Rethinking Memory Mechanisms of Foundation Agents in the Second Half: A Survey

Wei-Chieh Huang[1,†], Weizhi Zhang[1,†,‡], Yueqing Liang[2,†], Yuanchen Bei[3,*], Yankai Chen[1,18,*], Tao Feng[3,*], Xinyu Pan[4,*], Zhen Tan[5,*], Yu Wang[14,*], Tianxin Wei[3,*], Shanglin Wu[6,*], Ruiyao Xu[7,*], Liangwei Yang[22,*], Rui Yang[3,*], Wooseong Yang[1,*], Chin-Yuan Yeh[1,*], Hanrong Zhang[1,*], Haozhen Zhang[8,*], Siqi Zhu[3,*], Henry Peng Zou[1], Wanjia Zhao[21], Song Wang[9], Wujiang Xu[10], Zixuan Ke[22], Zheng Hui[11], Dawei Li[5], Yaozu Wu[13], Langzhou He[1], Chen Wang[1], Xiongxiao Xu[2], Baixiang Huang[6], Juntao Tan[22], Shelby Heinecke[22], Huan Wang[22], Caiming Xiong[22], Ahmed A. Metwally[23], Jun Yan[23], Chen-Yu Lee[23], Hanqing Zeng[24], Yinglong Xia[24], Xiaokai Wei[25], Ali Payani[26], Yu Wang[27], Haitong Ma[12], Wenya Wang[8], Chengguang Wang[15], Yu Zhang[16], Xin Eric Wang[17], Yongfeng Zhang[10], Jiaxuan You[3], Hanghang Tong[3], Xiao Luo[4], Xue Liu[18,19], Yizhou Sun[20], Wei Wang[20], Julian McAuley[14], James Zou[21], Jiawei Han[3], Philip S. Yu[1,¶], Kai Shu[6,¶]

[†] *Co-first authors* [*] *Core contributors (listed alphabetically)* [‡] *Project organizer* [¶] *Core supervisors*

[1] *UIC* [2] *IIT* [3] *UIUC* [4] *UW–Madison* [5] *ASU* [6] *Emory* [7] *Northwestern* [8] *NTU* [9] *UCF* [10] *Rutgers* [11] *Cambridge* [12] *Harvard* [13] *UTokyo* [14] *UCSD* [15] *UCSC* [16] *TAMU* [17] *UCSB* [18] *MBZUAI* [19] *McGill* [20] *UCLA* [21] *Stanford* [22] *Salesforce* [23] *Google* [24] *Meta* [25] *Roblox* [26] *Cisco* [27] *Capital One*

AgentMemoryWorld/Awesome-Agent-Memory

Figure 1: **Roadmap of Foundation Agent Memory**. A timeline illustrating the trend of foundation agent memory frameworks, categorized by memory substrates and subjects (user or agent-centric).

Contents

Abstract

The research of artificial intelligence is undergoing a paradigm shift from prioritizing model innovations over benchmark scores towards emphasizing problem definition and rigorous real-world evaluation. As the field enters the "second half," the central challenge becomes real utility in long-horizon, dynamic, and user-dependent environments, where agents face context explosion and must continuously accumulate, manage, and selectively reuse large volumes of information across extended interactions. Memory, with hundreds of papers released this year, therefore emerges as the critical solution to fill the utility gap. In this survey, we provide a unified view of foundation agent memory along three dimensions: *memory substrate* (internal and external), *cognitive mechanism* (episodic, semantic, sensory, working, and procedural), and *memory subject* (agent- and user-centric). We then analyze how memory is instantiated and operated under different agent topologies and highlight learning policies over memory operations. Finally, we review evaluation benchmarks and metrics for assessing memory utility, and outline various open challenges and future directions.

1 Introduction

The landscape of Artificial Intelligence (AI) has now undergone a fundamental paradigm shift: from prioritizing foundation model architecture and simplified benchmark performance to emphasizing problem definition and rigorous real-world evaluations. This marks the end of the "first half" of AI development, in which progress was primarily driven by training methods (Liu et al., 2024a), scaling (He et al., 2016; Achiam et al., 2023; Gu et al., 2025), and model architectures (Vaswani et al., 2017), which repeatedly pushed higher scores on standardized benchmarks (Wang et al., 2024n; Krizhevsky et al., 2012). In the first half of AI development, the field has evolved towards a dominant paradigm of scaling data and model size. A general recipe of massive pre-training (Minaee et al., 2024) followed by a post-training process (Ouyang et al., 2022) that solves traditional benchmarks with remarkable accuracy. Under well-defined training pipelines, Large Language Models (LLMs) and agents can achieve over 90% accuracy on benchmarks such as MMLU (Hendrycks et al., 2021b) or MATH (Hendrycks et al., 2021c). As a result, LLMs and agents have rapidly evolved from static predictors, like conventional machine learning models, into general-purpose agents capable of complex reasoning (Wu et al., 2025f), planning (Li et al., 2025i), and tool use (Huang et al., 2025a; Zou et al., 2025b) in various tasks and environments.

Despite the impressive capabilities demonstrated on standard benchmarks, a significant gap remains between the reported performances and the utility in many real-world tasks and environments (Yu et al., 2024b). The majority of evaluation protocols largely simplify experimental assumptions and design static, pre-defined rules, with relatively short and isolated task settings (Cobbe et al., 2021; Chen et al., 2021; Budzianowski et al., 2018). In particular, most recent agent evaluation benchmarks are coupled with short agentic execution times without multi-turn, long-term interaction (Wang et al., 2024k; Lu et al., 2025a). As a result, these evaluations no longer reflect the foundation agent's ability in reality, where interactions are inherently long-horizon, long-context, and deeply user-dependent with high-level complexity. As the field transitions towards more realistic settings, such as embodied agents (Li et al., 2024b), GUI automation (Ye et al., 2025a), deep research (Huang et al., 2025c), personal health-care (Zhan et al., 2024), and human-agent collaborations (Feng et al., 2024; Zou et al., 2025c), the complexity of the operational environment explodes, exposing agents to exceptionally large and dynamic contexts. In such settings, static, one-shot capabilities are insufficient. Instead, agents must accumulate, retain, and selectively reuse information across interactions. Memory thus emerges as the critical and natural solution to bridge the gap between idealized benchmark performance and real-world implementation and environment (Zhang et al., 2025o).

As the field enters the "second half" of AI development, the focus shifts from improving training recipes to solving the critical utility problem in reality (Bell et al., 2025; Yao et al., 2025). How to design a benchmark to evaluate an agent in the real environment has become one of the most important challenges (Xu et al., 2025c), particularly as agents strive to adapt along two primary empirical dimensions: **user-facing personalization** (Cai et al., 2025; Zhang et al., 2025m; Wu et al., 2025e) and **task-oriented specialization** (Ling et al.,

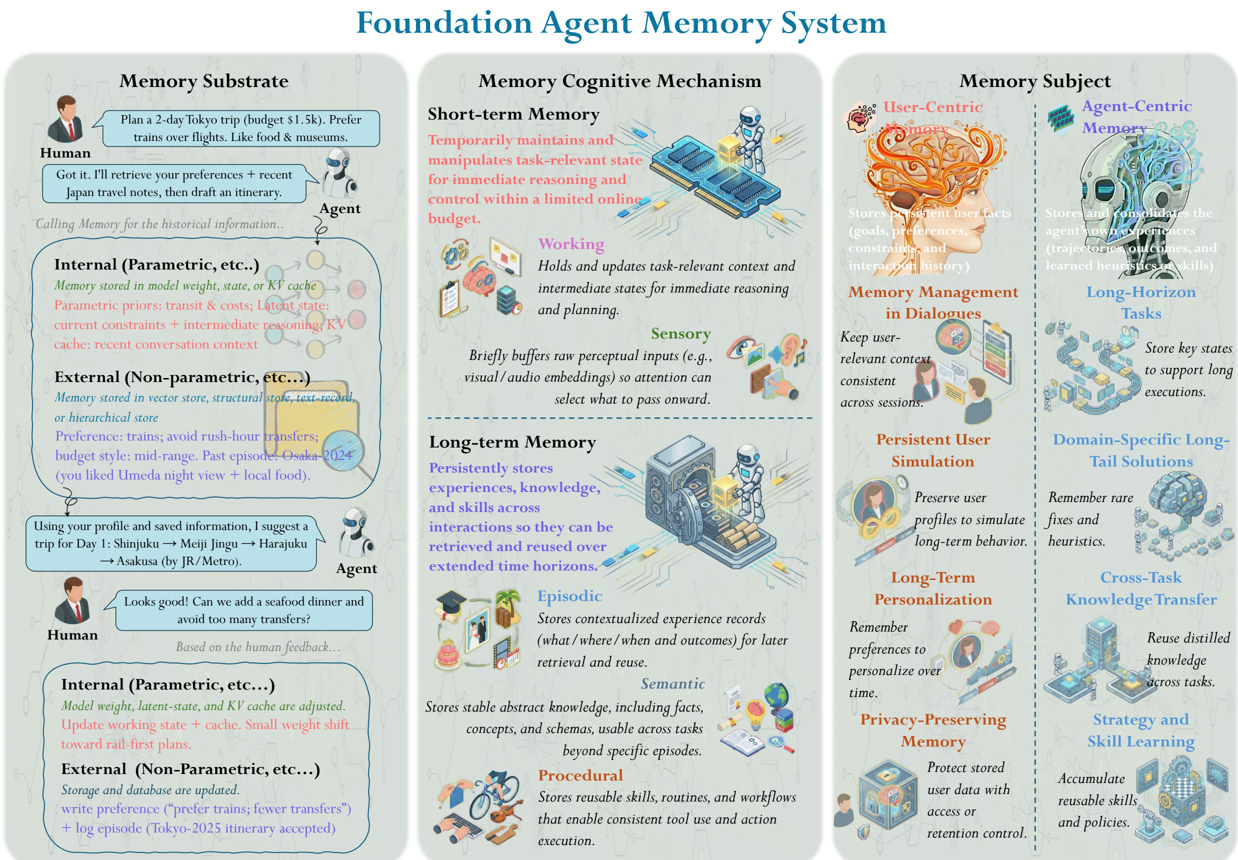

Figure 2: **The taxonomy of Foundation Agent Memory**. The **Memory Substrate (what form is represented)** for foundation agents includes the *internal* and *external* memory. In the **Memory Cognitive Mechanism (how memory functions)** perspective, memories are categorized into *episodic*, *semantic*, *sensory*, *working*, and *procedural* memory. Based on the **Memory Subject (who is supported)**, the memory is characterized into *user-centric* and *agent-centric* perspectives.

2025; Zhang et al., 2025i). In both dimensions, the interaction contexts of a specific user's long-term history or the vertical tasks like coding (Islam et al., 2024) and web search (He et al., 2024a) expand far beyond what can be accommodated by prompt-based mechanisms alone. As multi-session data from daily interactions or accumulated context from project work expands exponentially, reliance on a static memory mechanism is insufficient. As a result, memory architectures evolve from a static, predefined, and simple mechanism (Hao et al., 2023; Hu et al., 2023) towards a self-adaptive, self-evolving, and flexible unit (Liu et al., 2025g;f), to intelligently store, load, summarize, forget, and refine to keep the informative experience for downstream.

Although the rapid growth of foundation agent memory research has produced several surveys, important gaps remain in how agent memory is analyzed from a system design perspective in real-world agent utility backgrounds. Such memory design has shifted from short, isolated prompts to long-horizon interaction, where agents must operate over exploding context windows, multi-session workflows, and persistent real-world user relationships. Early works often organize memory primarily by task applications or management strategies (Zhang et al., 2025o; Du et al., 2025), or adopt neuroscience-inspired perspectives that project AI memory onto human memory through functional analogies and memory lifecycles (Wu et al., 2025g; Liang et al., 2025). While these approaches provide useful conceptual grounding, they do not systematically characterize underlying memory substrates or explicitly model the subject that memory serves within an agent system, which are significant when the context exceeds the foundation model's limitations. As a result, they fall short of distinguishing the optimization goals of agent memory and overlook complementary dimensions connections that are essential for designing and deploying autonomous agent systems in real-world applications. More recent work begins to broaden this conceptual landscape. In particular, Hu et al. (2025d) organizes agent memory along forms, functions, and temporal dynamics. Despite offering a valuable consolidation of memory, it remains largely partial, focusing mainly on how memory functions in agent-centric tasks rather than how memory should be designed, optimized, and deployed for users. Motivated by these, this survey analyzes foundation agent memory of hundreds of papers from three major complementary perspectives that connect memory with system-level design choices in increasingly complex environments.

Specifically, we introduce a unified taxonomy that is organized around three core design dimensions: the **memory substrate**, the **cognitive mechanism**, and the **memory subject**, as shown in Figure 2. We classify existing foundation agent memory works by substrate into external and internal, by cognitive mechanism into functional categories such as episodic, sensory, working, semantic, and procedural memory, and by subject into user-centric versus agent-centric subjects. From a system perspective, we further analyze how these foundation agent memory systems are operated under different agent topologies, distinguishing the fundamental memory operations in single-agent systems from those memory routings in multi-agent settings. Furthermore, we highlight the growing role of learning policies, showing how agents are increasingly trained to conquer memory management itself and thus to learn from their interaction histories and self-evolve over time. To reflect the impact of shifting environments on foundation agent memory design, we discuss scalability issues across context length and environment complexity, and we review the evaluation method and metrics used to measure memory performance and utility. Finally, we outline six open challenges in foundation agent memory to guide the next generation of foundation agent memory design.

2 Background

2.1 Large Language Models and Foundation Agents

Recent advances have pushed large language models beyond one-shot question answering into foundation agents (Park et al., 2023; Xi et al., 2025a; Luo et al., 2025; Liu et al., 2025a): systems capable of perceiving environments, reasoning through complex objectives, and executing actions to achieve assigned goals. Unlike traditional "one-shot" chatbots, an agent operates in a full loop: it interprets instructions, selects actions, observes outcomes, and updates its internal state or memory. This iterative interaction makes agents particularly well-suited to long-horizon, dynamic problem solving. This trend is already visible in emerging agentic products such as Deep Research and Manus, which emphasize multi-step execution and tool-augmented decision making over single-turn responses (Zhang et al., 2025l).

A foundation agent is typically an autonomous or semi-autonomous system that uses foundation models, such as LLMs, as its core decision modules, augmented with mechanisms for state estimation, action execution, and memory management. In this survey, we use the term *foundation agent* to denote an AI agent whose core decision-making is driven by a general-purpose foundation model, including large language models, vision–language models, or learned world models. Core capabilities commonly include planning for task decomposition and decision making (Yao et al., 2023; Huang et al., 2024b), tool use through external functions or models (Wu et al., 2025c; Yuan et al., 2025c; Lu et al., 2025a), multimodal perception (Liu et al., 2023a; Bai et al., 2025), and memory that spans both short-term context and longer-term horizons (Lumer et al., 2025; Wang et al., 2025s). Memory (Zhang et al., 2025o; Xiong et al., 2025d) becomes especially important because context windows are limited and environments evolve over time; accordingly, many agents rely on external memory stores, coupled with summarization (Lu et al., 2025b), reflection (Renze & Guven, 2024), or consolidation procedures that compress experience into reusable knowledge (Kang et al., 2025c; Yu et al., 2025b). Multi-agent systems (Talebirad & Nadiri, 2023; Wu et al., 2024b) extend this paradigm by assigning specialized roles and memories to multiple agents that communicate, coordinate (Lan et al., 2024; Estornell et al., 2025).

