summarization to recall more relative and comprehensive memories, while HippoRAG 2 [61] further combines original passages with phrase-based knowledge graphs to incorporate both conceptual

and contextual information. For vector memory, MemoryBank [17] adopts a dual-tower dense retrieval model, similar to Dense Passage Retrieval [90], to accurately identify relevant memories. The resulting vector representations are then indexed using FAISS [91] for efficient similarity-based retrieval.

Usage The use of personal memory can effectively empower downstream applications with personalization, enhancing the user’s individualized experience. For instance, the recalled relevant memory is used as contextual information to enhance the personalized recommendation and response capability of the conversational recommender agents [63–66], improving the personalized user experience. In addition to memory-augmented personalized dialogue and recommendation, personal memory can also be leveraged to enhance a wide range of applications, including software development [68], social-network simulation [69, 70], and financial trading [71].

To facilitate in-depth research on personal memory, a variety of memory-related benchmarks have emerged in recent years, including long-term conversational memory (MADial-Bench [74], LOCOMO [75], MSC [78]), everyday life memory (MemDaily [76]), memory-aware proactive dialogue (ChMapData [77]), multimodal dialogue memory (MMRC [79]), egocentric video understanding (Ego4D [80], EgoLife [62]), and long-context reasoning-in-a-haystack (BABI-Long [81, 82]).

3.2 Parametric Personal Memory

In addition to external non-parametric memory, a user’s personal memory can also be stored parametrically. Specifically, personal data can be used to fine-tune an LLM, embedding the memory directly into its parameters (i.e., parametric long-term memory) to create a personalized LLM. Alternatively, historical dialogues can be cached as prompts during inference (i.e., parametric short-term memory), enabling quick reuse in future interactions.

3.2.1 Memory Caching For Acceleration (Quadrant-III)

Personal parametric short-term memory typically refers to intermediate attention states produced by the LLM when processing personal data, which is usually utilized as memory caches to accelerate inference. Specifically, prompt caching [83] is usually used as an efficient data management technique that allows for the pre-storage of large amounts of personal data or information that may be frequently requested, such as a user’s conversational history. For instance, during multi-turn dialogues, the dialogue system can quickly provide the personal context information directly from the parametric memory cache, avoiding the need to recalculate or retrieve it from the original data source, saving both time and resources. Major platforms such as DeepSeek, Anthropic, OpenAI, and Google employ prompt caching to reduce API call costs and improve response speed in dialogue scenarios. Moreover, personal parametric short-term memory can enhance the performance of retrieval-augmented generation (RAG) through Contextual Retrieval [84], where prompt caching helps reduce the overhead of generating contextualized chunks. At present, research specifically targeting caching techniques for personal memory data remains limited. Instead, most existing work considers caching as a fundamental capability of system memory, particularly in the context of key-value (KV) management and KV reuse. A more detailed discussion of these aspects is provided in Section 4.

3.2.2 Personalized Knowledge Editing (Quadrant-IV)

Personal parametric long-term memory utilizes personalized Knowledge Editing technology [92], such as Parameter-Efficient Fine-Tuning (PEFT) [93], to encode personal data into the LLM’s parameters in a parametric manner, thereby facilitating the long-term, parameterized storage of memory. For instance, Character-LLM [85] enables the role-playing of specific characters, such as Beethoven, Queen Cleopatra, Julius Caesar, etc., by training large language models to remember the roles and experiences of these characters. AI-Native Memory [36] proposes using deep neural network models, specifically large language models (LLMs), as Lifelong Personal Models (LPMs) to parameterize, compress, and continuously evolve personal memory through user interactions, enabling a more comprehensive understanding of the user. MemoRAG [86] utilizes LLM parametric memory to store user conversation history and preferences, forming a personalized global memory that enhances personalization and enables tailored recommendations. Echo [87] is a large language

model enhanced with temporal episodic memory, designed to improve performance in applications requiring multi-turn, complex memory-based dialogues. The parameterization of personal long-term memory presents several challenges, notably the need to fine-tune models on individual user data, which demands substantial computational resources. This requirement significantly hinders the scalability and practical deployment of parametric approaches to long-term personal memory.

3.3 Discussion

In this section, we describe personal memory and related work from the perspectives of non-parametric and parametric approaches. Specifically, personal non-parametric short-term memory necessitates efficient mechanisms for memory encoding and management. Existing literature predominantly emphasizes the design and implementation of systems that facilitate the construction, management, retrieval, and effective utilization of a user’s personal non-parametric long-term memory. In contrast, personal parametric short-term memory can employ techniques such as prompt caching to reduce computational costs and enhance efficiency. Parametric long-term memory offers advantages in memory compression, thereby supporting a more comprehensive and global representation of the user’s accumulated experiences. Recent trends in the field indicate a growing interest in integrating both short-term and long-term memory paradigms, wherein parametric and non-parametric memory components complement and reinforce one another. The subsequent section will present a detailed discussion of system memory and its associated research developments.

4 System Memory

System memory constitutes a critical component of LLM-driven AI systems. It encompasses a sequence of intermediate representations or results generated throughout the task execution process. By leveraging system memory, LLM-driven AI systems can enhance their capabilities in reasoning, planning, and other higher-order cognitive functions. Moreover, the effective use of system memory contributes to the system’s capacity for self-evolution and continual improvement. In this section, we examine system memory and its associated research from both non-parametric and parametric perspectives.

Quadrant	Dimension	Feature	Models
V	System Non-Parametric Short-Term	Reasoning & Planning Enhancement	ReAct [24], RAP [94], Reflexion [95], Talker-Reasoner [96], TPTU [97]
VI	System Non-Parametric Long-Term	Reflection & Refinement	Buffer of Thoughts [98], AWM [99], Think-in-Memory [100], GITM [101], Voyager [102], Retroformer [103], Expel [104], Synapse [105], MetaGPT [106], Learned Memory Bank [107], M+ [108]
VII	System Parametric Short-Term	KV Management	LookupFFN [109], ChunkKV [110], vLLM [111], FastServe [112], StreamingLLM [113], Orca [114], DistServe [115], LLM.int8() [116], FastGen [117], Train Large, Then Compress [118], Scissorhands [119], H ₂ O [120], Mooncake [121], MemServe [122], SLM Serving [123], IMPRESS [124], AdaServe [125], MPIC [126], IntelLLM [127]
		KV Reuse	KV Cache [128], Prompt Cache [83], Contextual Retrieval [84], CacheGen [129], ChunkAttention [130], RAGCache [131], SGLang [132], Ada-KV [133], HCache [134], Cake [135], EPIC [136], RelayAttention [137], Marconi [138], IKS [139], FastCache [140], Cache-Craft [141], KVLink [142], RAGServe [143], BumbleBee [144]
VIII	System Parametric Long-Term	Parametric Memory Structures	Memorizing Transformer [145], Focused Transformer [146], MAC [147], MemoryLLM [148], WISE [149], LongMem [150], LM2 [151], Titans [152]

Table 3: System Memory

4.1 Contextual System Memory

From a temporal perspective, non-parametric short-term system memory refers to a series of reasoning and action results generated by large language models during task execution. This form of memory supports enhanced reasoning and planning within the context of the current task, thereby contributing to improved task accuracy, efficiency, and overall completion rates. In contrast, non-parametric long-term system memory represents a more abstracted and generalized form of short-term memory. It encompasses the consolidation of prior successful experiences and mechanisms of self-reflection based on historical interactions, which collectively facilitate the continual evolution and adaptive enhancement of LLM-driven AI systems.