Foundation agents are now being explored for different real-world applications, such as workflow automation (Xiong et al., 2025c), tutoring (Wang et al., 2025l), web and GUI interaction (He et al., 2024a; Wang et al., 2025e), embodied control in simulated or real environments Fan et al. (2022); Yang et al. (2025c), and early forms of agentic scientific assistance (Ren et al., 2025; Pantiukhin et al., 2025; Zheng et al., 2025d). Despite

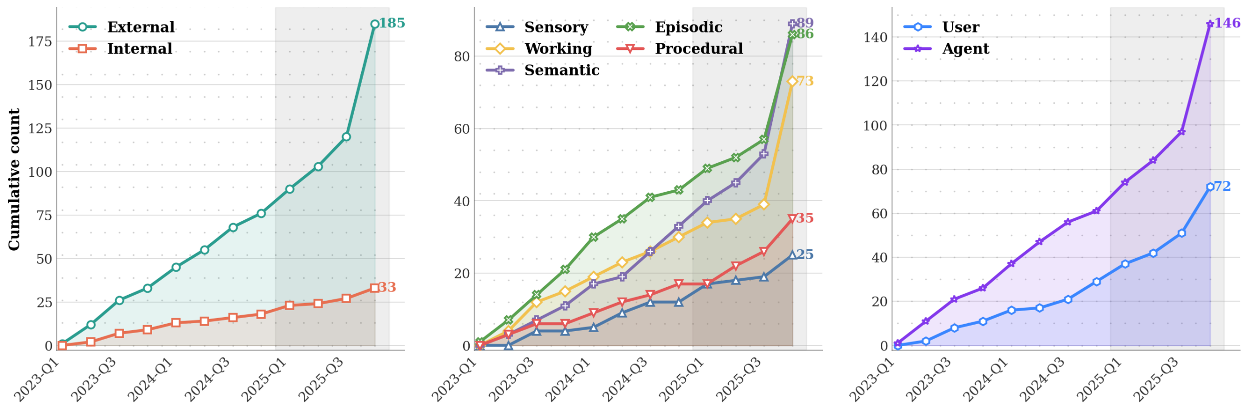

Figure 3: **Cumulative publication trends of memory-related research in LLM agents (2023 Q1 – 2025 Q4)**. The plots illustrate the distribution of 218 collected papers across three key dimensions: memory substrate (left), memory cognitive mechanism (middle), and memory subject (right). The shaded region highlights the rapid acceleration of research output observed in 2025.

this progress, several foundational challenges remain. First, long-horizon reliability (Huang et al., 2024b; Xi et al., 2025b) requires preventing compounding errors, behavioral loops, and unreliable replanning. Second, evaluation (Yehudai et al., 2025) should move beyond static QA toward benchmarks that measure dynamic, interactive capabilities, such as tool use, multi-step decision making, and long-horizon feedback. Third, alignment and safety (Hua et al., 2024; Yuan et al., 2024a; Tian et al., 2023; Zhang et al., 2024c) become increasingly important as autonomy and tool access expand, demanding stronger guarantees of controllability. Addressing these challenges could ultimately turn today's tool-using assistants into agents that are capable and trustworthy to solve more complex tasks and stay predictable and controllable in real-world deployments.

2.2 Memory

Memory generally refers to a system's ability to retain, organize, and exploit information over time. In the context of LLM-based foundation agents, memory is used to explain how agents go beyond single-turn contexts to support long-term interaction, behavioral consistency, and experience accumulation (Luo et al., 2025). While biological studies often characterize memory as persistent, experience-driven neural change (e.g., synaptic plasticity and consolidation (Hebb, 2005; Bliss & Lømo, 1973; Tonegawa et al., 2015)), for agent systems, the more relevant insight lies in how memory is designed, realized, and used in practice to support different functions, representations, and targets.

In human cognitive models, memory is commonly understood as a set of interacting subsystems organized across different time scales. Short-term memory supports the temporary retention and manipulation of information during ongoing processing. Sensory memory briefly buffers raw perceptual input, enabling downstream processing (Sperling, 1960), while working memory operates under strict capacity constraints and supports online information manipulation, reasoning, and control (Baddeley, 2020; 2000). Long-term memory supports information retention over extended periods and comprises multiple functionally distinct systems. Episodic memory stores specific experiences situated in time and context, semantic memory accumulates abstract facts and conceptual knowledge (Tulving, 1972; 1985), and procedural memory captures skills, habits, and action policies that are typically expressed implicitly through performance rather than explicit recall (Cohen & Squire, 1980; Squire, 1992). In biological systems, functional distinctions in memory are ultimately grounded in physical substrates, such as synaptic plasticity (Hebb, 2005; Bliss & Lømo, 1973; Martin et al., 2000) and circuit-level changes that give rise to enduring memory traces or engrams (Tonegawa et al., 2015; Josselyn & Tonegawa, 2020). While these cognitive and biological perspectives provide essential grounding, foundation agent memory systems introduce additional design considerations, including explicitly distinguishing whose information memory is designed to capture and support.

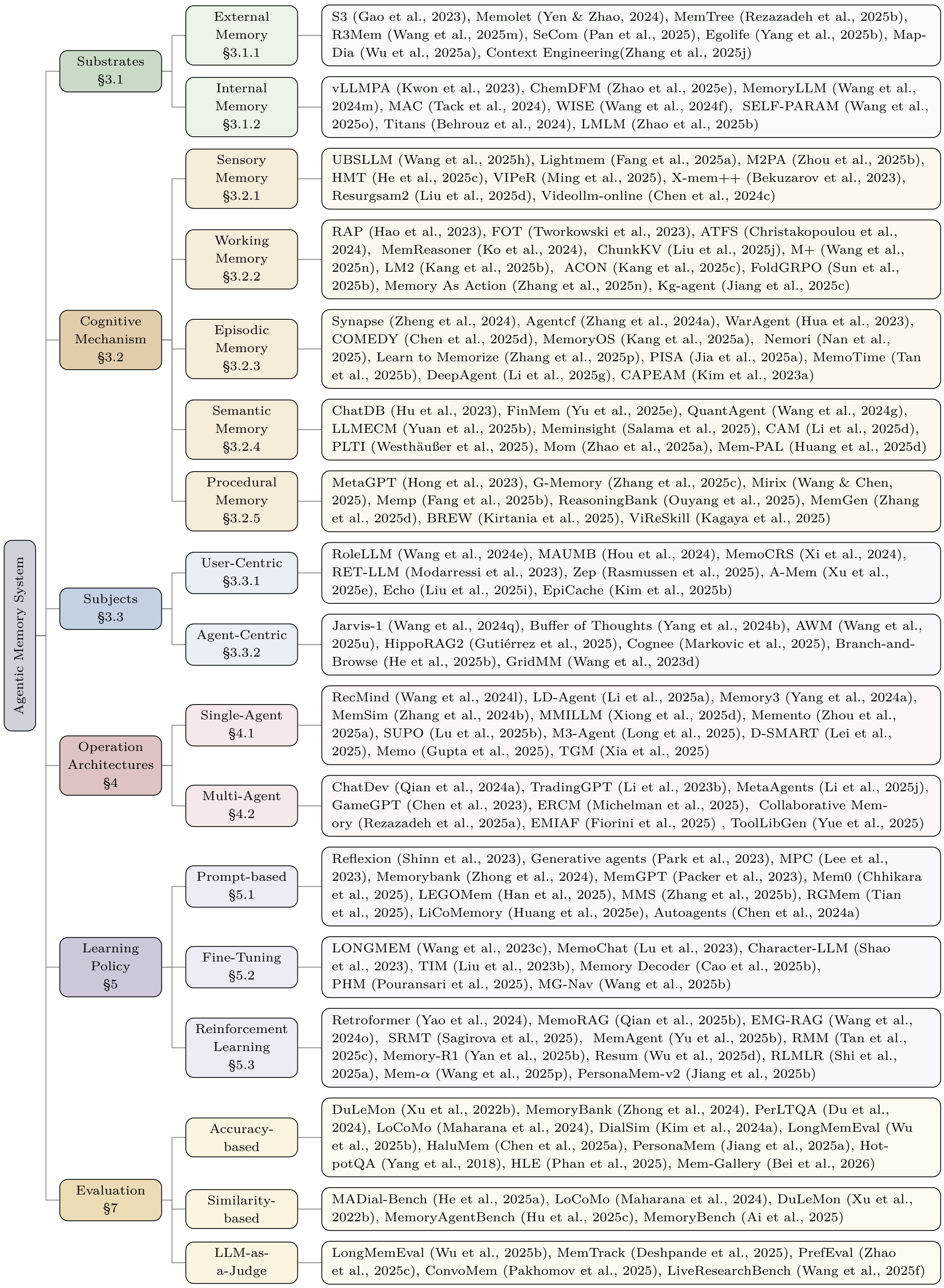

Figure 4: A taxonomy of the Foundation Agent Memory System

3 Taxonomy of Memory in Foundation Agents

We conducted a systematic literature collection by querying Google Scholar with memory-related keywords (e.g., agent memory, long-term memory, context management, personalization memory) and manually scrutinizing proceedings of major computer science conferences and journals (including top-tier NLP, ML, IR, and AI venues). From an initial pool of several hundred papers, we curated the most relevant contributions through iterative screening, ultimately selecting 218 key articles published between 2023 Q1 and 2025 Q4. The resulting trends, illustrated in Figure 3, demonstrate a dramatic escalation in research activity. Most notably, the publication volume witnesses an exponential surge throughout 2025, culminating in a significant peak in Q4. The trends show that the improvement need of the intelligence memory design to support the agent complete the long-horizon, long-context task in the more complex environment.

To better distinguish memory design, we categorize memory in foundation agents along three orthogonal perspectives, as shown in Figure 2: (1) **Memory Substrates**, or so-called storage format, describing what form memory is represented in different settings, are presented in Section 3.1; (2) **Memory Cognitive Mechanisms**, describing the functional role memory serves in the pipeline or workflow, are described in Section 3.2; (3) **Memory Subjects**, whose information the memory is designed to capture and support, are elaborated in Section 3.3. In addition, we present the taxonomy in Figure 4. The operation and management of memory in single- and multi-agent systems is introduced in Section 4. The learning policy is introduced in Section 5. The scaling of the foundation agent system is introduced in Section 6. The Evaluation of the foundation agent is introduced in Section 7. The application of the foundation agent system is introduced in Section 8. The future challenge of the foundation agent system is highlighted in Section 9.

3.1 Memory Substrates

Memory substrates serve as the essential mechanisms for retaining historical knowledge produced during interactions between foundation agents and humans in various tasks and environments. Based on the framework, storage mediums, and persistence mechanisms of current research, we categorize these substrates into external and internal memory. The definitions and implementations of these categories are elaborated in Sections 3.1.1 and Section 3.1.2, respectively.

3.1.1 External Memory

External Memory

External memory refers to any memory substrate that stores the knowledge, information, and past experience **outside the agent model's parameter or state**. The agent can explicitly read from and write to the external memory via retrieval and update operations, enabling scalable, easy-to-update, cross-session retention of knowledge and interaction history.

The external memory represents information storage systems that store information in an **vector index**, **text-record**, **structural store**, and **hierarchical store** (Lewis et al., 2020; Zhang et al., 2023b). It operates independently of the gradient-updated weights within the neural network. And, it is characterized by a clear separation between the computation, performed within the LLM's internal parameters, and the knowledge, stored in an external database or memory module (Mallen et al., 2023; Zhong et al., 2024). This separated computation and retrieval design allows agents to access extensive and continually updated information without requiring expensive retraining of the foundation model (Omidi et al., 2025; Aratchige & Ilmini, 2025). In addition, external memory is both scalable and flexible. With just a few changes, it can be easily added, replaced, or inserted into most frameworks. Moreover, unlike internal memory (Saha et al., 2021), where knowledge may be overwritten or adjusted due to weight updates from new experience, external memory can preserve past information in its original form, helping reduce hallucinations or knowledge cutoff by utilizing an additional storage module (Castrillo et al., 2025).

However, this design also has drawbacks and trade-offs. Normally, inference takes a longer time since the agent has to retrieve information from external storage. This problem will be severe when running in multi-turn iterations or in complex environments. In addition, the retrieval can be unreliable. If the similarity or ranking algorithms are not well-developed, they may provide results that are irrelevant or unhelpful for the current task. This irrelevant information would introduce noise that may reduce the agent's utility (Xu et al., 2025a). As the system operates for longer periods, with more turns, and when the memory grows, the escalating costs of storage and indexing may further diminish efficiency and performance (Chen et al., 2025e). As a result, a well-constructed external memory usually demands solid procedures for regulating memory size and quality, including summarization, selective retention, forgetting, deduplication, and periodic pruning of low-value or obsolete items (Du et al., 2025; He et al., 2024b).

Vector Index. Vector index is implemented by embedding memory items or query into a shared vector space, running approximate nearest-neighbor search to fetch top-k items, and appending those snippets to the prompt as grounding context (Lewis et al., 2020; Zou et al., 2025a). The dominant implementation of external memory in the current work of agents or LLMs is the vector database, utilized within the RAG framework (Quinn et al., 2025; Li et al., 2025i; Zhang et al., 2025l). This approach operates memory as a geometric problem. By embedding textual or multi-modal information into high-dimensional vectors, it represents each memory item as a point in continuous space. New queries are also transformed into vectors with the same embedding mechanism, and the agents search for the closest points, the data representing the closest semantic information. Current techniques, such as hierarchical navigable small-world (HNSW) (Liu et al., 2025h), inverted file (IVF) (Rege et al., 2023), or product quantization (PQ) (Huang & Huang, 2024) indexing, are developed and optimized to find the most semantically similar documents or past interactions efficiently in a high-dimensional embedding space. This approximate-nearest-neighbor search retrieves a small set of context snippets, which are mapped back to their original form and fed into the agent or LLM to integrate its response with relevant, previously unseen information (Guu et al., 2020).

Text-record. Text-record memory treats an agent's long-term memory as a set of explicit text documents rather than embeddings or graphs (Zhang et al., 2025o). It is implemented as persistent human-readable text artifacts (e.g., a running "core summary" plus episodic logs) that are periodically summarized or edited, and selectively copied into the prompt when generating the next response. A common implementation keeps a persistent core summary and augments it with semantic and episodic lists. The semantic list stores discrete facts and can be modified via insert, update, and delete operations. In contrast, the episodic list is a chronological ledger of timestamped events and interactions. During memory updates, the agent modifies these structures (Zhu et al., 2025c; Wang & Chen, 2025). During response generation, it extracts a limited selection of pertinent information and instances, then reintegrates them into the prompt with the primary summary. The architecture facilitates transparency and fast integration by structuring memory as human-readable summaries and lists (Yue et al., 2025; Park et al., 2023). However, it requires meticulous summarization and pruning to maintain a succinct core summary and manageable lists.

Structural Store. Structural store is implemented by storing memories in explicit schemas and retrieving them via symbolic queries or traversal before formatting the returned records for the LLM's context. The structured memory architecture (Jia et al., 2025a; Bei et al., 2025), based on the topological design, can be partitioned into *relational tables*, *graph-based structured*, or *tree-structure memories*. **Relational tables** use SQL to retrieve row records from a table. They can store facts and preferences as structured records, separate short-term transactions and long-term events into different tables, and use joins and indexes for efficient retrieval between different tables (Wang et al., 2024d). **Graph-based memory** represents episodes or pieces of information as nodes and edges within a knowledge graph (Anokhin et al., 2024). A knowledge graph is essentially a semantic network. It typically organizes information into nodes with entities and edges capturing the relationships between them. When new information arrives, the system combines semantic embeddings, keyword search with traversal to detect edge changes, resolve conflicts, and maintain the valid graph structure (Jiang et al., 2025c). **Tree-based memory** organizes knowledge hierarchically, with each node in the tree storing an aggregated summary of textual content and a corresponding semantic embedding, and deeper branches representing more specific details (Sarthi et al., 2024). When new information arrives, the system traverses the tree and updates or creates nodes based on embedding similarity to decide where the information is stored. This dynamic mechanism supports multi-level abstraction and efficient retrieval to support long conversation tasks (Rezazadeh et al., 2025b). The structured memory, with its different types of formats, can capture sequential history and relationships and support multi-level abstraction. This technique offers the agent flexibility and control over how memories are stored, updated, and queried.

Hierarchical Store. Hierarchical memory separates memory into multiple storage units, each with its own schema and storage strategy. **Instead of keeping everything in a single flat archive, the agent maintains dedicated memory modules** (Rezazadeh et al., 2025b; Zhai et al., 2025). For instance, a core memory for persistent persona and user facts, an episodic memory for time-series or sensitive events, a semantic memory for abstract concepts, a procedural memory for step-by-step instructions, a resource memory for documents and media, and a knowledge store for sensitive or private information (Yang et al., 2025a; Wang et al., 2025p). Each module utilizes a data structure suited to its content. For example, episodic memory might store records with `event_type`, `summary`, `details`, and `timestamp`; semantic memory may organize entries by `type`, `summary`, and `source`; procedural memory encodes workflows as JSON sequences; and the knowledge store applies strict access controls to secure credentials and API keys. A meta-memory manager (Wang & Chen, 2025; Zhang et al., 2023a) coordinates these modules, directing queries to the correct store and utilizing specialized retrieval methods (e.g., *embedding similarity* (Xu et al., 2025d), *BM25 match* (Hu et al., 2025c), *string match* (Alla et al., 2025)) for each storage unit. This multi-store substrate has advantages such as modularity, separation of concerns, and scalability, making it easier to design specialized schemas and retrieval paths while supporting nuanced personalization across long-term interactions. However, coordinating multiple memory agents increases the system's architectural complexity. It can potentially add latency and additional resource cost when querying several stores, and requires well-designed mechanisms to maintain data consistency and synchronize updates between different turns in the long-horizon task (Sun & Zeng, 2025; Li et al., 2025k; Maragheh & Deldjoo, 2025).

3.1.2 Internal Memory

Internal Memory

Internal memory refers to the information stored **directly within the model's architecture**, encompassing both the persistent knowledge embedded in its parameters (i.e., parametric memory) and the working states utilized during inference.

Weights. Weight memory is implemented by writing memory into parameters, so the model later recalls the information without retrieving external context. **Knowledge or experience is embedded directly into the neural network's parameters** (Mallen et al., 2023) through *pre-training*, *post-training*, or *targeted parameter editing* (Wang et al., 2025q; Ampel et al., 2025; Wang et al., 2024h). Because this knowledge is internal, these models recall facts and past events efficiently and robustly without communicating with an external store. However, maintaining and updating internal weights is challenging. Therefore, recent research has divided weight updates into three main strategies: *continual learning*, *model editing*, and *distillation*. **Continual learning** methods incrementally update the model's weights with the new information while attempting to avoid catastrophic forgetting (De Lange et al., 2021). Methods like regularizing changes to prevent overwriting important weights and applying soft masks to selectively update parameters can largely preserve the previous domain knowledge while recognizing new information (Serra et al., 2018). **Model editing** strategies treat specific factual associations as localized memories and adjust the corresponding parameters: one line of work locates decisive mid-layer weights and applies rank-one updates to change a single factual association, whereas another extends this idea to update thousands of memories by applying small updates across multiple layers (Abdali et al., 2024; Meng et al., 2022a). For instance, some methods show that carefully fine-tuning the model with augmented data can achieve comparable editing performance. Another method inserts a few new neurons to correct mistakes sequentially without affecting unrelated behavior. Another approach adds calibration memory slots that store corrected facts without altering the original parameters (Chhetri et al., 2025; Luo & Specia, 2024). **Distillation** techniques compress contextual knowledge into the parameters themselves. It can internalize prompt- or context-dependent behaviors into model parameters by training a student to reproduce a prompted teacher's outputs (e.g., prompt-to-weights or in-context distillation). Concretely, some approaches optimize the model so it behaves as if a fixed prompt were implicitly present at inference time (Cao et al., 2025a; Padmanabhan et al., 2023). While such parameter-based memorization can make recall efficient at inference, modifying weights is computationally costly and may introduce interference, overwriting, or distortion of previously learned knowledge (Meng et al., 2022b).

Latent-State. Latent-state memory is implemented by carrying forward and reusing intermediate hidden states across steps or segments. Earlier activations can directly impact later computation even when the raw tokens are no longer in the window. The intermediate activations or hidden-state tensors are produced as information flows through layers during the forward pass process (Ibanez-Lissen et al., 2025). These hidden states are the per-layer-wise representations that combine the fixed learned parameters with the current prompt, forming the model's working state that drives the next-token prediction. Unlike weight-based memory, state memory is not durable. It is usually created during runtime activations, then used, and typically discarded or reset at the end of a request (Jelassi et al., 2024). Therefore, it supports within-session coherence and reasoning but does not persist knowledge across sessions unless exported externally. The key trade-off is resource cost. The environment must hold not only parameters but also intermediate states in memory, and this activation value scales with model depth, batch size, and precision. Efficient and optimized latent-state management reduces inference latency and memory cost. One representative direction reconstructs hidden representations across layers to reinforce earlier context at a controllable cost (Dillon et al., 2025). In contrast to reconstruction, other work caches a compact subset of hidden-state vectors or memory tokens across segments, thereby extending the effective context length beyond a fixed window while reducing memory and computation costs (Dai et al., 2019). Beyond reusing existing states, latent-state memory can also be generated. A memory trigger and memory weaver synthesize machine-native latent tokens that are fed back into the model to enrich downstream reasoning (Zhang et al., 2025d). Meanwhile, structured modulation of hidden-state transitions maintains latent trajectories aligned with prior context, which can reduce semantic drift in long sequences (Carson & Reisizadeh, 2025). Additionally, some architectures integrate compressive memory into the attention mechanism to store and reuse key-value pairs from previous segments, enabling the processing of very long inputs with bounded memory (Munkhdalai et al., 2024). Other strategies treat the hidden state as a fast-weight updated through small optimization steps during inference to refine internal representations over extended contexts (Zhu et al., 2025b).

KV Cache. In transformer-based LLMs, the KV cache is a transient, inference-time memory that speeds up autoregressive decoding (Kwon et al., 2023; Liu et al., 2023c). It is implemented by caching per-layer attention keys and values from previous tokens during decoding and reusing them for subsequent tokens. During self-attention, each token is projected into keys and values (Ge et al., 2023; Pope et al., 2023). Without KV caching, these matrices are recomputed for all previous tokens at every step, leading to unwanted resource waste and slowing generation for long sequences (Cai et al., 2024; Feng et al., 2025b). The KV cache stores the keys and values from earlier tokens and, at each new step, only computes them for the freshly generated token and retrieves the rest from the cache. This mechanism significantly accelerates inference, especially on long outputs or large and deep models, but comes at the cost of higher memory usage and implementation complexity (Pope et al., 2023; Kwon et al., 2023). Recent research proposes several high-impact approaches for compressing the KV cache. One approach recognizes that a small portion of tokens contribute most to attention scores and dynamically evicts the rest, balancing "heavy hitters" with recently generated tokens. This method achieves large throughput improvements while retaining only a fraction of the cache (Zhang et al., 2023c). Another technique identifies consistent attention patterns within a prompt's observation window and clusters important features to enable substantial speed and memory savings for long input sequences (Li et al., 2024d). Another approach observes that layers differ in how many key-value vectors they truly need, and instead of giving every layer the same cache size, ranks vectors by importance and uses a binary search to decide how many to keep per layer under a global budget, so only the most informative vectors are retained (Wang et al., 2024a). These current studies demonstrate that the KV cache can be used efficiently to improve model performance significantly.

3.1.3 Tradeoffs Across Memory Substrates

Different memory substrates can have different advantages and drawbacks. They have trade-offs in terms of access speed, scalability, adaptability, and reliability in ways that show up quickly on long-horizon tasks. In practice, the choice often comes down to fast, precise internal state versus scalable external storage that can become noisy as it grows.

Internal memory, including parametric memory encoded in model weights, usually offers high-speed access and tight integration with reasoning. However, it is expensive to update and can suffer from catastrophic forgetting due to frequent modifications. It therefore fits stable, general-purpose knowledge better than rapidly changing or user-specific information. Latent substrates such as hidden activations or KV caches are fast and transient, making them well-suited to within-session state. However, their ephemeral nature and linear scaling with sequence length limit capacity and make them unsuitable for cross-session retention.

External memory, consisting of vector databases or stores, scales naturally with experience and supports flexible editing without retraining. However, retrieval adds latency and makes performance sensitive to indexing and retrieval quality. Prior work shows that storing excessive or low-quality items can inject noise and increase the difficulty of targeting the informative items, which degrades tightly coupled decision-making (Liu et al., 2024b). To better tackle the issue, explicit memory management mechanisms such as pruning, summarization, and hierarchical organization are required and necessary.

Overall, no single substrate dominates across settings and environments. Effective system design increasingly adopts hybrid memory architectures, using internal or parametric memory for inherent knowledge or facts, latent memory for fast short-term reasoning, and external memory for scalable experience storage. This pattern reflects a broader shift from a static or persistent, defined knowledge storage mechanism to a more dynamic and adaptive approach that can handle the complexities of real-world tasks.

3.2 Memory Cognitive Mechanisms

Human memory provides a conceptual scaffold for analyzing memory in LLM-based agents (Kim et al., 2023b; Li & Li, 2024). Cognitive psychology distinguishes multiple interacting systems that explain how information is perceived, maintained, and reused (Baddeley, 2020; Tulving, 1972). While a wide range of cognitive memory types have been proposed in the literature, we focus on a set of five atomic cognitive memory systems that are particularly relevant for LLM-based agents: **sensory**, **working**, **episodic**, **semantic**, and **procedural memory**. Table 1 summarizes these five memory systems by mapping their core functional roles to representative research directions and illustrative agent-level implementations. These five systems constitute a minimal and architecturally complete decomposition of cognitive memory, whereas other memory constructs such as autobiographical (Conway & Pleydell-Pearce, 2000) or prospective memory (Einstein & McDaniel, 1990) can be understood as compositions or functional abstractions built upon them. Accordingly, our taxonomy is organized around these five atomic memory types, which cover both short-term and long-term memory mechanisms commonly realized, either explicitly or implicitly, in current agent architectures.

3.2.1 Sensory Memory

Sensory Memory

Sensory memory refers to the temporary retention of incoming perceptual signals, allowing attention and selection mechanisms to operate before higher-level processing occurs, by briefly holding raw inputs long enough for the system to decide what to attend to next.

In current foundation agents, sensory memory is typically not explicitly modeled. This is largely due to the highly abstracted nature of textual inputs, where perceptual processing has already been collapsed into symbolic or linguistic representations. In contrast to memory systems that encode stable knowledge, sensory memory functions as a transient interface between perception and cognition, operating over very short timescales and across multiple sensory modalities (Atkinson & Shiffrin, 1968). However, in multimodal or embodied agents, analogous mechanisms emerge in the form of short-lived perceptual buffers, such as caches of visual, auditory, or interaction embeddings. These buffers function as a sensory stage by temporarily retaining minimally processed observations before they are filtered, summarized, or routed to working memory for downstream reasoning and control.

Table 1: Mapping five cognitive memory systems to their functional roles and corresponding memory design directions in foundation agents, with illustrative agent-level examples.

Cognitive Type	Functional Role	Core research focuses in LLM agents (representative works)	Implementations
Short-Term Memory			
Sensory Memory	**What is perceived**. Brief retention of recent visual, auditory, or other sensory inputs before further processing.	**Perceptual buffering and lightweight caching** for multimodal streams, such as short-lived embedding queues and perceptual state buffers (He et al., 2025c; Fang et al., 2025a; Zhou et al., 2025b; Di et al., 2025). **Temporal gating and selection** mechanisms that stabilize noisy or high-bandwidth observations for downstream reasoning and control (Mon-Williams et al., 2025; Bjorck et al., 2025; Black et al., 2025; Ravi et al., 2024; Liu et al., 2025d).	**Keep the last 2–5 seconds of audio and video frames** (or recent sensor embeddings) to smooth perception and handle brief occlusions.
Working Memory	**What is currently handled**. Temporary holding and manipulation of current information.	**Pre-write representation shaping** that reduces what must enter the active context, including compression, folding, and abstraction-aware representations (Labate et al., 2025; Kang et al., 2025c; Wu et al., 2025d; Sun et al., 2025b; Ye et al., 2025b). **Online state maintenance under fixed budgets**, including update, eviction, and runtime control of context and KV states during execution (Zhang et al., 2025n; Yuan et al., 2025a; Kim et al., 2025b; Liu et al., 2025j; Kwon et al., 2023; Ni et al., 2025).	**An in-progress reasoning state (chain of thought):** "goal: refine the survey; earlier sections set the framing; the next revision should preserve framing consistency."
Long-Term Memory			
Episodic Memory	**What happened**. Contextual record of specific experiences.	**Episode recording and structuring**, including what to write, how to organize events and trajectories, and multi-scale episode formation (Rajesh et al., 2025; Yeo et al., 2025a;b; Anokhin et al., 2024). **Retrieval and reflection at decision time**, including adaptive triggering, episode selection, and retention policies that shape long-term accessibility (Yeo et al., 2025b; Latimer et al., 2025; Li et al., 2025f; Sarin et al., 2025; Alqithami, 2025).	**A past interaction log:** "last time you preferred a 2-page summary; the previous plan failed due to missing API keys," stored with its time and situational context.
Semantic Memory	**What is known**. Conceptual and factual knowledge about the world.	**Knowledge induction and organization** into reusable representations, including memory graphs, schemas, and compact neural representations (Zhao et al., 2025a; Jia et al., 2025a; Li et al., 2025d; Rasmussen et al., 2025; Behrouz et al., 2024; Wang et al., 2024m; Pouransari et al., 2025). **Knowledge access and reliability control** during reasoning, including selective activation, validation, and continual revision under distribution shift (Jimenez Gutierrez et al., 2024; Wang et al., 2025m; Rezazadeh et al., 2025b; Yan et al., 2025b; Wang et al., 2024f; Alqithami, 2025).	**A knowledge base:** entities or facts (e.g., project info, preferences, definitions) retrieved by query and checked for reliability.
Procedural Memory	**How to act**. Skills and action patterns.	**Skill induction and packaging**, learning reusable procedures from experience, tools, or interaction traces (Hong et al., 2023; Fang et al., 2025b; Han et al., 2025; Zhang et al., 2025d; Xia et al., 2025). **Skill execution, composition, and adaptation**, invoking and refining procedures under changing contexts over long horizons (Ouyang et al., 2025; Li et al., 2025h; Tablan et al., 2025; Terranova et al., 2025; Wang & Chen, 2025).	**A reusable workflow or tool skill:** "search → read → extract → cite," or "debug with sanitizer," invoked as a routine.

Only a limited number of LLM agent works explicitly instantiate sensory memory as a distinct stage in their memory architectures. Hierarchical and multi-stage designs such as HMT (He et al., 2025c), LightMem (Fang et al., 2025a), and M2PA (Zhou et al., 2025b) model sensory memory as an initial buffer that retains recent, minimally processed inputs before selection, compression, or consolidation into downstream memory components. Beyond these explicit formulations, implicit sensory memory realizations are more commonly realized in streaming and embodied agents. Systems such as SAM2 (Ravi et al., 2024) and ReSurgSAM2 (Liu et al., 2025d) maintain short-term perceptual queues over recent video frames, while ReKV (Di et al., 2025) and V-Rex (Kim et al., 2025a) rely on streaming KV caches or memory queues to retain recent tokens or visual representations during online inference. Although these mechanisms are rarely labeled as sensory memory, they serve an analogous function by buffering recent observations to support perceptual continuity and computational efficiency, and are typically tightly coupled with working and semantic memory rather than implemented as standalone cognitive modules.

Despite its limited explicit treatment in current LLM agent research, sensory memory is likely to become increasingly important as agents are deployed in multimodal, embodied, and robotic settings. In such environments, agents must process continuous, high-bandwidth sensory streams, such as video frames, audio signals, and proprioception, under real-time and memory constraints (Black et al., 2025; Bjorck et al., 2025; Mon-Williams et al., 2025). In these extensive embodied applications, sensory memory is often instantiated through sensory-level buffering and gating mechanisms, such as short-lived perceptual embedding buffers, attention-driven filtering, and temporal integration, rather than explicitly labeled cognitive modules. These designs reduce redundant computation, stabilize partially observable environments, and support principled consolidation from raw sensory input into working and episodic memory. As a result, more explicit modeling of sensory memory may become a key design for scalable embodied and robotic foundation agents.

3.2.2 Working Memory

Working Memory

Working memory refers to a short-term memory mechanism that supports the temporary storage and active manipulation of information necessary for complex tasks such as reasoning, comprehension, and learning, enabling information to be actively maintained during ongoing operations.

In the LLM-based agent setting, the core goal of working memory (Baddeley, 2020) is to maintain and manipulate task-relevant state under strict online capacity constraints, such that multi-step reasoning and action can proceed without interruption. Since LLMs are inherently stateless, working memory provides the mechanism through which such state is explicitly carried and updated across interaction steps.

Working memory in foundation agents is most commonly instantiated through the active context, which includes the prompt context, intermediate reasoning traces, tool outputs, and runtime states such as key-value caches that are accessible during inference. Within this instantiation, existing approaches differ in where they intervene during execution. One line of work focuses on how task-relevant state is represented before or as it is written into the active context. By compressing (Kang et al., 2025c; Wu et al., 2025d), restructuring (Sun et al., 2025b; Ye et al., 2025b), or abstracting interaction history (Labate et al., 2025), these methods aim to delay or avoid context saturation at the source. Specifically, (Labate et al., 2025) replaces large intermediate outputs with lightweight references, while (Kang et al., 2025c; Wu et al., 2025d; Sun et al., 2025b; Ye et al., 2025b) periodically summarize or fold completed reasoning segments to maintain a compact working context. A second line of work addresses the problem after working state has already accumulated. These approaches study how task-relevant state can be continuously maintained, updated, or evicted under a fixed online budget during execution. They include agent-level policies that explicitly decide memory updates (Zhang et al., 2025n; Yuan et al., 2025a) and system-level mechanisms that manage runtime states such as key-value caches and scheduling (Kim et al., 2025b; Liu et al., 2025j; Kwon et al., 2023; Ni et al., 2025).

In summary, working memory in foundation agents serves as the agent's online workspace, realized through the active context under strict capacity constraints. Rather than treating longer context windows as the sole solution, existing work shows that sustaining coherent long-horizon reasoning depends on selectively retaining and manipulating task-relevant state during execution. This shift reframes working memory from a passive context buffer to an actively managed computational resource. As agents scale to longer horizons and more complex interaction settings, progress in working memory will increasingly hinge on principled state selection and transformation, rather than unbounded context expansion.

3.2.3 Episodic Memory

Episodic Memory

Episodic memory refers to a form of long-term memory dedicated to the persistent storage of an agent's interactive experiences. It records specific events situated in particular temporal and environmental contexts, typically organized as interaction trajectories, action sequences, and associated feedback.