4.1.1 Reasoning & Planning Enhancement (Quadrant-V)

Analogous to human cognition, the reasoning and planning processes of large language models (LLMs) give rise to a sequence of short-term intermediate outputs. These outputs may reflect task-related attempts, which can be either successful or erroneous. Regardless of their correctness, such intermediate results serve as informative and constructive references that can guide subsequent task execution. This form of system non-parametric short-term memory plays a pivotal role in LLM-driven AI systems. Empirical evidence demonstrates that leveraging this memory structure significantly enhances the reasoning and planning capabilities of LLMs. For instance, ReAct [24] integrates reasoning and action by generating intermediate reasoning steps alongside corresponding actions, enabling the model to alternate between thought and execution. This approach facilitates intelligent planning and adaptive decision-making in complex problem-solving scenarios. Similarly, Reflexion [95] introduces mechanisms for dynamic memory and self-reflection, allowing the LLM to self-evaluate and iteratively refine its behavior based on prior errors or limitations. This self-improvement loop promotes enhanced performance in future tasks, resembling a continuous learning and optimization process.

4.1.2 Reflection & Refinement (Quadrant-VI)

The development of system non-parametric long-term memory parallels the human process of learning from both successes and failures. It involves the reflection upon and refinement of accumulated short-term memory traces. This memory mechanism enables the system not only to retain and replicate effective strategies from past experiences but also to extract valuable lessons from failures, thereby minimizing the likelihood of repeated errors. Through continuous updating and optimization, the system incrementally enhances its decision-making capabilities and improves its responsiveness to novel challenges. Moreover, the progressive accumulation of long-term memory empowers the system to address increasingly complex tasks with greater adaptability and resilience. For instance, Buffer of Thoughts (BoT) [98] refines the chain of thoughts from historical tasks to form thought templates, which are then stored in a memory repository, guiding future reasoning and decision-making processes. Agent Workflow Memory (AWM) [99] introduces reusable paths, called workflows, and guides subsequent task generation by selecting different workflows. Think-in-Memory (TiM) [100] continuously generates new thoughts based on conversation history, which is more conducive to reasoning and computation compared to raw observational data. Ghost in the Minecraft (GITM) [101] uses reference plans recorded in memory, allowing the agent planner to more efficiently handle encountered tasks, thereby improving task execution success rates. Voyager [102] refines skills based on environmental feedback and stores acquired skills in memory, forming a skill library for future reuse in similar situations (e.g., fighting zombies vs. fighting spiders). Retroformer [103] leverages recent interaction trajectories as short-term memory and reflective feedback from past failures as long-term memory to guide decision-making and reasoning. ExpeL [104] enhances task resolution by drawing on contextualized successful examples and abstracting insights from both successes and failures through comparative and pattern-based analysis of past experiences.

4.2 Parametric System Memory

The parametric system memory refers to the temporary storage of knowledge information in parametric forms, such as KV Cache [128], during the inference process (short-term memory), or the long-term editing and storage of knowledge information in the model parameters (long-term memory). The former, parametric short-term system memory, corresponds to human working memory,

enabling cost reduction and efficiency improvement in large language model inference. The latter, parametric long-term system memory, corresponds to human semantic memory, facilitating the efficient integration of new knowledge.

4.2.1 KV Management & Reuse (Quadrant-VII)

Parametric short-term system memory primarily focuses on the management and reuse of attention keys (Key) and values (Value) in LLMs, aiming to address issues such as high inference costs and latency during the reasoning process. KV management optimizes memory efficiency and inference performance through techniques such as KV cache organization [111], compression [110], and quantization [116]. In particular, vLLM [111] is a high-efficiency LLM serving system built on PagedAttention, a virtual memory-inspired attention mechanism that enables near-zero KV cache waste and flexible sharing across requests, substantially improving batching efficiency and inference throughput. ChunkKV [110] is a method for compressing the key-value cache in long-context inference with LLMs by grouping tokens into semantic chunks, retaining the most informative ones, and enabling layer-wise index reuse, thereby reducing memory and computational costs while outperforming existing approaches on several benchmarks. LLM.int8() [116] is a mixed-precision quantization method that combines vector-wise Int8 quantization with selective 16-bit handling of emergent outlier features, enabling memory-efficient inference of large language models (up to 175B parameters) without performance degradation.

Meanwhile, KV reuse focuses on reusing inference-related parameters through token-level KV Cache [128] and sentence-level Prompt Cache [83], which helps reduce computational costs and improve the efficiency of large language model (LLM) usage. Specifically, KV Cache [128] stores the attention keys (Key) and values (Value) generated by the neural network during sequence generation, allowing them to be reused in subsequent inference steps. This reuse accelerates attention computation in long-text generation and reduces redundant computation. In contrast, Prompt Cache [83] operates at the sentence level by caching previous input prompts along with their corresponding output results. When similar prompts are encountered, the LLM can retrieve and return cached responses directly, saving computation and accelerating response generation. By avoiding frequent recomputation of identical or similar contexts, KV reuse enables more efficient inference and significantly reduces computational overhead. Additionally, it enhances the flexibility and responsiveness of LLM-based systems in handling continuous or interactive tasks. Building on these ideas, RAGCache [131] introduces a multilevel dynamic caching system tailored for Retrieval-Augmented Generation (RAG), which caches intermediate knowledge states, optimizes memory replacement policies based on LLM inference and retrieval patterns, and overlaps retrieval with inference to significantly reduce latency and improve throughput.

Parametric short-term system memory overlaps somewhat with the previously mentioned parametric short-term personal memory in terms of technical approach. The difference lies in their focus: parametric short-term personal memory is more concerned with improving the processing of individual input data, while parametric short-term system memory focuses on optimizing the storage and reuse of system-level context during task execution. The former primarily addresses how to quickly process and adapt to an individual’s input information, whereas the latter aims to reduce inference costs in multi-turn reasoning and enhance the consistency and efficiency of global tasks.

4.2.2 Parametric Memory Structures (Quadrant-VIII)

From the perspective of large language models (LLM) as long-term parametric memory, LLMs are not merely tools that provide immediate responses based on input and output; they can also store and integrate information over long time spans, forming an ever-evolving knowledge system. LLMs based on the Transformer [153] architecture are capable of memorizing knowledge information, primarily due to the self-attention mechanism in the Transformer-based model and the large-scale parameterized training approach. By training on vast corpora, LLMs learn extensive world knowledge, language patterns, and solutions to various tasks. Additionally, LLMs can modify, update, or refine the internal knowledge through parameterized knowledge editing, allowing for more precise task handling or responses that better align with user needs. MemoryLLM [148] has the ability to self-update and inject memory with new knowledge, effectively integrating new information and demonstrating excellent model editing performance and long-term information retention capabilities. WISE [149] is a lifelong editing framework for large language models that employs a dual-

parametric memory design, with the main memory preserving pretrained knowledge and the side memory storing edited information. It leverages a routing mechanism to dynamically access the appropriate memory during inference and uses knowledge sharding to distribute and integrate edits efficiently, ensuring reliability, generalization, and locality throughout continual updates. The core function of parameterized knowledge editing [92] is to enable large language models (LLMs) with dynamic and flexible knowledge updating capabilities, allowing them to respond to constantly changing task requirements, domain knowledge, and new information from the real world. This allows LLMs to remain efficient and accurate across various application scenarios and be customized and optimized according to user or environmental needs.