The primary role of episodic memory in foundation agents is to preserve historical interaction contexts and outcomes over extended time horizons, enabling agents to reference past experiences when relevant to ongoing interactions or decision making (Tulving, 1983; 2002). By maintaining access to concrete past experiences, episodic memory supports background reconstruction and cross-session continuity in dynamic environments. This allows agents to ground their behavior in previously observed situations, maintain consistency across repeated interactions, and recover relevant context that would otherwise be lost between sessions.

Episodic memory in foundation agents is most commonly instantiated as an explicit experience repository that accumulates interaction histories across sessions and can be accessed by the agent when needed. From a methodological perspective, existing work on episodic memory in foundation agents can be grouped into two dominant lines of research based on their primary focus. One line of work focuses on episode recording, namely how historical interactions across sessions are organized into coherent episodic records that preserve event structure and situational context. For cross-session interaction histories, (Rajesh et al., 2025) emphasizes selective episode writing and structured organization of past interactions, making what to store and how to organize episodic content explicit. For long video understanding, (Yeo et al., 2025a) constructs episodic records by organizing events together with their temporal and causal relations, enabling coherent episode-level representations of extended visual experiences. More broadly, (Yeo et al., 2025b) represents episodic events at multiple temporal scales, allowing episode formation and access to adapt to different levels of granularity. (Anokhin et al., 2024) links episodic observations with semantic anchors, enabling episodic recall during planning. A second line of work focuses on retrieval and reflection, namely how stored episodes are triggered, selected, and leveraged at decision time. (Yeo et al., 2025b) formulates episodic retrieval as an adaptive process that iteratively selects a memory source and temporal scale conditioned on the query and retrieval history. (Latimer et al., 2025) defines explicit recall and reflection operations that retrieve episodes based on their relevance to the current reasoning context and use them to guide subsequent behavior. In tool-use settings, (Li et al., 2025f) retrieves episodic experience by matching the current execution state against structured representations induced from past trajectories. For multi-session dialogue, (Sarin et al., 2025) triggers episodic retrieval using session-level context and user state cues to recall relevant episodic summaries across sessions. Finally, (Alqithami, 2025) shows that retention policies under fixed memory budgets directly shape which episodes remain retrievable over long horizons.

Overall, episodic memory research in foundation agents centers on preserving concrete interaction experiences across sessions and enabling selective access to those experiences at decision time. Existing work primarily addresses two methodological questions: episode recording and episodic retrieval or reflection. Open problems include how agents should define episode boundaries, regulate the influence of episodic recall during reasoning, and manage long-term retention as episodic memory scales.

3.2.4 Semantic Memory

Semantic Memory

Semantic memory refers to a form of long-term memory dedicated to the storage of abstract facts, general concepts, and structured knowledge. It provides agents with decontextualized information that remains stable over time and can be reused across different situations and objectives.

In foundation agent architectures, semantic memory functions as a stable knowledge base that supports **knowing-what** factual reasoning (Tulving, 1972). Its content is typically derived through the distillation and decontextualization of recurring facts accumulated in episodic memory. When an agent encounters similar knowledge patterns across multiple tasks, fragmented factual information from individual experiences is consolidated into universal concepts, entity relations, or attribute summaries via summarization mechanisms. This semanticization process equips the agent with a knowledge substrate analogous to an encyclopedia or technical manual. By enabling direct access to verified conceptual knowledge rather than repeated retrieval of raw interaction logs, semantic memory provides a coherent and reliable factual foundation for complex reasoning and decision-making. Importantly, in long-horizon agent settings, semantic memory is not static but subject to continual access, revision, and control as new information accumulates.

Existing semantic memory approaches mainly differ in how abstract knowledge is constructed and stabilized, and how it is selected, validated, and maintained as agents reason over long horizons. One line of work focuses on knowledge induction and organization, studying how stable, decontextualized knowledge is distilled from interactions, documents, or external evidence and stored in forms that can be reliably reused by agents. Representative approaches organize semantic knowledge using hierarchical schemas (Zhao et al., 2025a; Jia et al., 2025a), memory graphs (Rasmussen et al., 2025; Wang et al., 2024o), or entry-centric semantic structures (Xu et al., 2025e; Li et al., 2025d), supporting long-term maintenance and structured access during reasoning. Other approaches encode semantic knowledge into neural memory modules (Behrouz et al., 2024; Wang et al., 2024m) or auxiliary parameters (Wang et al., 2025o; Pouransari et al., 2025), emphasizing compact representations and fast reuse without relying on explicit external structures. A second line of work focuses on how semantic knowledge is activated, validated, and updated during agent reasoning as new evidence accumulates. Representative approaches treat semantic access as a decision-time control process in which abstract knowledge is selectively chosen, checked for applicability, and applied to the current reasoning state, rather than retrieved through a fixed similarity lookup (Jimenez Gutierrez et al., 2024; Wang et al., 2025m; Rezazadeh et al., 2025b; Yan et al., 2025b). Other work emphasizes long-term semantic reliability by introducing mechanisms for preference drift detection (Sun et al., 2025a), retention or forgetting policies (Alqithami, 2025), and continual knowledge editing (Wang et al., 2024f) to prevent outdated or inconsistent knowledge from degrading agent behavior.

In summary, semantic memory complements working and episodic memory by providing a stable yet revisable knowledge substrate. It serves as the primary site for distilling fragmented experiences into static laws and facts. By stripping universal knowledge ontologies from specific episodes, it equips the agent with the common-sense foundation necessary for cross-domain tasks. As research progresses, semantic memory is evolving from simple document storage toward self-evolving semantic networks, ensuring that agents maintain an accurate and consistent knowledge system over long-term operation.

3.2.5 Procedural Memory

Procedural Memory

Procedural memory refers to a form of long-term memory dedicated to **how to perform tasks**. It encodes operational skills, execution strategies, and automated routines for specific scenarios. Unlike memory that stores factual knowledge, it abstracts complex action sequences into reusable patterns, enabling agents to complete tasks efficiently and coherently.

In foundation agent architectures, procedural memory is typically reflected in the execution layer, where repeated action sequences, decision rules, or workflows shape how high-level decisions are carried out. Rather than being tied to a single memory substrate, such knowledge may be distilled from prior executions, learned through optimization, or shared across agents. As execution experience accumulates, short-lived action states can be consolidated into reusable skills or routines, allowing agents to execute tasks more consistently over long interaction horizons (Fang et al., 2025b).

The instantiation of procedural memory exhibits an evolution from explicit non-parametric templates toward implicit parametric neural policies through diverse mechanisms. Experience distillation and metacognitive control involve transforming historical trajectories into reusable schemes; for instance, From Experience to Strategy distills interactions into structured strategies (Xia et al., 2025), while Adapting Like Humans focuses on metacognitive routines for error correction at test time (Li et al., 2025h). Learning optimization and policy refinement emphasize the internalizing of skills into parametric weights; for example, Retroformer

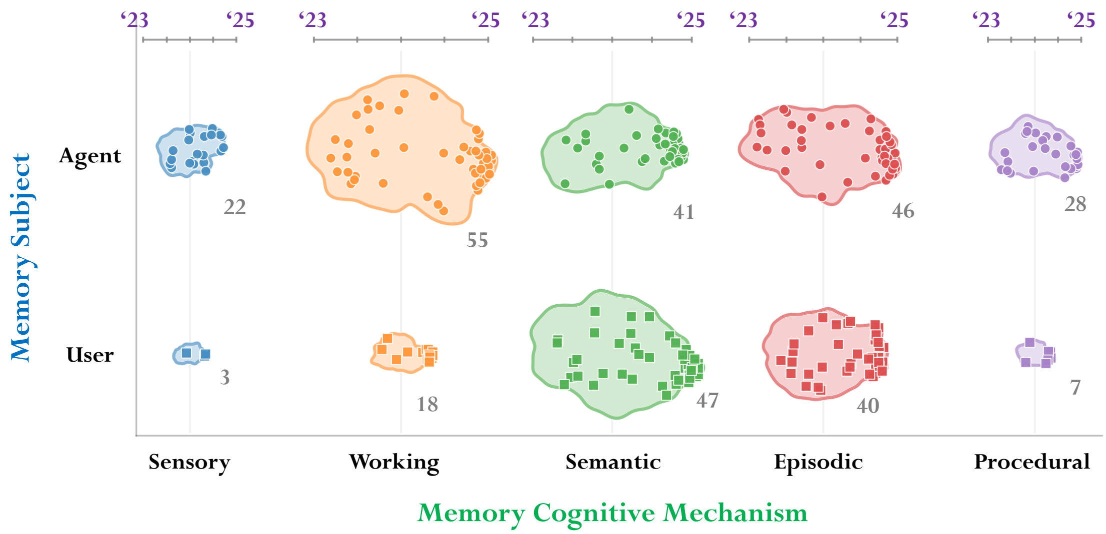

Figure 5: **Connections between memory cognitive mechanism and memory subject.** Each cluster corresponds to number of paper of memory cognitive mechanism (*sensory*, *working*, *semantic*, *episodic*, *procedural*) for agent- or user-centric memory work, and the area size is proportional to the paper number. The scatter distribution represents the publication time for the paper, from 2023 to 2025.

employs policy gradient optimization to refine agent actions (Yao et al., 2024), and BREW enhances task-handling through continuous training and refinement (Kirtania et al., 2025). Multi-agent coordination and shared practices establish common operational norms in collaborative environments, as explored in Smarter Together, MetaGPT, and MIRIX (Tablan et al., 2025; Hong et al., 2023; Wang & Chen, 2025). Workflow externalization and automation transform complex operations into explicit automation templates, as seen in Agent Workflow Memory and LEGOMem (Wang et al., 2025u; Han et al., 2025). Furthermore, self-evolving mechanisms in works like MemGen and ReasoningBank facilitate the accumulation of reasoning traces for long-term procedural evolution (Zhang et al., 2025d; Ouyang et al., 2025). Finally, while work like Evaluating Long-Term Memory signals growing interest in assessment, procedural memory evaluation remains less standardized than fact-centric settings (Terranova et al., 2025).

In summary, procedural memory complements semantic and episodic memory by providing reusable action abstractions. It is currently undergoing a pivotal transition from explicit non-parametric instruction retrieval to implicit parametric neural strategies. This evolution not only supports skill acquisition but also ensures the stable and efficient implementation of complex decisions in long-horizon autonomous agents.

3.3 Memory Subjects

Memory subjects characterize who the memory is primarily modeling and serving for. This dimension is orthogonal to memory substrate and cognitive mechanism, and is critical for understanding the optimization goal of the memory in foundation agents. Figure 5 illustrates how this subject-level distinction connects with cognitive memory mechanisms. Working memory, procedural memory, and sensory memory are predominantly agent-centric, reflecting their roles in supporting many real-world downstream tasks. Semantic and episodic memory appear in both user- and agent-centric settings. For users, they encode stable preferences and interaction histories, whereas in agent-centric memory, they support world modeling and experience accumulation. Together, we may find that memory subjects interleave with cognitive categories, offering a complementary perspective for organizing memory research beyond the single classifications as in prior surveys (Zhang et al., 2025o; Du et al., 2025) alone.

3.3.1 User-Centric Memory

User-Centric Memory

User-centric memory refers to an abstraction of **user-specific facts and preferences**, including biographical data, historical interactions, and expressed preferences, that could be leveraged across sessions and domains, to align with the user for coherent interactions and assistant task execution.

User-centric memory constitutes the foundational section of LLM agents towards user-related, especially personalization tasks (Liu et al., 2025e; Chen et al., 2024b; Zhang et al., 2025m). Beyond general instruction prompts, which only inform temporal task states, user-centric memory underpins the agent's action and response regarding the user identity and grounding user contexts for both short-term and long-term (Tan et al., 2025c). This component is particularly critical in domains, including but not limited to counseling (Litam et al., 2021), recommendation (Chen et al., 2018), and personal healthcare (Velliangiri et al., 2022), where the agent must demonstrate both instant and longitudinal awareness of the user's evolving preference or behavioral traits. Below, rather than imposing a strict taxonomy, we highlight the key objectives in which maintaining accurate and persistent user-centric factual memory is indispensable for enabling robust and coherent user-related tasks.

Memory Management in Dialogues. The dialogue system is one of the most popular real-world application scenarios for user-agent interactions (Chen et al., 2017; Xu et al., 2025e). Within the LLM's finite context window, existing dialogue systems can generally only process multi-turn conversations in limited sessions contexts (Li et al., 2025a). Prompting LLM agents with all details in prior conversations is computationally infeasible and often counterproductive due to noise accumulation and topic drift. Effective user memory thus requires principled mechanisms for relevant conversation context retrieval (Maharana et al., 2024), update (Xu et al., 2025e), summarization (Chhikara et al., 2025; Salama et al., 2025), and forgetting (Zhong et al., 2024). Recent real-world memory systems such as MemGPT (Packer et al., 2023) explicitly formalize memory hierarchies to manage long-term conversational state. In parallel, OpenAI's ChatGPT memory feature implements a persistent memory mechanism that retains user-relevant details while giving users explicit control over what is stored or forgotten (OpenAI, 2025). These developments illustrate a central challenge in dialogue memory management: determining which pieces of information will remain important for future dialogue accumulated during the user-agent conversations, such as stable preferences, life events, or sensitive constraints.

Persistent User Simulation. High-fidelity user simulators are essential for realistic interactive online platforms because they provide scalable, reproducible, and controlled environments for training and evaluation, without accessing extensive real-user data, thereby respecting privacy and reducing costs associated with live user studies and online A/B testing (Park et al., 2023; Zhu et al., 2024; Samuel et al., 2024; Zhang et al., 2025q). In online digital platforms such as recommender systems (Zhu et al., 2025a) and social networks (Gao et al., 2023), simulators help approximate long-term user behaviors and preferences, enabling evaluation of policy optimization, ranking strategies, and interventional impacts under dynamic conditions rather than static test collections (Jin et al., 2013; Zhang et al., 2025k). In such scenarios, user-centric memory support both longitudinal consistency and fine-grained preference evolution, as simulators need to maintain personalized profiles and adaptive interaction patterns over extended horizons.

Long-Term Personalization. Unlike user simulation, which primarily targets group- or platform-level benefits, and dialogue systems, that mainly capture preferences expressed within user-agent conversational contexts, long-term personalization focuses on optimizing the experience of a specific user over extended time horizons spanning days, months, or even years. By maintaining an explicit user profile or persistent personal knowledge base across sessions, agents can adapt linguistic style, decision-making, and behavior in a manner that consistently supports satisfaction of the same individual over time. Early work in long-term personalization (Salemi et al., 2024) generates contextually relevant responses with episodic memory construction. Subsequent research (Tan et al., 2024a) further explores encoding personal memory into parameter-efficient modules, such as LoRA-based parameters (Hu et al., 2022). However, these approaches either struggle to fully exploit rich personal user data or incur substantial computational overhead. More recent frameworks therefore emphasize more efficient memory mechanisms that can reconcile newly observed behaviors with existing memory structures for personalized alignment (Cai et al., 2025; Zhang et al., 2025m).

Privacy-Preserving Memory. Persistent user memory introduces substantial demands in privacy, security, and data governance, because sensitive attributes may be stored, retrieved, and inadvertently exposed through both training-time memorization and inference-time context leakage in memory-augmented agents (Mireshghallah & Li, 2025). Recent work has systematically demonstrated that memory modules in LLM agents are vulnerable to targeted extraction attacks (Wang et al., 2025a), and one can easily recover private user data stored in agent memory under a black-box threat model. In multi-agent settings, privacy risks are further compounded by heterogeneous agent roles and dynamic collaboration, which complicates the enforcement of consistent privacy protocols across interacting memory banks (Shi et al., 2025b). To ensure safe deployment, agents must support selective memory retention, secure storage, user-controlled deletion, and transparent auditing of what information is remembered or forgotten. Practically, it is often combined with differential privacy mechanisms (especially in personalized or federated adaptation), encryption-based storage and retrieval (private vector database), and explicit retention or access-control policies to mitigate leakage risks from both stored content and embeddings Tran et al. (2025); Shi et al. (2025b).

3.3.2 Agent-Centric Memory

Agent-Centric Memory

Agent-centric memory refers to **distilled knowledge, skills, and operational task priors** that an agent accumulates through its own history of task execution or gained via environment interaction. This supports long-context, long-horizon, and long-running tasks across real-world environments.

Unlike user-centric memory preserves personalized information mainly about the corresponding user, agent-centric memory encodes more general lessons and experiences in the agent's own history of solving real-world tasks (Luo et al., 2025; Shinn et al., 2023; Wang et al., 2025c; Zhang et al., 2025i). This memory is agent-centric, capturing lessons that are generally applicable rather than tied to any single user. In essence, it is how an agent "learns from experience and environments", retaining important facts, strategies, and world knowledge gained through previous experiences to improve future performance (Wei et al., 2025e;d; Zhang et al., 2025i). Different from user-centric memory to optimize the particular user satisfaction, agent-centric experience memory cares more about broader real-world problems, where the solution and results are intended to generalize across different users. This is crucial for long-horizon autonomy: an agent tackling complex, multi-step tasks or lifelong learning (Zheng et al., 2025c;b) must be able to remember and build upon what it has encountered before. Below, we outline the key motivations and scenarios that necessitate agent-centric memory approaches.

Long-Horizon Tasks. LLM agents frequently engage in tasks requiring hundreds or thousands of reasoning and action steps, such as coding (Jimenez et al., 2025), web navigation (Zhang et al., 2025l; Zhou et al., 2024), complex multi-turn decision-making Shani et al. (2024), or sequential tool use (Qin et al., 2024). In these settings, immediate working memory is easily overwhelmed by accumulated observations and intermediate reasoning. Agent-centric memory provides an externalized mechanism for storing and retrieving key intermediate states, enabling agents to operate beyond their native context window. For instance, MEM1 (Zhou et al., 2025c) introduces an end-to-end RL framework that maintains a compact internal state, enabling agents to consolidate relevant information while discarding redundant context and thereby operate with near-constant memory usage across arbitrarily long, multi-turn interactions. Complementary to MEM1, MemAgent (Yu et al., 2025b) proposes an RL-based memory agent tailored for long-text processing. It reads long inputs in segmented chunks and uses a fixed-length, overwriteable memory that is updated incrementally. This design enables LLM agents to scale to extremely long contexts with linear complexity and minimal performance degradation. More recent approaches include hierarchical memory modules (Hu et al., 2025b), context folding schemes (Sun et al., 2025b), and learned memory controllers (Zhang et al., 2025n) that decide what to store and when to compress or discard outdated information.

Domain-Specific Long-Tail Solutions. Many real-world problems exhibit long-tail phenomena (Zhang et al., 2021; Kandpal et al., 2023; Park & Tuzhilin, 2008), where rare but important patterns, error cases, or domain-specific heuristics occur infrequently in the training data. Agent-centric experience memory supports the retention of these rare insights and knowledge, enabling agents to reuse them efficiently when similar or related cases arise in the future (Li et al., 2024a). For example, in software debugging (Jimenez et al., 2025), most errors follow common patterns, yet real-world systems often fail due to highly specific configuration issues, dependency conflicts, or environment-dependent bugs. Similarly, in scientific research (Ghafarollahi & Buehler, 2025), while general reasoning patterns and experimental procedures are often shared within a discipline, many sub-area–specific experimental setups (e.g., specialized channel-coding configurations in wireless communications) or highly domain-dependent troubleshooting practices (e.g., field-specific protocols in archaeology) are rarely encountered across researchers, and are therefore unlikely to be sufficiently represented in pretraining data. Comparable long-tail dynamics also arise in complex information search (Wei et al., 2025a) and professional workflow automation (Zhang et al., 2025g), where agents must address narrowly scoped, context-dependent problems that benefit from storing customized one-off solutions and reapplying them over time (Yang et al., 2025a).

Cross-Task Knowledge Transfer. Long-term memory enables agents to accumulate durable knowledge across tasks and episodes, supporting continual improvement without catastrophic forgetting (Hatalis et al., 2023). At a high level, cross-task knowledge transfer concerns how agents abstract and retain task-agnostic cognitive knowledge from diverse interactions, enabling generalization across heterogeneous tasks, domains, and environments rather than improving execution within a single setting or environment. For example, Agent KB (Tang et al., 2025b) constructs a cross-domain experience framework that aggregates high-level strategies and execution lessons distilled from heterogeneous agent trajectories into a shared knowledge base, enabling agents to retrieve and reuse transferable problem-solving knowledge when facing novel tasks across different domains. Unlike specific trajectory replay, the goal of cross-task memory is to distill interaction experience into task-agnostic abstractions that generalize across environments and objectives. Another representative example is the action–thought patterns used across WebShop (Yao et al., 2022) and ALFWorld (Shridhar et al., 2021) in ReAct (Yao et al., 2023). Such abstractions also include reusable tool-use patterns (e.g., stable tool-use strategies (like Toolformer (Schick et al., 2023) and ToolLLM (Qin et al., 2024), and error-avoidance execution heuristics accumulated through repeated failures (Shinn et al., 2023). This enables LLM agents to progressively evolve into more capable and efficient problem solvers, exhibiting behavior that mirrors human-like expertise development across diverse tasks over time.

Strategy and Skill Learning. In contrast to cross-task knowledge transfer, strategy and skill learning for the agent-centric memory focuses on retaining environment-grounded procedural memories that encode how to efficiently execute multi-step actions within a specific interaction regime, such as a web interface, GUI system, or physical environment. For instance, Web agents (Wei et al., 2025e) in environments like WebArena (Zhou et al., 2024) learns how to reuse successful multi-step browsing policies rather than re-exploring interfaces from scratch. For GUI agents, such procedural memory stores interface-specific action sequences, such as menu traversal, widget manipulation, and error recovery strategies in constraints of desktop or mobile environments (Qin et al., 2025; Wang et al., 2025e). In embodied agents, procedural memory manifests as executable skills and control policies that respect physical dynamics and action preconditions (Fung et al., 2025; Yang et al., 2025c). These stored experiences can be reused as templates, demonstrations, or priors for solving tasks more efficiently. Memory-based skill learning allows agents to refine effective behaviors over repeated episodes and internalize world models at the level of action–outcome regularities. This capability is central to long-horizon autonomy and forms the basis for emerging research in lifelong learning (Zheng et al., 2025c), long-running agents (Yang et al., 2025a).

4 Memory Operation Mechanism

4.1 Memory Operations in the Single-Agent System

In a single-agent system, memory operation mechanisms define how a foundation agent actively constructs, updates, controls, and utilizes memory throughout long-horizon interaction and task execution. Rather

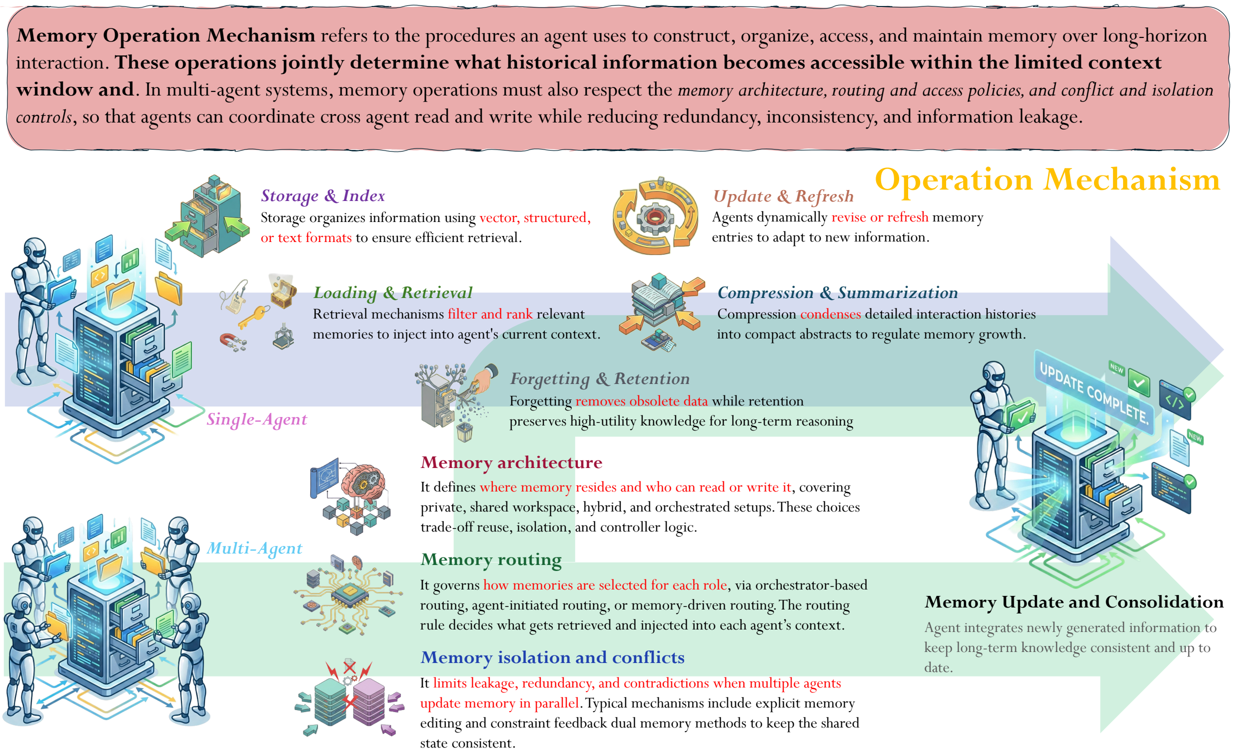

Figure 6: **The Operation Mechanism of Foundation Agent Memory System**. The diagram illustrates the complete operation mechanism of the foundation agent memory system. For single-agent system, it defines five core operations: *storage & index*, *loading & retrieval*, *update & refresh*, *compression & summarization*, and *forgetting & retention*, that govern how historical information is preserved and accessed to support downstream work. For multi-agent systems, the framework extends to address coordination challenges through *memory architecture definitions*, *routing protocols*, and *isolation and conflict resolution strategies*, ensuring data consistency and efficient collaboration across distributed agents.

than treating memory as a static repository, modern agents manipulate memory through a sequence of operations, including indexing, retrieval, updating, compression, summarization, forgetting, and pruning. These operations collectively regulate how past experience is incorporated into ongoing reasoning and future decision-making, forming the operational backbone of single-agent memory systems.

4.1.1 Storage and Index

As an agent's memory grows over time, how information is stored and organized becomes essential for ensuring that relevant information can be efficiently and reliably retrieved when needed. In single-agent systems, memory is typically indexed at write time by associating each entry with semantic embeddings and auxiliary metadata such as timestamps, task identifiers, entities, or tool usage (Lewis et al., 2020; Zhang et al., 2023b; Rezazadeh et al., 2025b). Vector-based storage remains the dominant paradigm in non-parametric agent memory, enabling efficient approximate nearest-neighbor search over episodic or semantic memory in retrieval-augmented agents (Guu et al., 2020; Quinn et al., 2025). Beyond flat vector indices, agents increasingly adopt structured storage formats, including relational tables, graph-based memory, and hierarchical tree representations, which enable relational queries, multi-level abstraction, and schema-aware access aligned with task structure (Anokhin et al., 2024; Sarthi et al., 2024; Rezazadeh et al., 2025b). In parallel, some systems maintain text-record memories that store explicit, human-readable summaries and chronological logs, relying on keyword matching or lightweight string-based retrieval to prioritize transparency and controllability (Park et al., 2023; Zhang et al., 2025o). Finally, storage is not limited to external memory modules: parametric memory embedded in model weights, transient working memory in the context window, and inference-time internal states such as KV caches collectively serve as implicit storage substrates that influence retrieval behavior and reasoning dynamics (Mallen et al., 2023; Kwon et al., 2023; Pope et al., 2023). As memory scales across longer horizons, the choice of storage format and indexing strategy directly affects retrieval precision, computational overhead, and downstream reasoning reliability.

4.1.2 Loading and Retrieval

To utilize stored experience during ongoing reasoning and decision-making, an agent relies on retrieving task-relevant memories while limiting the influence of irrelevant or outdated information. In single-agent systems, this process typically begins with lightweight loading operations that filter or preselect memory entries based on metadata such as recency, task scope, or memory type, followed by similarity-based retrieval over vectorized representations (Lewis et al., 2020; Mallen et al., 2023; Zhang et al., 2025o). Loaded memories are then ranked or refined using semantic similarity, heuristic constraints, or budget-aware selection strategies before being injected into the model's prompt or working context, forming the primary interface between external memory and the LLM's inference process (Park et al., 2023; Wang et al., 2025p). Prior studies indicate that retrieval quality has a substantial impact on agent performance. Retrieving excessive memory can introduce noise and context overload, while overly restrictive retrieval may prevent access to critical historical information (Xu et al., 2025e; Hu et al., 2025c). Consequently, effective loading and retrieval mechanisms aim to balance relevance, diversity, and context budget to support coherent long-horizon behavior.

4.1.3 Updates and Refresh

As agents interact with their environment over extended horizons, previously stored memory may become incomplete, outdated, or misaligned with newly observed information, making static or append-only memory representations insufficient. Update mechanisms enable a single agent to revise existing memory entries in response to new observations, external feedback, or reflective reasoning, allowing memory content to evolve rather than merely accumulate. In practice, updates are often triggered after task completion, evaluation signals, or detected inconsistencies, and may involve rewriting semantic summaries, merging overlapping episodic records, or adjusting the importance of stored information (Park et al., 2023; Tan et al., 2025c; Wang et al., 2023a). In parallel, *refresh* operations focus on adjusting the relative prominence and accessibility of memory without altering its core content, such as re-ranking salient entries, regenerating condensed summaries, or reinforcing frequently accessed memories to preserve their influence over future reasoning. Recent agent systems further demonstrate that reflective or self-evaluative processes can autonomously trigger both updating and refresh actions, leading to improved long-term coherence and task performance (Shinn et al., 2023; Ouyang et al., 2025). Together, these mechanisms allow memory representations to evolve dynamically, supporting adaptation in non-stationary environments while mitigating the accumulation of obsolete or misleading information.

4.1.4 Compression and Summarization

Long-horizon agent memory systems require mechanisms that regulate growth while preserving information essential for future reasoning and decision-making. Compression and summarization address this need by converting fine-grained episodic records into more compact and abstract representations that reduce redundancy and improve memory efficiency. In practice, many agent systems perform summarization periodically or after task completion, consolidating interaction histories into semantic or hierarchical memory suitable for long-term storage (Wang et al., 2025j; Chen et al., 2025d). Hierarchical consolidation further organizes compressed memory into multi-level or tree-structured representations, enabling scalable retrieval across different abstraction levels (Sarthi et al., 2024; Rezazadeh et al., 2025b). Dynamic Cheatsheets further instantiate this idea by maintaining a compact, task-adaptive summary that is continuously updated to surface only the most salient information for ongoing reasoning, reducing repeated large-scale retrieval and context expansion at inference time (Suzgun et al., 2025). While these mechanisms improve context utilization and scalability, prior work highlights an inherent trade-off between abstraction fidelity and long-term recall, making summarization strategies a critical design in single-agent memory systems (Lee et al., 2024; Wu et al., 2025d).

4.1.5 Forgetting and Retention

In agent systems, memory relevance evolves over time as task objectives change and environments become non-stationary, making indiscriminate memory accumulation increasingly misaligned with effective long-term reasoning. Forgetting mechanisms address this challenge by reducing the influence of obsolete or low-utility information through explicitly removing or suppressing memory entries that are no longer aligned with an agent's ongoing objectives. In practice, forgetting is commonly implemented via heuristic policies such as recency-based decay or importance thresholds, as well as learned strategies that optimize memory removal under explicit resource or efficiency constraints (Alla et al., 2025; Wang et al., 2025p). In contrast, retention mechanisms determine which memories should be preserved and prioritized over extended interaction horizons, ensuring that task-relevant knowledge remains accessible despite continual memory growth. Recent work emphasizes adaptive retention strategies, where agents dynamically adjust preservation priorities based on task context, feedback signals, or long-term performance objectives, enabling robust behavior in non-stationary environments (Yan et al., 2025b; Zheng et al., 2025a).

4.2 Memory Mechanism in the Multi-Agent System

In a multi-agent system, memory operation mechanisms describe how several agents build and reuse memory together during collaboration: each agent can keep its own private memory, and they can also share experience through a shared workspace. As in single-agent systems, this mechanism still includes basic operations such as indexing, retrieval, updating, compression, summarization, forgetting, and pruning. Beyond these basic operations, what matters more in multi-agent settings is cross-agent read and write rules: in each task, the system needs to choose suitable memory for agents with different roles. In addition, the system often needs extra operations to remove redundancy, resolve conflicts, and keep the memory consistent.

4.2.1 Memory Architecture

In multi-agent systems, memory can be organized in different ways, varying in which layer it is stored, how read and write permissions are defined across agents, and whether the system introduces a controller. These design choices determine whether routing can construct an efficient memory view and whether systemic issues such as information leakage may arise.

Private-only. In a private-only architecture, each agent has its own independent memory, and its read and write rights only take effect inside that agent's memory. In multi-agent collaboration, agents can only rely on their own memory and the current task context to work. This setup gives strong isolation and makes the system easier to check. However, the same memories are often created again in multiple private spaces, which can waste resources. Representative examples include RecAgent (Wang et al., 2025h), which instantiates one agent per user and keeps a private memory to avoid mixing different users' histories to better protect privacy; TradingGPT (Li et al., 2023b), where each trading agent maintains its own memory so it can follow a consistent risk preference and sector focus, and collaboration mainly happens through exchanging selected viewpoints rather than sharing full memories; and MetaAgents (Li et al., 2025j), which equips each role with an isolated memory of its past thoughts and decisions so the role stays stable and consistent, while any information gained from dialogue is written back locally.