4.3 Discussion

In this section, we describe system memory and related work from the perspectives of non-parametric and parametric approaches. Non-parametric short-term system memory can enhance the reasoning and planning abilities for current tasks, while non-parametric long-term system memory enables the reuse of successful experiences and the self-reflection based on historical experience, facilitating the evolution of LLM-driven AI system capabilities. On the other hand, parametric short-term system memory can reduce costs and improve efficiency in large language model inference, and long-term parametric system memory can store and integrate information over long time spans, forming a continuously evolving knowledge system. In the next section, we will summarize the issues and challenges in memory research in the era of large language models and point out potential future directions for development.

5 Open Problems and Future Directions

Although substantial progress has been made in current memory research across the three dimensions—object, form, and time—as well as within the eight corresponding quadrants, numerous open issues and challenges remain. Building upon recent advancements and recognizing existing limitations, we outline the following promising directions for future research:

From Unimodal Memory to Multimodal Memory In the era of large language models, LLM-driven AI systems are gradually expanding from being able to process only a single type of data (such as text) to handle multiple types of data simultaneously (such as text, images, audio, video, and even sensor data). This transition enhances perceptual capabilities and enables robust performance in complex real-world tasks. For example, in the medical field, by combining text (medical records), images (medical imaging), and speech (doctor-patient conversations), AI systems can more accurately understand and diagnose medical conditions. Multimodal memory systems can integrate information from different sensory channels into a unified understanding, thereby approaching human cognitive processes more closely. Moreover, the expansion of multimodal memory also opens up possibilities for more personalized and interactive AI applications [154]. For instance, personal AI assistants can not only communicate with users through text but also interpret users’ emotions by recognizing facial expressions, voice intonations, or body language, thus providing more personalized and empathetic responses.

From Static Memory to Stream Memory Static memory can be viewed as a batch-processing approach to memory storage. It accumulates information or experiences in discrete batches, typically processing, storing, and retrieving them at specific intervals or predetermined points in time. As an offline memory model, static memory emphasizes the systematic organization and consolidation of large volumes of information, making it well-suited for long-term knowledge retention and structured learning. In contrast, stream memory operates in a continuous, real-time manner. Analogous to data stream processing, it handles information as it arrives, prioritizing immediacy and adaptability. As an online or real-time memory model, stream memory focuses on the dynamic updating of information and rapid responsiveness to evolving contexts. These two memory paradigms are not mutually exclusive and often function complementarily: while static memory supports the accumulation of stable, long-term knowledge, stream memory enables agile adaptation to ongoing tasks and real-time information demands.

From Specific Memory to Comprehensive Memory The human memory system comprises multiple interconnected subsystems—such as sensory memory, working memory, explicit memory, and implicit memory—each fulfilling distinct functions and contributing to the overall cognitive process. In the context of large language models (LLMs), current memory architectures often concentrate on narrow or task-specific components, such as short-term memory for immediate inference or domain-specific knowledge storage. While such targeted memory mechanisms can enhance performance in specific scenarios, their limited scope constrains the system’s overall flexibility, generalization, and adaptability. Looking forward, the development of comprehensive and collaborative memory systems is essential. These systems should integrate diverse memory types and support efficient interaction, self-organization, and continual updating, enabling LLMs to manage increasingly complex and dynamic tasks. By more closely emulating the multi-layered, multi-dimensional, and adaptive characteristics of human memory, such architectures have the potential to significantly advance the general intelligence and autonomy of LLM-based AI systems.

From Exclusive Memory to Shared Memory At present, the memory of each LLM-driven AI system operates independently, typically confined to a specific domain and tailored to processing isolated tasks or environments. However, as AI technologies continue to evolve, memory systems are expected to become increasingly interconnected, transcending domain boundaries and enabling enhanced collaboration among models. For instance, a large language model specialized in the medical domain could share its memory or knowledge base with another model focused on finance, facilitating cross-domain knowledge transfer and cooperative task solving. Such a shared memory paradigm would not only improve the efficiency and adaptability of individual systems but also empower multiple LLMs to dynamically access and leverage one another’s domain-specific expertise. This shift toward collaborative memory architectures could give rise to a more intelligent, resource-efficient network of AI systems capable of addressing complex, multi-domain challenges. Ultimately, shared memory is poised to broaden the scope of AI applications and accelerate its integration into increasingly diverse and demanding real-world scenarios.

From Individual Privacy to Collective Privacy With the increasing prevalence of data sharing in the AI era, the focus of privacy protection is gradually shifting from the traditional notion of individual privacy to the broader and emerging concept of collective privacy. Conventional privacy frameworks primarily aim to safeguard personal data, preventing unauthorized access, leakage, or misuse of individually identifiable information. However, in the context of large language models, individual data is often aggregated into group-level datasets for large-scale analysis and prediction. Collective privacy concerns the protection of the rights and interests of groups or communities whose data is used collectively, raising questions about how to prevent misuse, profiling, or excessive surveillance at the group level. As memory systems in AI become more advanced and interconnected, ensuring collective privacy will emerge as a critical challenge. Addressing this issue will require innovative techniques that can effectively balance the trade-off between data utility and privacy preservation [155].

From Rule-Based Evolution to Automated Evolution Traditional AI systems evolve by reflecting on past experiences—such as reusing successful strategies—based on accumulated knowledge and historical data. However, this evolutionary process often depends on manually crafted rules and heuristic adjustments to enable such self-reflection. While rule-based evolution can be effective, it inherently limits the system’s flexibility, scalability, and efficiency, with the quality and generalizability of the rules directly constraining the system’s adaptive capabilities. Looking ahead, AI systems are expected to achieve automated evolution, dynamically adjusting and optimizing themselves by leveraging both personal and system-level memories in response to changing data and environmental contexts. Such systems will be capable of autonomously identifying performance bottlenecks and initiating self-improvement without relying on explicit, human-defined rules. This transition toward self-directed adaptation will significantly enhance system responsiveness, reduce the need for human intervention, and enable a more intelligent, dynamic, and continuously self-evolving paradigm.

6 Conclusion

Memory plays a pivotal role in the advancement of AI systems in the era of large language models (LLMs). It not only shapes the degree of personalization in AI behavior but also influences key capabilities such as adaptability, reasoning, planning, and self-evolution. This article systematically examines the relationship between human memory and memory mechanisms in LLM-driven AI systems, exploring how principles of human cognition can inspire the design of more efficient and flexible memory architectures. We begin by analyzing various categories of human memory—including perceptual memory, working memory, and long-term memory—and compare them with existing memory models in AI. Building upon this, we propose an eight-quadrant classification framework grounded in three dimensions: object, form, and time, offering a theoretical foundation for the construction of multi-level and comprehensive memory systems. Furthermore, we review the current state of memory development in AI from both personal memory and system memory perspectives. Finally, we identify key open challenges in contemporary AI memory design and outline promising directions for future research in the LLM era. We believe that, with continued technological progress, AI systems will increasingly adopt more dynamic, adaptive, and intelligent memory architectures, thereby enabling more robust applications across complex, real-world tasks.