Shared-workspace. Unlike private-only memory, shared-workspace designs use a common pool that all agents can read and write. Agents share intermediate results through this shared state, so they do not need heavy peer-to-peer messaging, which reduces communication cost. However, the shared pool may quickly become noisy and therefore requires filtering mechanisms, as well as coordination strategies to avoid conflicts when multiple agents update the same information. As representative shared-workspace designs, MetaGPT (Hong et al., 2023) uses a simple shared pool to publish role agents' intermediate artifacts, and each agent applies its role profile to filter the pool and pull only relevant memory into context. InteRecAgent (Huang et al., 2025b) makes the shared state more task-specific by establishing a Candidate Bus: tools repeatedly read the current candidate set and write back filtered candidates, so the set is progressively narrowed which can avoid prompt length overflow. MAICC (Jiang et al., 2025d) further scales the workspace into a shared experience pool with offline data and an online replay buffer, where agents query with their current sub-trajectory and retrieve top-k similar trajectories to reuse in-context.

Hybrid. This setup keeps both a private layer and a shared layer. It uses a policy to decide whether new information is written to the private space or to the shared space. At each call, it builds a permission-limited memory view for the agent, so the agent only sees what it is allowed to see. This gives a balance between maximizing memory reuse and keeping sensitive information isolated. For example, in Collaborative Memory (Rezazadeh et al., 2025a), the system separates all memory fragments into private memory and shared memory. During writing, a write policy decides whether the information is generally useful or user-specific, and then writes it to the shared or private layer. During reading, a read policy uses an access graph provided by the system to build a permission-limited memory view for the current call. Similarly, in MirrorMind (Zeng et al., 2025), each Author Agent maintains a private memory corresponding their research interests, while sharing foundational disciplinary knowledge through a public memory. This hybrid design forms the structure of an AI Scientist.

Orchestrated. The previous three categories mainly differ in where memory lives and who can see it. By contrast, orchestrated designs introduce an explicit controller that coordinates agents in a hierarchical workflow and mediates memory access. Concretely, the controller decomposes the task, assigns subtasks to role agents and decides what each agent can read or write. This centralized coordination is well-suited for multi-stage workflows and strongly constrained settings, but it may also introduce a bottleneck and additional system complexity. ChatDev (Qian et al., 2024a) exemplifies this pattern by running role agents under a predefined ChatChain (design, coding, testing), where stage outputs serve as structured handoffs to reduce cross-stage context overload. MIRIX (Wang & Chen, 2025) similarly orchestrates memory maintenance: a Meta Memory Manager routes updates/retrieval to specialized Memory Managers and aggregates their results into a unified response. Notably, this control layer is orthogonal to storage layout and can be combined with private-only, shared-workspace, or hybrid memory stores. For instance, MGA (Cheng et al., 2025) organizes GUI interaction as an observe, plan and ground agent pipeline. A Planner acts as the controller, while lower-level agents can submit intermediate states in a shared workspace, with the planner selecting what history is retrieved and injected at each step. Similarly, AgentFlow also combines orchestration with a shared workspace: a planner controls all modules and writes key intermediate results into a shared memory for later retrieval (Li et al., 2025l).

4.2.2 Memory Routing

Given an architecture, routing describes a set of memory allocation and access rules. When a new task arrives, the system needs to build a separate memory view for agents with different roles: it decides which past memories should be retrieved and how they should be injected into each agent's context. We group routing methods by where the routing decision is made: a central orchestrator, individual agents, or the memory store itself via retrieval and matching.

Orchestrator-based Routing. This refers to a setting where a centralized orchestrator makes routing decisions in a unified way. It maintains the global task state and collaboration progress, breaks a complex goal into subtasks, and then assigns subtasks to different worker agents based on their roles and abilities, while also distributing the required memory and deciding the execution order. These decisions can be updated dynamically as the task state change. This method emphasizes a centralized global workflow, but the cost is that the orchestrator may become a bottleneck for performance and robustness, creating a single point of load and failure risks. For example, in LEGOMem (Han et al., 2025), the orchestrator keeps the global state, generates the next subtask, selects an agent to execute it, and updates the state using the agent's summary; memory is scheduled in the same way, with task memories injected to the orchestrator for planning and subtask memories routed to the selected agent for execution. GameGPT (Chen et al., 2023) shifts the focus to a workflow routing: a manager defines a multi-stage pipeline and requires each stage to write key intermediate outputs into a shared workspace P_t, so later stages can inherit and reuse them. Finally, Westhäußer et al. (2025) extends orchestration to memory-source routing, where the orchestrator using MCP selects which memory sources to call and injects the most relevant snippets; if the injected evidence is insufficient, If the injected information is not enough, the Self-Validator asks for more retrieval and updates the context, enabling centralized memory control.

Agent-Initiated Routing. Compared with orchestrator-based routing, this method routing decisions are not assigned by a centralized orchestrator. Instead, each agent initiates them locally based on its role and task. Information is usually published into a shared memory pool first, and then agents use constraint mechanisms to select the memories they need, forming their own memory views. This method is often more flexible, but it also depends more on good filtering design to avoid noise, conflict, or missing important information in the shared pool. In SRMT (Sagirova et al., 2025), each agent keeps a personal memory vector and decides locally how much to read from shared memory: at each step it uses cross attention over its memory to select from the shared memory sequence and updates its personal memory. Taking a more explicit filtering route, S^3 (Gao et al., 2023) treats shared memory as a platform-wide message stream and scores items with factors such as forgetting, relevance, and source credibility, retaining only a small subset as the agent's own memory view. In the Talker-Reasoner setup (Christakopoulou et al., 2024), a shared store (*mem*) is written by the Reasoner and read by the Talker, and the Talker can choose what to read or wait for a fresh update before replying, showing that agent can decide when to read from memory.

Memory-driven Routing. Different from the above two, routing here is mostly done by retrieval from the memory store. The system represents the current task as a query, and then performs "retrieval, scoring and reranking, selection" in the memory store to obtain the most relevant subset of memories and inject it into the context. Sometimes it can also use structured links between memories (like graph-based expansion) to extend the retrieved results into a more complete set of experience pieces. As a typical form of memory-driven routing, G-Memory (Zhang et al., 2025c) organizes multi-agent histories as a hierarchical graph, retrieves relevant nodes for a new task, expands them through neighborhood links, and compresses the result into a core subgraph before trimming it into role-specific memory views. CRMWeaver (Lai et al., 2025) instead routes at the level of reusable guidelines, it retrieves the most relevant workflow guideline from past successful trajectories and injects it into the current context, and writes back a new guideline when no match exists. Finally, in Spark (Tablan et al., 2025), each coding problem is treated as a query, and a retrieval agent analyzes intent and selects the most relevant documentation and past experience traces from a shared store.

4.2.3 Memory Isolation and Conflicts

Building on multi-agent architecture and memory routing, memory isolation is very important because memory is read and written by multiple agents rather than updated in a single, sequential loop. Multiple agents may produce conclusions in parallel, and their accessible resources and permissions often change over time. The system writes all the information into the same shared pool without any separation and make it visible to every agent, and there might be consistency conflicts: different agents may write facts that contradict each other, and outdated information may not be removed and still retrieved later, which can mislead reasoning. Agents Thinking Fast and Slow (Christakopoulou et al., 2024) reports that the Talker may produce wrong or rushed answers because it can read an outdated belief state before the Reasoner updates it. Here, memory conflicts are mainly handled in two ways: controlling at write time, or gradually improving consistency through an iterative loop.

Write Control for Memory Isolation. A direct strategy is to enforce isolation at the memory write and update stage. In each interaction round, an agent first compares newly extracted candidate facts against the existing memory state and selectively updates memory through a controlled evaluation mechanism, rather than blindly appending new information. In Memory-R1 (Yan et al., 2025b), the memory manager agent is the only agent allowed to mutate memory. This includes four kinds of discrete editing actions, like *ADD*, *UPDATE*, *DELETE*, and *NOOP*. The *ADD* writes a new entry; *UPDATE* rewrites or merges an existing entry (especially when the new information is a refinement or correction of the old one, the system tends to keep the version with more information); *DELETE* removes an entry that is clearly contradicted by new evidence or becomes outdated; and *NOOP* means the information is already covered or is not important for long-term memory, so no update is made. For each user query, the memory manager observes the current memory state and decides which specific memory slots to operate on. As a result, irrelevant memory entries remain untouched. A related form of write-time isolation is adopted in WebCoach (Liu et al., 2025b), where the memory store is updated only after an episode is completed, so partial trajectories are never written into long-term memory, which is isolated and preventing memory conflicts.

Memory Consistency with Feedback Loop. In contrast to write control, this line of work treats memory conflicts as an iterative optimization process for consistency. Generally, the multi-agent memory system enforces the task's hard requirements in a stable constraint memory that later iterations can consult, and it keeps a growing feedback memory that records failures from earlier rounds to support learning in subsequent iterations. EvoMem (Fan et al., 2025b) is a representative example. Its verifier compares each candidate with the stored constraint memory and outputs a score. The system accepts a solution only when the score reaches one hundred.

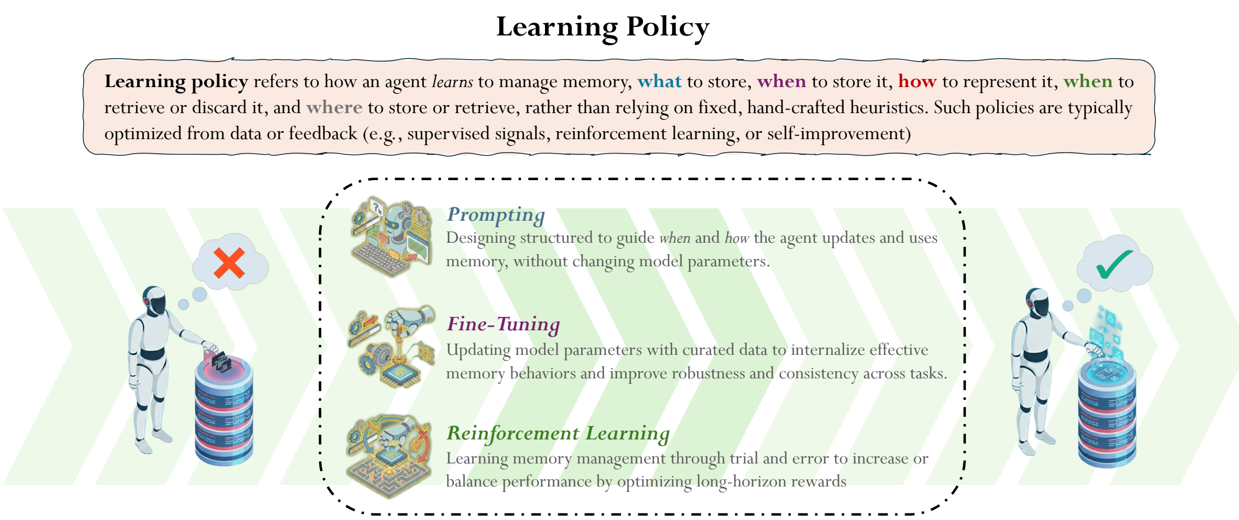

Figure 7: **Learning Policy for Foundation Agent Memory System**. We illustrate how learning policies guide agents in deciding what to store, when to store it, how to represent it, and when and where to retrieve or discard memories. The figure summarizes three common approaches, including *prompting*, *fine-tuning*, and *reinforcement learning*, that progressively improve memory decisions from imprecise memory management toward effective and accurate memory management.

5 Memory Learning Policy

This section examines how agents acquire strategies to perform memory operations (reading, writing, pruning) effectively. Rather than relying on hard-coded heuristics, these policies dictate when and what to remember based on task requirements. We categorize these policies into three distinct paradigms based on the learning signal: prompting, fine-tuning, and reinforcement learning.

5.1 Prompt-based Memory Learning

This paradigm parameterizes the memory policy as natural language prompts. The agent executes these prompts to determine when to access, modify, or prune memory. The primary advantages of this approach include the elimination of expensive model fine-tuning and the high interpretability of the policy. We further categorize this paradigm into Static Prompt-based Control and Prompting Optimization.

5.1.1 Static Prompt-based Control

Static prompt-based control encodes memory policies as fixed, human-designed rules that remain invariant during execution. Memory decisions are specified at design time through prompt templates or predefined schemas, offering strong interpretability and predictable behavior, but lacking the ability to adapt based on interaction feedback or distributional shifts.

Existing work on static prompt-based memory control can be grouped into three design targets: *static memory OS and organization*, where memory is treated as a structured container or a operation system with a fixed form and enforced through hierarchical partitioning, indexing, summarization, or schema-based representations to mitigate long-horizon context degradation (e.g., SCM (Wang et al., 2023a), MemGPT (Packer et al., 2023), LiCoMemory (Huang et al., 2025e), MemoChat (Lu et al., 2023), A-Mem (Xu et al., 2025e), D-SMART (Lei et al., 2025), MemWeaver (Yu et al., 2025d), and Zep (Rasmussen et al., 2025)); *static memory control in single-agent settings*, where access and retention are constrained by persona identities or domain-specific priors encoded in prompts rather than learned relevance signals (e.g., RoleLLM (Wang et al., 2024e), ChatHaruhi (Li et al., 2023a), Mem-PAL, WarAgent (Hua et al., 2023), MemoTime (Tan et al., 2025b), Fin-Mem (Yu et al., 2025e), TradingGPT (Li et al., 2023b), and MemoCRS (Xi et al., 2024)); and *static memory assignment and coordination in multi-agent systems*, where memory is distributed across agents through predefined roles, modular decomposition, and structured communication protocols without learning-based coordination (e.g., MIRIX (Wang & Chen, 2025), LEGOMem (Han et al., 2025), G-Memory (Zhang et al., 2025c), MADRA (Wang et al., 2025g), GameGPT (Chen et al., 2023), and ChatDev (Qian et al., 2024a)).

5.1.2 Dynamic Prompt-based Control

Dynamic prompt-based control explores whether memory policies encoded in natural language prompts can be adapted at test time based on experience and feedback, without updating model parameters. Rather than fixing memory behavior at design time, this paradigm treats memory control as a language-mediated and continually revisable process.

Existing work in this space centers on a set of closely related research questions. One line of work asks whether memory usage policies can be corrected through reflection on past outcomes, prompting agents to analyze failures or successes and convert these insights into revised memory instructions that guide future behavior (e.g., Reflexion (Shinn et al., 2023), ReasoningBank (Ouyang et al., 2025), WebCoach (Liu et al., 2025b), and QuantAgent (Wang et al., 2024g)). Another line investigates whether memory representations themselves can be dynamically optimized to improve information efficiency under limited context budgets, treating compression, denoising, and structural reorganization as adaptive, prompt-driven processes (e.g., ACON (Kang et al., 2025c), ACE, SeCom (Pan et al., 2025), Nemori (Nan et al., 2025), CAM (Li et al., 2025d), EvoMem (Fan et al., 2025b), and ViLoMem (Bo et al., 2025)). A further question concerns whether dynamic prompting can distill accumulated experiences into reusable procedural knowledge, such as reasoning templates, execution scripts, or tool-usage strategies that generalize beyond episodic recall (e.g., BoT (Yang et al., 2024b), Memp (Fang et al., 2025b), and ToolMem (Xiao et al., 2025)). Despite their adaptivity, these methods remain fundamentally language-mediated and lack explicit credit assignment, limiting their capacity for long-term policy optimization compared to fine-tuning and reinforcement learning–based approaches.

5.2 Fine-tuning: Parameterized Memory Policies

Beyond prompt-based adaptation, supervised fine-tuning (SFT) internalizes memory policies into model parameters, enabling more stable and reusable memory behaviors. From a policy learning perspective, SFT-based approaches investigate how memory policies are internalized, stabilized, and executed efficiently once embedded into model weights.

5.2.1 Policy Internalization into Parameters

A defining characteristic of SFT-based memory control is that memory policies are internalized into model parameters, transforming memory from an external context manipulation problem into a parametric policy representation. Rather than relying on prompts or explicit buffers at inference time, these approaches embed memory-related behaviors directly into the weight space, enabling stable and reusable memory usage across tasks.

Within this paradigm, existing work explores several closely related research questions concerning how memory control policies can be embedded into model parameters and how such internalization should be structured. Some approaches focus on internalizing memory content itself, distilling short-term contextual information into long-term parametric representations (Wang et al., 2024m; 2025o)). Others emphasize internalizing memory access and retrieval behaviors by learning parameterized interfaces or lightweight modules that mediate interaction with external or structured memory stores, rather than absorbing raw content into weights (e.g., Memory3 (Yang et al., 2024a) and MLP Memory (Wei et al., 2025c)). A related line of work investigates how such parameterized memory policies can be organized hierarchically, enabling scalable invocation of large memory components while decoupling long-tail knowledge from core reasoning abilities (Pouransari et al., 2025). Despite their differences, these approaches share the objective of shifting memory control from prompt-level manipulation to parameterized decision rules learned through supervision.