References

- [1] Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):186345, 2024.
- [2] Yuanchun Li, Hao Wen, Weijun Wang, Xiangyu Li, Yizhen Yuan, Guohong Liu, Jiacheng Liu, Wenxing Xu, Xiang Wang, Yi Sun, et al. Personal llm agents: Insights and survey about the capability, efficiency and security. *arXiv preprint arXiv:2401.05459*, 2024.
- [3] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 1(2), 2023.
- [4] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology*, 15(3):1–45, 2024.
- [5] Bo Chen, Xinyi Dai, Huifeng Guo, Wei Guo, Weiwen Liu, Yong Liu, Jiarui Qin, Ruiming Tang, Yichao Wang, Chuhan Wu, et al. All roads lead to rome: Unveiling the trajectory of recommender systems across the llm era. *arXiv preprint arXiv:2407.10081*, 2024.
- [6] Chen Zhang, Xinyi Dai, Yaxiong Wu, Qu Yang, Yasheng Wang, Ruiming Tang, and Yong Liu. A survey on multi-turn interaction capabilities of large language models. *arXiv preprint arXiv:2501.09959*, 2025.
- [7] Zeyu Zhang, Xiaohe Bo, Chen Ma, Rui Li, Xu Chen, Quanyu Dai, Jieming Zhu, Zhenhua Dong, and Ji-Rong Wen. A survey on the memory mechanism of large language model based agents. *arXiv preprint arXiv:2404.13501*, 2024.
- [8] Xun Jiang, Feng Li, Han Zhao, Jiaying Wang, Jun Shao, Shihao Xu, Shu Zhang, Weiling Chen, Xavier Tang, Yize Chen, et al. Long term memory: The foundation of ai self-evolution. *arXiv preprint arXiv:2410.15665*, 2024.
- [9] Lauralee Sherwood, Robert Thomas Kell, and Christopher Ward. *Human physiology: from cells to systems*. Thomson/Brooks/Cole, 2004.
- [10] Lilian Weng. Llm-powered autonomous agents. *lilianweng.github.io*, Jun 2023.
- [11] Andrew E Budson and Elizabeth A Kensinger. *Why we forget and how to remember better: the science behind memory*. Oxford University Press, 2023.
- [12] Alan Baddeley. *Working memory, thought, and action*, volume 45. OuP Oxford, 2007.

- [13] Bernal Jiménez Gutiérrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. Hipporag: Neurobiologically inspired long-term memory for large language models. *arXiv preprint arXiv:2405.14831*, 2024.
- [14] George Applegarth, Christian Weatherstone, Maximilian Hollingsworth, Henry Middlebrook, and Marcus Irvin. Exploring synaptic resonance in large language models: A novel approach to contextual memory integration. *arXiv preprint arXiv:2502.10699*, 2025.
- [15] Samuel J Gershman, Ila Fiete, and Kazuki Irie. Key-value memory in the brain. *arXiv preprint arXiv:2501.02950*, 2025.
- [16] Bang Liu, Xinfeng Li, Jiayi Zhang, Jinlin Wang, Tanjin He, Sirui Hong, Hongzhang Liu, Shaokun Zhang, Kaitao Song, Kunlun Zhu, Yuheng Cheng, Suyuchen Wang, Xiaoqiang Wang, Yuyu Luo, Haibo Jin, Peiyan Zhang, Ollie Liu, Jiaqi Chen, Huan Zhang, Zhaoyang Yu, Haochen Shi, Boyan Li, Dekun Wu, Fengwei Teng, Xiaojun Jia, Jiawei Xu, Jinyu Xi-ang, Yizhang Lin, Tianming Liu, Tongliang Liu, Yu Su, Huan Sun, Glen Berseth, Jianyun Nie, Ian Foster, Logan Ward, Qingyun Wu, Yu Gu, Mingchen Zhuge, Xiangru Tang, Haohan Wang, Jiaxuan You, Chi Wang, Jian Pei, Qiang Yang, Xiaoliang Qi, and Chenglin Wu. Advances and challenges in foundation agents: From brain-inspired intelligence to evolutionary, collaborative, and safe systems, 2025.
- [17] Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. Memorybank: Enhancing large language models with long-term memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19724–19731, 2024.
- [18] OpenAI. Memory and new controls for chatgpt. *openai.com*, February 2024.
- [19] Apple. Introducing apple intelligence, the personal intelligence system that puts powerful generative models at the core of iphone, ipad, and mac. *apple.com*, June 2024.
- [20] mem0ai. mem0: The memory layer for personalized ai. *mem0.ai*, July 2024.
- [21] ModelScope. Memoryscope: Equip your llm chatbot with a powerful and flexible long term memory system. *github.com*, September 2024.
- [22] Richard C Atkinson and Richard M Shiffrin. Human memory: A proposed system and its control processes. In *Psychology of learning and motivation*, volume 2, pages 89–195. Elsevier, 1968.
- [23] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.
- [24] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022.
- [25] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474, 2020.
- [26] OpenAI. Introducing chatgpt. *openai.com*, November 2022.
- [27] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.
- [28] Anthropic. Introducing claude. *anthropic.com*, March 2023.
- [29] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

- [30] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.
- [31] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.
- [32] Fei Mi, Yitong Li, Yulong Zeng, Jingyan Zhou, Yasheng Wang, Chuanfei Xu, Lifeng Shang, Xin Jiang, Shiqi Zhao, and Qun Liu. Pangu-bot: Efficient generative dialogue pre-training from pre-trained language model. *arXiv preprint arXiv:2203.17090*, 2022.
- [33] Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, et al. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*, 2024.
- [34] Andreas Köpf, Yannic Kilcher, Dimitri Von Rütte, Sotiris Anagnostidis, Zhi Rui Tam, Keith Stevens, Abdullah Barhoum, Duc Nguyen, Oliver Stanley, Richárd Nagyfi, et al. Openassistant conversations-democratizing large language model alignment. *Advances in Neural Information Processing Systems*, 36:47669–47681, 2023.
- [35] Microsoft. Recall overview. *microsoft.com*, February 2025.
- [36] Jingbo Shang, Zai Zheng, Jiale Wei, Xiang Ying, Felix Tao, and Mindverse Team. Ai-native memory: A pathway from llms towards agi. *arXiv preprint arXiv:2406.18312*, 2024.
- [37] Memory. Beyond short-term memory: How memory makes chatbots remember. *github.com*, April 2024.
- [38] langchain ai. Langgraph memory service. *github.com*, October 2024.
- [39] GoodAI. Charlie mnemonic. *github.com*, March 2024.
- [40] memodb io. Memobase: User profile-based memory for genai apps. *memobase.io*, January 2025.
- [41] Letta-AI. Letta. *github.com*, September 2024.
- [42] Cognee.ai. Cognee. *github.com*, October 2024.
- [43] Gibbeum Lee, Volker Hartmann, Jongho Park, Dimitris Papailiopoulos, and Kangwook Lee. Prompted llms as chatbot modules for long open-domain conversation. *arXiv preprint arXiv:2305.04533*, 2023.
- [44] Ali Modarressi, Ayyoob Imani, Mohsen Fayyaz, and Hinrich Schütze. Ret-llm: Towards a general read-write memory for large language models. *arXiv preprint arXiv:2305.14322*, 2023.
- [45] Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G Patil, Ion Stoica, and Joseph E Gonzalez. Memgpt: Towards llms as operating systems. *arXiv preprint arXiv:2310.08560*, 2023.
- [46] Jingwei Sun, Zhixu Du, and Yiran Chen. Knowledge graph tuning: Real-time large language model personalization based on human feedback. *arXiv preprint arXiv:2405.19686*, 2024.
- [47] Ruifeng Yuan, Shichao Sun, Yongqi Li, Zili Wang, Ziqiang Cao, and Wenjie Li. Personalized large language model assistant with evolving conditional memory. *arXiv preprint arXiv:2312.17257*, 2023.
- [48] Zhuoshi Pan, Qianhui Wu, Huiqiang Jiang, Xufang Luo, Hao Cheng, Dongsheng Li, Yuqing Yang, Chin-Yew Lin, H Vicky Zhao, Lili Qiu, et al. On memory construction and retrieval for personalized conversational agents. *arXiv preprint arXiv:2502.05589*, 2025.

- [49] Hongkang Yang, Zehao Lin, Wenjin Wang, Hao Wu, Zhiyu Li, Bo Tang, Wenqiang Wei, Jinbo Wang, Zeyun Tang, Shichao Song, et al. Memory3: Language modeling with explicit memory. *arXiv preprint arXiv:2407.01178*, 2024.
- [50] Rana Salama, Jason Cai, Michelle Yuan, Anna Currey, Monica Sunkara, Yi Zhang, and Yassine Benajiba. Meminsight: Autonomous memory augmentation for llm agents. *arXiv preprint arXiv:2503.21760*, 2025.
- [51] Junru Lu, Siyu An, Mingbao Lin, Gabriele Pergola, Yulan He, Di Yin, Xing Sun, and Yunsheng Wu. Memochat: Tuning llms to use memos for consistent long-range open-domain conversation. *arXiv preprint arXiv:2308.08239*, 2023.
- [52] Zhen Tan, Jun Yan, I Hsu, Rujun Han, Zifeng Wang, Long T Le, Yiwen Song, Yanfei Chen, Hamid Palangi, George Lee, et al. In prospect and retrospect: Reflective memory management for long-term personalized dialogue agents. *arXiv preprint arXiv:2503.08026*, 2025.
- [53] Hao Li, Chenghao Yang, An Zhang, Yang Deng, Xiang Wang, and Tat-Seng Chua. Hello again! llm-powered personalized agent for long-term dialogue. *arXiv preprint arXiv:2406.05925*, 2024.
- [54] Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. A-mem: Agentic memory for llm agents. *arXiv preprint arXiv:2502.12110*, 2025.
- [55] Joon Sung Park, Joseph O’Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22, 2023.
- [56] Zheng Wang, Zhongyang Li, Zeren Jiang, Dandan Tu, and Wei Shi. Crafting personalized agents through retrieval-augmented generation on editable memory graphs. *arXiv preprint arXiv:2409.19401*, 2024.
- [57] Qingyue Wang, Liang Ding, Yanan Cao, Zhiliang Tian, Shi Wang, Dacheng Tao, and Li Guo. Recursively summarizing enables long-term dialogue memory in large language models. *arXiv preprint arXiv:2308.15022*, 2023.
- [58] Nuo Chen, Hongguang Li, Juhua Huang, Baoyuan Wang, and Jia Li. Compress to impress: Unleashing the potential of compressive memory in real-world long-term conversations. *arXiv preprint arXiv:2402.11975*, 2024.
- [59] Chenxu Hu, Jie Fu, Chenzhuang Du, Simian Luo, Junbo Zhao, and Hang Zhao. Chatdb: Augmenting llms with databases as their symbolic memory. *arXiv preprint arXiv:2306.03901*, 2023.
- [60] Yuki Hou, Haruki Tamoto, and Homei Miyashita. "my agent understands me better": Integrating dynamic human-like memory recall and consolidation in llm-based agents. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–7, 2024.
- [61] Bernal Jiménez Gutiérrez, Yiheng Shu, Weijian Qi, Sizhe Zhou, and Yu Su. From rag to memory: Non-parametric continual learning for large language models. *arXiv preprint arXiv:2502.14802*, 2025.
- [62] Jingkang Yang, Shuai Liu, Hongming Guo, Yuhao Dong, Xiamengwei Zhang, Sicheng Zhang, Pengyun Wang, Zitang Zhou, Binzhu Xie, Ziyue Wang, et al. Egolife: Towards egocentric life assistant. *arXiv preprint arXiv:2503.03803*, 2025.
- [63] Yunjia Xi, Weiwen Liu, Jianghao Lin, Bo Chen, Ruiming Tang, Weinan Zhang, and Yong Yu. Memocrs: Memory-enhanced sequential conversational recommender systems with large language models. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 2585–2595, 2024.