5.2.2 Parameterized Policy Stabilization and Boundary Control

Beyond internalization, SFT enables the stabilization of memory policies by learning explicit boundaries on what should be written, corrected, or suppressed once memory control is embedded into parameters. Rather than merely expanding memory capacity, these approaches aim to prevent error accumulation, concept drift, and persona inconsistency under long-term use.

A common theme across these works is the use of supervision to regularize memory updates and enforce boundary constraints. Some approaches train models to perform reflection or self-analysis before committing information to memory, encouraging the storage of high-level, noise-resistant representations aligned with task intent (Liu et al., 2023b; Zhang et al., 2025p; Chen et al., 2025d)). Others emphasize identifying and repairing erroneous or outdated memory by learning when existing knowledge should be revised or overridden through correction signals, verifier feedback, or routing mechanisms (e.g., WISE (Wang et al., 2024f), SuperIntelliAgent (Lin et al., 2025), CRMWeaver (Lai et al., 2025)). In long-horizon interactive settings, defensive boundary control further constrains memory updates to preserve role or identity consistency by restricting which experiences can be retained or reused, like Character-LLM (Shao et al., 2023). Collectively, these methods treat reflection and self-correction not as isolated prompt-level techniques, but as mechanisms for learning stable and bounded memory policies through supervision.

5.2.3 Parameterized Policy Efficiency and Retrieval Refinement

Beyond learning what to store and how to stabilize memory, SFT is also used to refine how parameterized memory policies are executed at inference time, particularly during memory reading and retrieval. Rather than relying on exhaustive context access, these approaches treat retrieval itself as a learning policy and optimize how queries are formulated, iteratively refined, and applied to compressed memory representations. Through supervised training, models learn to generate precise retrieval cues for targeted memory access (Qian et al., 2025b), to execute multi-hop or progressive retrieval that refines queries across reasoning steps (Ko et al., 2024), and to optimize compression-aware retrieval by internalizing or reversibly refining memory representations (Cao et al., 2025b; Wang et al., 2025m), thereby improving reasoning robustness while reducing inference-time overhead. Despite these gains, the resulting retrieval policies remain fixed after training and do not incorporate explicit credit assignment over extended decision horizons.

5.3 Reinforcement Learning for Memory Policies

Reinforcement learning introduces a fundamentally different paradigm for memory control by enabling memory policies to be optimized through interaction and reward feedback. Unlike prompt-based or supervised approaches, RL allows downstream task outcomes to influence earlier memory-related decisions, making memory construction itself a learnable policy. Existing work can be understood as progressively extending the temporal scope over which reinforcement signals shape memory behavior.

5.3.1 Step-Level Memory Decisions

At the shortest temporal scope, reinforcement learning is applied to memory control by treating memory management as a sequence of step-level decisions. In this setting, memory operations are modeled as actions selected by a learning policy and optimized based on their immediate or short-horizon impact on task reward.

One line of work studies how memory editing can be formalized as an explicit action space. Memory-R1 (Yan et al., 2025b) defines atomic memory operations such as adding, updating, deleting, or skipping entries, and learns a memory policy that selects among these actions based on task-level rewards. MemAct (Zhang et al., 2025n) extends this formulation by incorporating finer-grained editing actions, including trimming and summarization, directly into the agent's unified policy space. A closely related problem concerns step-level memory decisions under explicit capacity constraints. MemAgent (Yu et al., 2025b) and RMM (Tan et al., 2025c) address this setting by learning, through interaction-driven feedback, which information should be written into a fixed-size memory buffer when processing extremely long contexts. Mem-α (Wang et al., 2025p) generalizes this paradigm by framing memory construction itself as a sequential decision-making problem, where agents learn, via reinforcement learning, how to populate and update structured multi-component memories (e.g., core, semantic, and episodic memory) to maximize long-horizon task performance. Together, these approaches frame step-level memory control as the optimization of local memory actions through reinforcement signals, without explicitly modeling the long-term effects of memory state construction.

5.3.2 Trajectory-Level Memory Representation

As tasks extend over longer horizons, the value of memory decisions often emerges only through their cumulative influence on future reasoning and action selection. Reinforcement learning enables this setting by allowing delayed task outcomes to shape how trajectory-level memory states are constructed, updated, and maintained by a learning policy.

Within this scope, existing work studies how compact, decision-sufficient memory representations can be learned when interaction histories can no longer be evaluated step by step. Rather than treating memory as a transient buffer, these approaches view trajectory-level memory as part of the agent's Markov state, whose quality is assessed through downstream decision performance (Chen et al., 2025c; Wu et al., 2025d). A closely related question concerns how long interaction histories should be abstracted into such representations. Several studies treat summarization, folding, or compression as policy decisions whose effectiveness can only be evaluated through reinforcement signals propagated from future outcomes (Lu et al., 2025b; Sun et al., 2025b; Li et al., 2025g). Trajectory-level memory also raises the issue of how memory states should evolve over time as new interactions unfold. MemSearcher (Yuan et al., 2025a) addresses this problem by maintaining an iteratively updated compact memory state and propagating advantages across contexts to refine memory representations. Together, these works characterize trajectory-level memory as a learned state representation whose utility is defined by its long-term contribution to decision making under reinforcement learning.

5.3.3 Cross-Episode and Multi-Agent Memory

When memory extends beyond individual trajectories, it no longer serves only immediate reasoning but accumulates experience whose value emerges across repeated episodes or interactions. At this scope, reinforcement learning is essential, as only long-term and cross-episode reward signals can determine which memories should persist, adapt, or be revised by the memory policy.

Research at this level focuses on how experience should be represented, reused, and coordinated once memory spans multiple episodes or agents. Rather than preserving raw interaction histories, cross-episode memory aims to distill higher-level decision-relevant knowledge, such as reusable strategies, self-correction rules, or abstracted behavioral patterns, whose utility is evaluated through repeated reinforcement signals. This perspective underlies approaches such as MCTR (Li et al., 2025h) and graph-based experience abstraction (Xia et al., 2025), which encode experience as transferable decision knowledge learned through interaction. Crucially, such experience is retrieved and applied in a context-dependent manner governed by reinforcement learning, as exemplified by reflective retrieval policies in Retroformer (Yao et al., 2024) and Memento (Zhou et al., 2025a). As memory further extends across agents or representation spaces, reinforcement learning enables feedback to propagate beyond individual trajectories. This includes latent or non-textual memory representations such as MemGen (Zhang et al., 2025d), retrieval-path optimization and callback mechanisms in SUMER (Sumers et al., 2023) and ReMemR1 (Shi et al., 2025a), as well as shared or decentralized memory policies in multi-agent systems such as MAICC (Jiang et al., 2025d) and SRMT (Sagirova et al., 2025). These studies frame cross-episode and multi-agent memory as the broadest scope of reinforcement-learning-based memory control, where memory policies evolve through accumulated interaction and reward feedback rather than predefined rules or one-shot supervision.

6 Scaling: Memory, Contexts, and Environments

As LLMs and agents are deployed in increasingly realistic settings, their contexts grow rapidly and can explode along three scaling axes during interactions with the open-world environments: interaction horizon, environmental complexity, and system quantity. While traditional evaluations often rely on static, context-limited settings that overlook environmental dynamics, real-world utility requires the ability to accumulate, retain, and update knowledge across extended timeframes and heterogeneous data structures. This section explores how memory emerges as the essential architectural solution to this scaling challenge, transforming from a simple interaction log into a sophisticated context-management system that enables robust skill learning, long-term personalization, and collective coordination in multi-agent ecosystems.

6.1 Context-Limited Simple Environment

The majority of LLM and agent benchmarks today are still configured in *context-limited simple environments*, where an agent is placed in a compact and closed world, interacts for only a short horizon task instance (Hendrycks et al., 2021b) or synthetic non-real users (Maharana et al., 2024). While such settings facilitate reproducibility and experimental comparison, they substantially under-specify the memory demands faced by real-world agents. Therefore, strong benchmark performance often reflects proficiency in in-context pattern matching or short-term reasoning, rather than the ability to accumulate, retain, and reuse knowledge across extended interactions and evolving preferences of user contexts. This mismatch leads to a notable utility gap: agents that achieve high scores on existing benchmarks frequently fail to exhibit long-term adaptation, user personalization, or task-specific skill reuse in open-ended, dynamic environments.

A substantial portion of existing LLM and agent evaluations remains confined to static, context-limited environments. Classic general question answering benchmarks, including SQuAD (Rajpurkar et al., 2018), HotpotQA (Yang et al., 2018), and KILT (Petroni et al., 2021), operate over frozen Wikipedia snapshots, while MS MARCO (Craswell et al., 2021), Natural Questions (Kwiatkowski et al., 2019), SearchQA (Dunn et al., 2017), and TriviaQA (Joshi et al., 2017) replace live information access with pre-collected queries and passages. These designs yield stable, bounded, and well-controlled information sources under which modern systems achieve near-saturating performance. However, retrieval and reasoning are strictly episodic and instance-isolated: agents neither maintain cross-query state nor confront temporal drift, source inconsistency, or evolving knowledge. As a result, such benchmarks impose minimal requirements on persistent memory, primarily testing short-term retrieval and in-context reasoning rather than long-term knowledge accumulation or adaptive context management. Similar limitations extend to tool-use and web interaction benchmarks. Frameworks such as ToolBench (Qin et al., 2024), WebShop (Yao et al., 2022), and WorkArena (Drouin et al., 2024) rely on fixed APIs or self-hosted environments that abstract away authentication, failure recovery, and long-term user or task state. Even systems that interact with the live web, including WebGPT (Nakano et al., 2021) and WebVoyager (He et al., 2024a), remain constrained by curated site lists and strict interaction budgets. Sequential and interactive environments such as ScienceWorld (Wang et al., 2022), ALFWorld (Shridhar et al., 2021), TextWorld (Côté et al., 2018), and Jericho (Hausknecht et al., 2020), along with programming benchmarks like APPS (Hendrycks et al., 2021a), SWE-bench (Jimenez et al., 2025), SWE-Bench Pro (Deng et al., 2025), and SWE-Lancer (Miserendino et al., 2025), increase task complexity but remain episodic, reset-centric, and evaluation-driven. Consequently, these benchmarks systematically fail to assess knowledge accumulation across episodes and long-horizon consistency, thereby obscuring the critical role of memory for robust real-world agent deployment.

6.2 Context-Exploded Real-World Environments

Unlike context-limited benchmark evaluations in research works, real-world deployments expose agents to environments where context scales along multiple axes. It accumulates over long interaction horizons, becomes increasingly complex due to structured and heterogeneous system state, and spans multiple environments, agents, and tools.

6.2.1 Scaling with Interactions

User–Agent Interactions Memory over Extended Horizons. In persistent user–agent interactions, agents are required to maintain coherent behavior and decision consistency aligned with the user across extended dialogue horizons. Unlike short, self-contained dialogues, information introduced in early rounds, such as user preferences, identity attributes, or implicit task constraints, may not immediately influence responses but often becomes decisive in later stages of interaction. Therefore, interaction history functions less as static input and more as a latent internal state that evolves over time. This requirement for information persistence arises from the stability of user personas and preferences across different turns (Li et al., 2016), the long-term structure of dialogue tasks involving planning or sustained goals, and the importance of identity and goal consistency for user trust and usability (Zhang et al., 2018). However, under fixed context window constraints, long-horizon dialogue systems frequently exhibit failure modes such as early-context forgetting and progressive context drift as interaction length increases (Liang et al., 2020). These failures are not isolated reasoning errors but the cumulative consequence of long-term context mismanagement. Existing mitigation strategies, such as sliding context windows, heuristic truncation, and summary-based memory mechanisms (Park et al., 2023), as well as explicit long-term structures like user profiles or persona memories (Xu et al., 2022a), improve scalability but often degrade recall reliability. In practice, seemingly peripheral information may be irreversibly discarded despite its potential future relevance, exposing a fundamental trade-off between context budget control and long-term information accessibility (Liu et al., 2024b).

Accumulated Memory for Multi-Turn Tool Use. Context explosion can be exacerbated in agents that rely on multi-turn tool use and reasoning architectures. Beyond conversational history, tool-based agents must retain tool inputs, execution outputs, and intermediate states, many of which are repeatedly referenced in subsequent reasoning steps (Schick et al., 2023). In frameworks such as ReAct (Yao et al., 2023) and Planner–Executor architectures (Wang et al., 2023b), this accumulation is particularly severe: planning traces, tool feedback, and reflective reasoning are explicitly preserved to maintain coherence and decision consistency (Shinn et al., 2023). Consequently, context size grows linearly or even exponentially with interaction length. In ReAct-style agents, explicit reasoning traces themselves become part of the context, enabling interpretability and complex reasoning but simultaneously introducing a substantial and persistent contextual burden. Additionally, for many real applications, the contexts of tool outputs like the web search (Wei et al., 2025e) and database query (Jing et al., 2025) can also accumulate dramatically. Naively removing or compressing these traces risks breaking causal dependencies between reasoning and action steps and undermining subsequent decisions (Yao et al., 2023). Thus, while multi-turn tool use substantially enhances agent capability, it also exposes scalability limits on finite context windows, underscoring a core challenge for long-horizon agent design (Liu et al., 2024b).

6.2.2 Scaling with Environment Complexity

As environments scale in complexity, an agent's context must accommodate heterogeneous data modalities, asynchronous tool interactions, and protocol- or permission-constrained external state. In such settings, context can no longer be treated as a linear interaction trace appended to a prompt. Real-world deployments require reasoning over structured artifacts such as API responses, files, databases, logs, and configuration states, whose semantics depend on schemas, provenance, update rules, and temporal validity. Naively flattening these artifacts into token sequences is both inefficient and structurally lossy, undermining interpretability, precise retrieval, and targeted updates (Modarressi et al., 2024). Consequently, increasing environmental complexity shifts memory from an implicit byproduct of prompt accumulation to an explicit, system-level component responsible for structured context management. The memory challenge thus transitions from merely retaining past information to maintaining coherent, queryable, and updatable representations of environment state across diverse sources and lifetimes.

Recent agent systems address this challenge by externalizing memory beyond the prompt and exposing explicit read–write interfaces that decouple storage from reasoning (cauri, 2025). Externalized memory enables schema-aware retrieval, versioning, targeted edits, and access control—operations that are difficult or infeasible within prompt-based context alone (Yakobi & Sadon, 2025). Protocol-based interfaces such as the Model Context Protocol (Anthropic, 2024) and skill-oriented abstractions like Claude Agent Skills (Anthropic, 2025) further formalize how agents interact with external state by defining query semantics, update policies, and scope boundaries, thereby improving modularity, auditability, and behavioral safety (Keen, 2025). As environmental complexity increases, persistence becomes unavoidable: user models, system configurations, and task artifacts represent durable state that must remain consistent across sessions while respecting governance constraints such as privacy, permissions, and rollback (Sarin et al., 2025; Wu & Shu, 2025). These demands introduce additional challenges, including schema drift, concurrent updates, and path-dependent evaluation, underscoring the need for principled abstractions that treat memory as a first-class infrastructure for context management in complex, open-world environments.

6.2.3 Scaling with Environment Quantity

Memory as Interfaces of Tool Environments. In open-world settings, context complexity does not only from prolonged interaction within a single environment, but also from the need to operate across multiple heterogeneous tool environments, each defined by distinct state records and access protocols. For example, a personal embodied assistant may alternate between a physical household environment, where it performs long-horizon behaviors grounded in perceptual feedback, and a digital web environment, where it retrieves information such as weather forecasts through browser-based interactions. Supporting such behavior requires memory mechanisms that can preserve, differentiate, and reconcile environment-specific states across diverse action and observation modalities (Glocker et al., 2025; Hong et al., 2025). In OSWorld (Xie et al., 2024), a GUI agent may issue search queries in a browser, retrieve results from a news application, and consult documents in a file viewer, with each interaction producing environment-specific observations and state transitions. While tools enable localized interaction within each environment, memory provides the interface function that persists, organizes, and contextualizes information across environment boundaries. As a result, memory systems are required not merely to log past tool calls, but to maintain structured and separated representations of environment state that support long-horizon reasoning without breaking the context window (Burtsev et al., 2020; Rae et al., 2019).

Memory in Multi-Human-Agent Systems. As agent systems scale to include multiple agents and human participants, effective coordination increasingly depends on structured communication and shared memory mechanisms rather than naïve context sharing, which quickly fragments under limited context windows (Chen et al., 2025f). Recent approaches introduce agent-aligned or semi-shared memory abstractions that encode relational histories, inter-agent dependencies, and task-relevant state across long interaction horizons. For example, Intrinsic Memory Agents equip agents with role-specific memory templates that preserve specialized perspectives while enabling integration into a shared contextual substrate, substantially improving long-horizon planning stability (Yuen et al., 2025). Beyond memory, structured communication protocols play a central role in coordination: hierarchical and role-aware dialogue schemes reduce noise and bias in inter-agent exchanges (Wang et al., 2025r), while cognitively adaptive orchestration frameworks dynamically adjust communication patterns based on inferred collaborator states (Zhang et al., 2025h). In open-world deployments involving multiple humans and agents, coordination further extends to alignment, conflict resolution, and collective decision-making under disagreement. Empirical studies show that debate-based protocols and decision rules such as consensus or majority voting significantly influence performance across reasoning and knowledge tasks (Kaesberg et al., 2025; Samanta et al., 2025), while adaptive and consensus-free debate mechanisms balance computational cost, robustness, and conformity effects (Fan et al., 2025a; Cui et al., 2025b). As these systems scale, implicit graph-structured representations increasingly underpin orchestration and communication, organizing relational dependencies without explicit symbolic graphs and reframing context management as a problem of distributed memory, coordination, and alignment in open-world systems (Qian et al., 2025a; Zhang et al., 2025f;e).