- [64] Yancheng Wang, Ziyang Jiang, Zheng Chen, Fan Yang, Yingxue Zhou, Eunah Cho, Xing Fan, Xiaojiang Huang, Yanbin Lu, and Yingzhen Yang. Recmind: Large language model powered agent for recommendation. *arXiv preprint arXiv:2308.14296*, 2023.
- [65] Lei Wang, Jingsen Zhang, Xu Chen, Yankai Lin, Ruihua Song, Wayne Xin Zhao, and Ji-Rong Wen. Recagent: A novel simulation paradigm for recommender systems. *arXiv preprint arXiv:2306.02552*, 2023.
- [66] Xu Huang, Jianxun Lian, Yuxuan Lei, Jing Yao, Defu Lian, and Xing Xie. Recommender ai agent: Integrating large language models for interactive recommendations. *arXiv preprint arXiv:2308.16505*, 2023.
- [67] Bing Wang, Xinnian Liang, Jian Yang, Hui Huang, Shuangzhi Wu, Peihao Wu, Lu Lu, Zejun Ma, and Zhoujun Li. Enhancing large language model with self-controlled memory framework. *arXiv preprint arXiv:2304.13343*, 2023.
- [68] Chen Qian, Wei Liu, Hongzhang Liu, Nuo Chen, Yufan Dang, Jiahao Li, Cheng Yang, Weize Chen, Yusheng Su, Xin Cong, et al. Chatdev: Communicative agents for software development, 2024. URL <https://arxiv.org/abs/2307.7924>, 2024.
- [69] Yuan Li, Yixuan Zhang, and Lichao Sun. Metaagents: Simulating interactions of human behaviors for llm-based task-oriented coordination via collaborative generative agents. *arXiv preprint arXiv:2310.06500*, 2023.
- [70] Chen Gao, Xiaochong Lan, Zhihong Lu, Jinzhu Mao, Jinghua Piao, Huandong Wang, Depeng Jin, and Yong Li. S³: Social-network simulation system with large language model-empowered agents. *arXiv preprint arXiv:2307.14984*, 2023.
- [71] Yang Li, Yangyang Yu, Haohang Li, Zhi Chen, and Khaldoun Khashanah. Tradinggpt: Multi-agent system with layered memory and distinct characters for enhanced financial trading performance. *arXiv preprint arXiv:2309.03736*, 2023.
- [72] Ryan Yen and Jian Zhao. Memolet: Reifying the reuse of user-ai conversational memories. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*, pages 1–22, 2024.
- [73] Ching-Yun Ko, Sihui Dai, Payel Das, Georgios Kollias, Subhajit Chaudhury, and Aurelie Lozano. Memreasoner: A memory-augmented llm architecture for multi-hop reasoning. In *The First Workshop on System-2 Reasoning at Scale, NeurIPS’24*, 2024.
- [74] Junqing He, Liang Zhu, Rui Wang, Xi Wang, Reza Haffari, and Jiaxing Zhang. Madial-bench: Towards real-world evaluation of memory-augmented dialogue generation. *arXiv preprint arXiv:2409.15240*, 2024.
- [75] Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. Evaluating very long-term conversational memory of llm agents. *arXiv preprint arXiv:2402.17753*, 2024.
- [76] Zeyu Zhang, Quanyu Dai, Luyu Chen, Zeren Jiang, Rui Li, Jieming Zhu, Xu Chen, Yi Xie, Zhenhua Dong, and Ji-Rong Wen. Memsim: A bayesian simulator for evaluating memory of llm-based personal assistants. *arXiv preprint arXiv:2409.20163*, 2024.
- [77] Bowen Wu, Wenqing Wang, Haoran Li, Ying Li, Jingsong Yu, and Baoxun Wang. Interpersonal memory matters: A new task for proactive dialogue utilizing conversational history. *arXiv preprint arXiv:2503.05150*, 2025.
- [78] Jing Xu, Arthur Szlam, and Jason Weston. Beyond goldfish memory: Long-term open-domain conversation. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5180–5197, Dublin, Ireland, May 2022. Association for Computational Linguistics.

- [79] Haochen Xue, Feilong Tang, Ming Hu, Yexin Liu, Qidong Huang, Yulong Li, Chengzhi Liu, Zhongxing Xu, Chong Zhang, Chun-Mei Feng, et al. Mmrc: A large-scale benchmark for understanding multimodal large language model in real-world conversation. *arXiv preprint arXiv:2502.11903*, 2025.
- [80] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18995–19012, 2022.
- [81] Yuri Kuratov, Aydar Bulatov, Petr Anokhin, Ivan Rodkin, Dmitry Sorokin, Artyom Sorokin, and Mikhail Burtsev. Babilong: Testing the limits of llms with long context reasoning-in-a-haystack, 2024.
- [82] Yuri Kuratov, Aydar Bulatov, Petr Anokhin, Dmitry Sorokin, Artyom Sorokin, and Mikhail Burtsev. In search of needles in a 10m haystack: Recurrent memory finds what llms miss, 2024.
- [83] In Gim, Guojun Chen, Seung-seob Lee, Nikhil Sarda, Anurag Khandelwal, and Lin Zhong. Prompt cache: Modular attention reuse for low-latency inference. *Proceedings of Machine Learning and Systems*, 6:325–338, 2024.
- [84] Anthropic. Introducing contextual retrieval. *anthropic.com*, September 2024.
- [85] Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. Character-llm: A trainable agent for role-playing. *arXiv preprint arXiv:2310.10158*, 2023.
- [86] Hongjin Qian, Peitian Zhang, Zheng Liu, Kelong Mao, and Zhicheng Dou. Memorag: Moving towards next-gen rag via memory-inspired knowledge discovery. *arXiv preprint arXiv:2409.05591*, 2024.
- [87] WenTao Liu, Ruohua Zhang, Aimin Zhou, Feng Gao, and JiaLi Liu. Echo: A large language model with temporal episodic memory. *arXiv preprint arXiv:2502.16090*, 2025.
- [88] David Kadavy. *Digital Zettelkasten: Principles, Methods, & Examples*. Kadavy, Inc., 2021.
- [89] Jaap MJ Murre and Joeri Dros. Replication and analysis of ebbinghaus’ forgetting curve. *PLoS one*, 10(7):e0120644, 2015.
- [90] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick SH Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *EMNLP (1)*, pages 6769–6781, 2020.
- [91] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547, 2019.
- [92] Song Wang, Yaochen Zhu, Haochen Liu, Zaiyi Zheng, Chen Chen, and Jundong Li. Knowledge editing for large language models: A survey. *ACM Computing Surveys*, 57(3):1–37, 2024.
- [93] Zeyu Han, Chao Gao, Jinyang Liu, Jeff Zhang, and Sai Qian Zhang. Parameter-efficient fine-tuning for large models: A comprehensive survey. *arXiv preprint arXiv:2403.14608*, 2024.
- [94] Shibo Hao, Yi Gu, Haodi Ma, Joshua Jiahua Hong, Zhen Wang, Daisy Zhe Wang, and Zhit-ing Hu. Reasoning with language model is planning with world model. *arXiv preprint arXiv:2305.14992*, 2023.
- [95] Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36, 2024.
- [96] Konstantina Christakopoulou, Shibl Mourad, and Maja Matarić. Agents thinking fast and slow: A talker-reasoner architecture. *arXiv preprint arXiv:2410.08328*, 2024.

- [97] Jingqing Ruan, Yihong Chen, Bin Zhang, Zhiwei Xu, Tianpeng Bao, Guoqing Du, Shiwei Shi, Hangyu Mao, Ziyue Li, Xingyu Zeng, et al. Tptu: large language model-based ai agents for task planning and tool usage. *arXiv preprint arXiv:2308.03427*, 2023.
- [98] Ling Yang, Zhaochen Yu, Tianjun Zhang, Shiyi Cao, Minkai Xu, Wentao Zhang, Joseph E Gonzalez, and Bin Cui. Buffer of thoughts: Thought-augmented reasoning with large language models. *arXiv preprint arXiv:2406.04271*, 2024.
- [99] Zora Zhiruo Wang, Jiayuan Mao, Daniel Fried, and Graham Neubig. Agent workflow memory. *arXiv preprint arXiv:2409.07429*, 2024.
- [100] Lei Liu, Xiaoyan Yang, Yue Shen, Binbin Hu, Zhiqiang Zhang, Jinjie Gu, and Guannan Zhang. Think-in-memory: Recalling and post-thinking enable llms with long-term memory. *arXiv preprint arXiv:2311.08719*, 2023.
- [101] Xizhou Zhu, Yuntao Chen, Hao Tian, Chenxin Tao, Weijie Su, Chenyu Yang, Gao Huang, Bin Li, Lewei Lu, Xiaogang Wang, et al. Ghost in the minecraft: Generally capable agents for open-world environments via large language models with text-based knowledge and memory. *arXiv preprint arXiv:2305.17144*, 2023.
- [102] Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*, 2023.
- [103] Weiran Yao, Shelby Heinecke, Juan Carlos Niebles, Zhiwei Liu, Yihao Feng, Le Xue, Rithesh Murthy, Zeyuan Chen, Jianguo Zhang, Devansh Arpit, et al. Retroformer: Retrospective large language agents with policy gradient optimization. *arXiv preprint arXiv:2308.02151*, 2023.
- [104] Andrew Zhao, Daniel Huang, Quentin Xu, Matthieu Lin, Yong-Jin Liu, and Gao Huang. Expel: Llm agents are experiential learners. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19632–19642, 2024.
- [105] Longtao Zheng, Rundong Wang, Xinrun Wang, and Bo An. Synapse: Trajectory-as-exemplar prompting with memory for computer control. In *The Twelfth International Conference on Learning Representations*, 2023.
- [106] Sirui Hong, Xiawu Zheng, Jonathan Chen, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, et al. Metagpt: Meta programming for multi-agent collaborative framework. *arXiv preprint arXiv:2308.00352*, 2023.
- [107] Julie Michelman, Nasrin Baratalipour, and Matthew Abueg. Enhancing reasoning with collaboration and memory. *arXiv preprint arXiv:2503.05944*, 2025.
- [108] Yu Wang, Dmitry Krotov, Yuanzhe Hu, Yifan Gao, Wangchunshu Zhou, Julian McAuley, Dan Gutfreund, Rogerio Feris, and Zexue He. M+: Extending memoryllm with scalable long-term memory. *arXiv preprint arXiv:2502.00592*, 2025.
- [109] Zhanpeng Zeng, Michael Davies, Pranav Pulijala, Karthikeyan Sankaralingam, and Vikas Singh. Lookupffn: making transformers compute-lite for cpu inference. In *International Conference on Machine Learning*, pages 40707–40718. PMLR, 2023.
- [110] Xiang Liu, Zhenheng Tang, Peijie Dong, Zeyu Li, Bo Li, Xuming Hu, and Xiaowen Chu. Chunkkv: Semantic-preserving kv cache compression for efficient long-context llm inference. *arXiv preprint arXiv:2502.00299*, 2025.
- [111] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pages 611–626, 2023.
- [112] Bingyang Wu, Yinmin Zhong, Zili Zhang, Shengyu Liu, Fangyue Liu, Yuanhang Sun, Gang Huang, Xuanzhe Liu, and Xin Jin. Fast distributed inference serving for large language models. *arXiv preprint arXiv:2305.05920*, 2023.

- [113] Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. *arXiv preprint arXiv:2309.17453*, 2023.
- [114] Gyeong-In Yu, Joo Seong Jeong, Geon-Woo Kim, Soojeong Kim, and Byung-Gon Chun. Orca: A distributed serving system for {Transformer-Based} generative models. In *16th USENIX Symposium on Operating Systems Design and Implementation (OSDI 22)*, pages 521–538, 2022.
- [115] Yinmin Zhong, Shengyu Liu, Junda Chen, Jianbo Hu, Yibo Zhu, Xuanzhe Liu, Xin Jin, and Hao Zhang. {DistServe}: Disaggregating prefill and decoding for goodput-optimized large language model serving. In *18th USENIX Symposium on Operating Systems Design and Implementation (OSDI 24)*, pages 193–210, 2024.
- [116] Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. Gpt3. int8 (): 8-bit matrix multiplication for transformers at scale. *Advances in neural information processing systems*, 35:30318–30332, 2022.
- [117] Suyu Ge, Yunan Zhang, Liyuan Liu, Minjia Zhang, Jiawei Han, and Jianfeng Gao. Model tells you what to discard: Adaptive kv cache compression for llms. *arXiv preprint arXiv:2310.01801*, 2023.
- [118] Zhuohan Li, Eric Wallace, Sheng Shen, Kevin Lin, Kurt Keutzer, Dan Klein, and Joey Gonzalez. Train big, then compress: Rethinking model size for efficient training and inference of transformers. In *International Conference on machine learning*, pages 5958–5968. PMLR, 2020.
- [119] Zichang Liu, Aditya Desai, Fangshuo Liao, Weitao Wang, Victor Xie, Zhaozhuo Xu, Anastasios Kyrillidis, and Anshumali Shrivastava. Scissorhands: Exploiting the persistence of importance hypothesis for llm kv cache compression at test time. *Advances in Neural Information Processing Systems*, 36:52342–52364, 2023.
- [120] Zhenyu Zhang, Ying Sheng, Tianyi Zhou, Tianlong Chen, Lianmin Zheng, Ruisi Cai, Zhao Song, Yuandong Tian, Christopher Ré, Clark Barrett, et al. H2o: Heavy-hitter oracle for efficient generative inference of large language models. *Advances in Neural Information Processing Systems*, 36:34661–34710, 2023.
- [121] Ruoyu Qin, Zheming Li, Weiran He, Mingxing Zhang, Yongwei Wu, Weimin Zheng, and Xinran Xu. Mooncake: A kvcache-centric disaggregated architecture for llm serving. *arXiv preprint arXiv:2407.00079*, 2024.
- [122] Cunchen Hu, Heyang Huang, Junhao Hu, Jiang Xu, Xusheng Chen, Tao Xie, Chenxi Wang, Sa Wang, Yungang Bao, Ninghui Sun, et al. Memserve: Context caching for disaggregated llm serving with elastic memory pool. *arXiv preprint arXiv:2406.17565*, 2024.
- [123] Pol G Recasens, Yue Zhu, Chen Wang, Eun Kyung Lee, Olivier Tardieu, Alaa Youssef, Jordi Torres, and Josep Ll Berral. Towards pareto optimal throughput in small language model serving. In *Proceedings of the 4th Workshop on Machine Learning and Systems*, pages 144–152, 2024.
- [124] Weijian Chen, Shuibing He, Haoyang Qu, Ruidong Zhang, Siling Yang, Ping Chen, Yi Zheng, Baoxing Huai, and Gang Chen. {IMPRESS}: An {Importance-Informed}{Multi-Tier} prefix {KV} storage system for large language model inference. In *23rd USENIX Conference on File and Storage Technologies (FAST 25)*, pages 187–201, 2025.
- [125] Zikun Li, Zhuofu Chen, Remi Delacourt, Gabriele Oliaro, Zeyu Wang, Qinghan Chen, Shuhuai Lin, April Yang, Zhihao Zhang, Zhuoming Chen, et al. Adaserve: Slo-customized llm serving with fine-grained speculative decoding. *arXiv preprint arXiv:2501.12162*, 2025.
- [126] Shiju Zhao, Junhao Hu, Rongxiao Huang, Jiaqi Zheng, and Guihai Chen. Mpic: Position-independent multimodal context caching system for efficient mllm serving. *arXiv preprint arXiv:2502.01960*, 2025.
- [127] TingLong Li and Qiuyu Shao. IntelLLM: Little hints make a big difference for LLM KV cache compression, 2024.

- [128] Reiner Pope, Sholto Douglas, Aakanksha Chowdhery, Jacob Devlin, James Bradbury, Jonathan Heek, Kefan Xiao, Shivani Agrawal, and Jeff Dean. Efficiently scaling transformer inference. *Proceedings of Machine Learning and Systems*, 5:606–624, 2023.
- [129] Yuhao Liu, Hanchen Li, Kuntai Du, Jiayi Yao, Yihua Cheng, Yuyang Huang, Shan Lu, Michael Maire, Henry Hoffmann, Ari Holtzman, et al. Cachegen: Fast context loading for language model applications. *CoRR*, 2023.
- [130] Lu Ye, Ze Tao, Yong Huang, and Yang Li. Chunkattention: Efficient self-attention with prefix-aware kv cache and two-phase partition. *arXiv preprint arXiv:2402.15220*, 2024.
- [131] Chao Jin, Zili Zhang, Xuanlin Jiang, Fangyue Liu, Xin Liu, Xuanzhe Liu, and Xin Jin. Ragcache: Efficient knowledge caching for retrieval-augmented generation. *arXiv preprint arXiv:2404.12457*, 2024.
- [132] Lianmin Zheng, Liangsheng Yin, Zhiqiang Xie, Chuyue Sun, Jeff Huang, Cody Hao Yu, Shiyi Cao, Christos Kozyrakis, Ion Stoica, Joseph E. Gonzalez, Clark Barrett, and Ying Sheng. Sglang: Efficient execution of structured language model programs, 2024.
- [133] Yuan Feng, Junlin Lv, Yukun Cao, Xike Xie, and S Kevin Zhou. Ada-kv: Optimizing kv cache eviction by adaptive budget allocation for efficient llm inference. *arXiv preprint arXiv:2407.11550*, 2024.
- [134] Shiwei Gao, Youmin Chen, and Jiwu Shu. Fast state restoration in llm serving with hcache. *arXiv preprint arXiv:2410.05004*, 2024.
- [135] Shuwei Jin, Xueshen Liu, Qingzhao Zhang, and Z Morley Mao. Compute or load kv cache? why not both? *arXiv preprint arXiv:2410.03065*, 2024.
- [136] Junhao Hu, Wenrui Huang, Haoyi Wang, Weidong Wang, Tiancheng Hu, Qin Zhang, Hao Feng, Xusheng Chen, Yizhou Shan, and Tao Xie. Epic: Efficient position-independent context caching for serving large language models. *arXiv preprint arXiv:2410.15332*, 2024.
- [137] Lei Zhu, Xinjiang Wang, Wayne Zhang, and Rynson WH Lau. Relayattention for efficient large language model serving with long system prompts. *arXiv preprint arXiv:2402.14808*, 2024.
- [138] Rui Pan, Zhuang Wang, Zhen Jia, Can Karakus, Luca Zancato, Tri Dao, Yida Wang, and Ravi Netravali. Marconi: Prefix caching for the era of hybrid llms. *arXiv preprint arXiv:2411.19379*, 2024.
- [139] Derrick Quinn, Mohammad Nouri, Neel Patel, John Salihu, Alireza Salemi, Sukhan Lee, Hamed Zamani, and Mohammad Alian. Accelerating retrieval-augmented generation. *arXiv preprint arXiv:2412.15246*, 2024.
- [140] Jianian Zhu, Hang Wu, Haojie Wang, Yinghui Li, Biao Hou, Ruixuan Li, and Jidong Zhai. Fastcache: Optimizing multimodal llm serving through lightweight kv-cache compression framework. *arXiv preprint arXiv:2503.08461*, 2025.
- [141] Shubham Agarwal, Sai Sundaresan, Subrata Mitra, Debabrata Mahapatra, Archit Gupta, Rounak Sharma, Nirmal Joshua Kapu, Tong Yu, and Shiv Saini. Cache-craft: Managing chunk-caches for efficient retrieval-augmented generation. *arXiv preprint arXiv:2502.15734*, 2025.
- [142] Jingbo Yang, Bairu Hou, Wei Wei, Yujia Bao, and Shiyu Chang. Kvlink: Accelerating large language models via efficient kv cache reuse. *arXiv preprint arXiv:2502.16002*, 2025.
- [143] Siddhant Ray, Rui Pan, Zhuohan Gu, Kuntai Du, Ganesh Ananthanarayanan, Ravi Netravali, and Junchen Jiang. Ragserve: Fast quality-aware rag systems with configuration adaptation. *arXiv preprint arXiv:2412.10543*, 2024.
- [144] Lilly Kumari, Shengjie Wang, Tianyi Zhou, Nikhil Sarda, Anthony Rowe, and Jeff Bilmes. Bumblebee: Dynamic kv-cache streaming submodular summarization for infinite-context transformers. In *First Conference on Language Modeling*, 2024.

- [145] Yuhuai Wu, Markus N Rabe, DeLesley Hutchins, and Christian Szegedy. Memorizing transformers. *arXiv preprint arXiv:2203.08913*, 2022.
- [146] Szymon Tworkowski, Konrad Staniszewski, Mikołaj Patek, Yuhuai Wu, Henryk Michalewski, and Piotr Miłoś. Focused transformer: Contrastive training for context scaling. *Advances in Neural Information Processing Systems*, 36, 2024.
- [147] Jihoon Tack, Jaehyung Kim, Eric Mitchell, Jinwoo Shin, Yee Whye Teh, and Jonathan Richard Schwarz. Online adaptation of language models with a memory of amortized contexts. *arXiv preprint arXiv:2403.04317*, 2024.
- [148] Yu Wang, Yifan Gao, Xiushi Chen, Haoming Jiang, Shiyang Li, Jingfeng Yang, Qingyu Yin, Zheng Li, Xian Li, Bing Yin, et al. Memoryllm: Towards self-updatable large language models. *arXiv preprint arXiv:2402.04624*, 2024.
- [149] Peng Wang, Zexi Li, Ningyu Zhang, Ziwen Xu, Yunzhi Yao, Yong Jiang, Pengjun Xie, Fei Huang, and Huajun Chen. Wise: Rethinking the knowledge memory for lifelong model editing of large language models. *arXiv preprint arXiv:2405.14768*, 2024.
- [150] Weizhi Wang, Li Dong, Hao Cheng, Xiaodong Liu, Xifeng Yan, Jianfeng Gao, and Furu Wei. Augmenting language models with long-term memory. *Advances in Neural Information Processing Systems*, 36:74530–74543, 2023.
- [151] Jikun Kang, Wenqi Wu, Filippos Christianos, Alex James Chan, Fraser David Greenlee, George Thomas, Marvin Purtorab, and Andrew Toulis. Lm2: Large memory models for long context reasoning. In *Workshop on Reasoning and Planning for Large Language Models*, 2025.
- [152] Ali Behrouz, Peilin Zhong, and Vahab Mirrokni. Titans: Learning to memorize at test time. *arXiv preprint arXiv:2501.00663*, 2024.
- [153] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- [154] Wazeer Deen Zulfikar, Samantha Chan, and Pattie Maes. Memoro: Using large language models to realize a concise interface for real-time memory augmentation. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–18, 2024.
- [155] Bo Wang, Weiyi He, Pengfei He, Shenglai Zeng, Zhen Xiang, Yue Xing, and Jiliang Tang. Unveiling privacy risks in llm agent memory. *arXiv preprint arXiv:2502.13172*, 2025.