7 Evaluation

Evaluating foundation agent memory is fundamentally about measuring whether stored information and experience are accurate, useful, reliable, and efficiently accessible under long-horizon interactions. In the following, we summarize commonly used metrics in three different basic categories, including *accuracy-based*, *similarity-based*, and *LLM-as-a-judge*, in Section 7.1. In addition, we collect the commonly used benchmark to assess the foundation agent's performance, as shown in Section 7.2.

7.1 Metrics

Table 2 summarizes metrics most commonly used in foundation agent memory evaluations. There are multiple dimensions to evaluate foundation agents and their memory module performance. Some tasks have clear ground truth answers and can be scored with exact correctness (Du et al., 2024; Dunn et al., 2017), while others are open-ended (dialogue (Budzianowski et al., 2018), summarization (Maharana et al., 2024), preference following (Zhao et al., 2025c)) where multiple outputs are acceptable and reference-based scoring is not reliable. As a result, existing work typically combines outcome-level correctness with retrieval-oriented assessment and judge-based rubrics, depending on whether the benchmark exposes memory module, uses long-context prompting, or evaluates agents acting in an specific environment, scenario, or task (Patlan et al., 2025; Yadav et al., 2025).

Accuracy-based Metrics. When tasks have a clear objective outcome, accuracy-based metrics are used. For answering questions about long histories, accuracy or memory accuracy directly assesses the alignment of the final response with the ground truth answer (Yang et al., 2018; Zhong et al., 2024; Du et al., 2024). F1 eases exact matching by giving credit for partial overlap at the token level. This is especially prevalent when responses do not match exactly but have some variance (Trivedi et al., 2022; Deng et al., 2023). When the benchmark evaluates an explicit memory module, Recall@K, Mean Average Precision (MAP), and Normalized Discounted Cumulative Gain (NDCG)@K become central because they separate "did the system retrieve the right evidence" from "did the generator phrase the answer well." Recall@K checks to see if at least one relevant item is in the top-k retrieved results (Wu et al., 2025b). MAP and NDCG@K, on the other hand, also take into account the quality of the ranking and give systems that put relevant memories first a higher score (Kohar & Krishnan, 2025). For interactive agents, correctness is instead defined by environment evaluators, so Success Rate (SR) or Goal Completion (GC) represents whether the agent finishes tasks end-to-end (Zhou et al., 2024; Zheng et al., 2025a). For code and tool-use settings, Pass@K and Resolved Rate (RR) measure whether at least one of the top-k attempts solves the task or resolves an issue (Yao et al., 2025; Jimenez et al., 2025). In addition, memory-centric benchmarks increasingly add failure-mode assessment that are hard to capture with end-task accuracy alone. Memory Integrity quantifies whether extracted memories cover the required memory points, and False Memory Rate measures how often systems introduce fabricated or incorrect memories during storage, update, or use (Chen et al., 2025a).

Similarity-based Metrics. Similarity-based metrics are most prevalent in dialogue generation and summarization, where the output is free-form, and correctness cannot be fully checked by the accuracy or F1 score (Chen et al., 2024d). BLEU, ROUGE, and ROUGE-L measure lexical overlap with a reference, which is useful for tracking surface similarity but can underestimate valid paraphrases and overestimate fluent yet ungrounded responses (Gehrmann et al., 2023; Ai et al., 2025). Distinct-n complements overlap metrics by measuring lexical diversity, discouraging repetitive generations that can inflate similarity scores without improving faithfulness (He et al., 2025a). When lexical overlap is too strict, BERTScore provides an embedding-based approximation of semantic similarity (Maharana et al., 2024), and FactScore evaluates memory faithfulness by checking agreement at the level of atomic factual claims (Xu et al., 2024b), which is particularly relevant when summarization is used as a compression mechanism for long histories (Saxena et al., 2025). Perplexity is also used in some long dialogue benchmarks as a likelihood-based metric for generation quality over sessions (Xu et al., 2022a), but it remains an indirect indicator for memory performance, because it does not verify whether the model's content is grounded in the correct historical evidence (Durmus et al., 2022).

Table 2: **Metrics used in Foundation Agent Memory Evaluations**.

Metric	Short Description	Representative Benchmark(s)
Accuracy-Based Metrics		
Accuracy / Memory Accuracy	Proportion of instances answered correctly, computed at the benchmark granularity (e.g., question-, turn-, session-, or task-level).	HotpotQA (Yang et al., 2018), PerLTQA (Du et al., 2024), MemoryBank (Zhong et al., 2024)
F1 Score	Harmonic mean of precision and recall, computed at token level and optionally aggregated to question, turn, or session level.	MuSiQue (Trivedi et al., 2022), MSC (Xu et al., 2022a), PerLTQA (Du et al., 2024)
Recall@K	Evaluates retrieval success by checking if relevant evidence appears within the top-k results.	LongMemEval (Wu et al., 2025b), PerLTQA (Du et al., 2024)
Mean Average Precision (MAP)	Averages precision at ranks of relevant items, then averages across queries.	PerLTQA (Du et al., 2024), PMR (Kohar & Krishnan, 2025)
NDCG@K	Measures ranked relevance at cutoff K by prioritizing top positions.	LongMemEval (Wu et al., 2025b), PMR (Kohar & Krishnan, 2025)
Success Rate (SR) / Goal Completion (GC)	Fraction of interactive tasks completed under environment-defined success checkers, typically measured at the task or episode level.	WebArena (Zhou et al., 2024), OSWorld (Xie et al., 2024), MineDojo (Fan et al., 2022)
Pass@K / Resolved Rate (RR)	Probability a correct solution appears among top-k samples (Pass@K), or issues resolved end-to-end (RR).	HumanEval (Chen et al., 2021), SWE-Bench (Jimenez et al., 2025)
Memory Integrity (MI)	Completeness of memory extraction, measured by coverage or recall over memory points.	HaluMem (Chen et al., 2025a)
False Memory Rate (FMR)	Rate of introducing hallucinated memories, including fabricated or incorrect updates.	HaluMem (Chen et al., 2025a)
Similarity-Based Metrics		
ROUGE	Measures overlap between generated and reference text using ROUGE. It is widely used for summarization.	LoCoMo (Maharana et al., 2024), MemoryBench (Ai et al., 2025)
BLEU	n-gram precision overlap with reference. It is commonly used for dialogue generation.	DuLeMon (Xu et al., 2022b), LoCoMo (Maharana et al., 2024)
Distinct-n	Calculates the ratio of unique n-grams to measure lexical diversity and discourage repetition.	DuLeMon (Xu et al., 2022b), MADial-Bench (He et al., 2025a)
BERTScore	Embedding-based semantic similarity between candidate and reference.	LoCoMo (Maharana et al., 2024), MADial-Bench (He et al., 2025a)
FactScore	Fact-level faithfulness: extracts atomic claims and measures the fraction supported by retrieved evidence.	LoCoMo (Maharana et al., 2024), Face4Rag (Xu et al., 2024b)
Perplexity	Likelihood-based metric of predicting reference text, typically token-level and aggregated over sessions.	MSC (Xu et al., 2022a), DuLeMon (Xu et al., 2022b)
LLM-as-a-Judge Metrics (JUDGE)		
Response Correctness	A strong LLM judges whether the response answers the query or satisfies constraints.	LongMemEval (Wu et al., 2025b), MemTrack (Deshpande et al., 2025)
Faithfulness / Groundedness	Judge checks that response claims are grounded in retrieved context or memory, or check whether the citations support the claims.	SeekBench (Shao et al., 2025), LiveResearchBench (Wang et al., 2025f)
Preference Following	Judge evaluates preference following by checking whether the output satisfies user-stated preferences or constraints.	PrefEval (Zhao et al., 2025c), BEAM (Tavakoli et al., 2025), ConvoMem (Pakhomov et al., 2025)

LLM-as-a-judge Metrics. LLM-as-a-judge metrics are used when ground truth answers, or references, are incomplete, when multiple responses are acceptable, or when evaluation requires a rubric that is hard to encode as string matching (Yu et al., 2025a). Response Correctness asks a strong model to decide whether the answer satisfies the user query (Deshpande et al., 2025), which is convenient for open-ended responses but introduces judge dependence and sensitivity to prompting. Faithfulness or Groundedness uses a judge to verify that claims are supported by retrieved context or memory (and in some settings, whether provided citations support the response) (Shao et al., 2025; Wang et al., 2025f), which helps distinguish helpful but hallucinated answers from evidence-supported ones. Preference Following uses a judge to determine whether the output respects explicit user constraints or stated preferences (Zhao et al., 2025c; Tavakoli et al., 2025; Pakhomov et al., 2025), which is essential for personalization benchmarks where correctness is defined by user alignment rather than a single factual label (Wang et al., 2024b).

Across these metrics, a practical implication is that evaluation becomes more attributable to memory mechanisms when the benchmark can separate retrieval or selection from generation (Chen et al., 2025a). Retrieval and ranking metrics (Recall@K/MAP/NDCG@K) assess whether the memory interface surfaces the right items, while integrity and hallucination-oriented metrics (MI/FMR) expose failure modes that may not change end-task accuracy until they accumulate (Zhang et al., 2025a). In contrast, similarity-based metrics remain useful for tracking fluency and summarization quality, but they should be paired with grounding checks (FactScore or judge-based faithfulness) to avoid rewarding ungrounded paraphrases (Aralikatte et al., 2021). In addition, some benchmarks also emphasize cost and feasibility (e.g., capacity and efficiency) (Tan et al., 2025a), motivating reporting not only what the agent remembers, but also the computational and storage trade-offs required to achieve that performance.

7.2 Benchmarks

To assess the memory improvement on the foundation agent task in diverse scenarios, we categorize existing benchmarks into two primary domains: **user-centric** benchmarks and **agent-centric** benchmarks. User-centric benchmarks, such as MSC (Xu et al., 2022a) and MemoryBank (Zhong et al., 2024), primarily evaluate conversational consistency, evaluating an agent's ability to retain persona information, recall user preferences, and sustain coherent interactions across multi-session or multi-turn dialog. On the other hand, agent-centric benchmarks, including OSWorld (Xie et al., 2024) and Webarena (Zhou et al., 2024), focus on the functional application of memory for complex problem-solving, measuring success rates in tasks that require multi-hop reasoning and tool usage. We present the commonly used user-centric and agent-centric benchmarks in Section 7.2.1 and Section 7.2.2, respectively.

7.2.1 User-centric Evaluation Benchmark

User-centric benchmarks evaluate a foundation agent in personalized dialogue, where the goal is to maintain uniformity with a specific user's evolving profile and the shared interaction history over long horizons. Compared to agent-centric tasks, whose evaluation metrics are largely user-invariant (Mo et al., 2025), user-centric settings are inherently user-dependent (Zhao et al., 2025d). The agent must decide what to store from the conversation, retrieve relevant information when needed, revise it when the user changes their preference, and stay calibrated when evidence is missing (Terranova et al., 2025). Table 3 summarizes representative benchmarks with the interaction scale (#sessions, #questions, maximum context length), data resource (`Real`/`SIM`/`MIX`), and memory ability and evaluation coverage using ✓/✓/✗.

We define ten user-centric memory abilities to characterize what an agent must remember and use over long-horizon interactions (Wu et al., 2025b; Pakhomov et al., 2025): Fact Extraction (FE) is the ability to identify and store reusable facts from dialogue (e.g., user attributes, constraints, key events) so they can be recalled later; Multi-Session Reasoning (MR) requires integrating evidence that is distributed across multiple sessions and turns rather than contained in a single section; Temporal Reasoning (TR) covers reasoning over time series (ordering, timestamps, recency) and selecting the correct state when information changes; Update & Refresh (UR) captures explicitly revising memory when new evidence contradicts old content (overwriting outdated facts and following the latest state under conflicts); Compression & Summarization (CS) is the ability to condense long interaction histories into compact memory representations that remain faithful and usable; Forgetting & Retention (FR) reflects maintaining long-range information while selectively forgetting obsolete or irrelevant content to reduce interference; User Facts & Preferences (UP) focuses on user-centric memory subjects (persona, preferences, relationships, recurring habits or events) and their evolution; Assistant Facts (AS) tracks the assistant's own prior statements, recommendations, or commitments so the agent maintains coherence with what it previously said; Implicit Inference and Connection (IC) measures whether the agent can link scattered clues and perform multi-hop or implicit inference (e.g., applying a previously mentioned limitation to a new recommendation without being reminded); and Abstain & Boundary Han-

Table 3: **Overview of user-centric memory benchmarks. Memory Abilities**— FE: Fact Extraction ; MR: Multi-Session Reasoning ; TR: Temporal Reasoning ; UR: Update & Refresh ; CS: Compression & Summarization ; FR: Forgetting & Retention ; UP: User Facts & Preferences ; AS: Assistant Facts ; IC: Implicit Inference ; AB: Abstain & Boundary Handling. **Marks**— ✔: *explicitly covered*. The benchmark defines this ability as a target with dedicated task types or annotations and evaluates it directly; ✓: *partially or indirectly covered*. The ability may be required in some instances or implied by the setup, but lacks dedicated labeling or ability specific evaluation, so coverage is weak or not cleanly attributable; ✗: *not covered*. No corresponding task, annotation, rubric, and the evaluation does not require or measure this ability. **Data source**— `Real` (human-authored or real-world conversations), `SIM` (fully synthetic or simulated), `MIX` (mixture of real and synthetic). **Evaluation metrics**— AR: Accurate Retrieval; TTL: Test-Time Learning; LRU: Long-Range Understanding; SF: Selective Forgetting; MM-R: Multi-Message Relevance; RC: Response Correctness; CC: Contextual Coherence; MA: Memory Accuracy; MI: Memory Integrity; FMR: False Memory Rate; MRS: Memory Retention Score; JUDGE: LLM-as-a-judge.

Name	#Sess.	#Q	Max Tok.	FE	MR	TR	UR	CS	FR	UP	AS	IC	AB	Res.	Link	Evaluation
MSC	5K	–	1K	✓	✔	✓	✓	✔	✗	✔	✓	✓	✗	`Real`		Perplexity
DuLeMon	27,501	–	1K	✔	✓	✗	✔	✓	✗	✔	✔	✓	✗	`Real`		F1, Recall@k, BLEU, Distinct-n
MemoryBank	300	194	5K	✔	✔	✓	✔	✔	✔	✔	✓	✓	✗	`SIM`		Retrieval Accuracy, RC, CC
PerLTQA	3,409	8,593	1M	✗	✓	✗	✗	✗	✗	✔	✓	✔	✗	`SIM`		Accuracy, F1, Recall@k, MAP
LoCoMo	1K	7,512	10K	✓	✔	✔	✗	✔	✓	✔	✔	✔	✔	`MIX`		F1, ROUGE, BLEU, MM-R
DialSim	∼1,300	1M	367K	✗	✔	✔	✓	✓	✓	✓	✔	✔	✔	`MIX`		Accuracy
LOCCO	3,080	2,981	–	✓	✗	✓	✗	✗	✔	✓	✗	✗	✗	`SIM`		Accuracy, MRS
MemoryAgentBench	130	207	1.44M	✓	✔	✔	✔	✔	✔	✔	✗	✔	✓	`SIM`		Accuracy, F1, AR, TTL, LRU, SF
LongMemEval	50K	500	1.5M	✔	✔	✔	✔	✓	✓	✔	✔	✔	✔	`MIX`		JUDGE, Recall@K, NDCG@K
HaluMem	1,387	3,467	1M	✔	✔	✔	✔	✗	✓	✔	✓	✔	✔	`MIX`		MA, MI, FMR
PersonaMem	60	∼6K	1M	✓	✔	✔	✗	✓	✔	✔	✗	✔	✗	`SIM`		Accuracy
PrefEval	–	3,000	100K	✓	✔	✔	✓	✗	✓	✔	✗	✔	✓	`MIX`		JUDGE, Accuracy
MemBench	65K	53K	100K	✔	✔	✔	✔	✓	✓	✔	✔	✔	✗	`MIX`		Accuracy, Capacity, Efficiency
MemoryBench	20K	–	–	✓	✔	✔	✔	✔	✓	✔	✗	✔	✓	`MIX`		Accuracy, F1
ConvoMem	300	75,335	3M	✔	✔	✔	✔	✗	✓	✔	✔	✔	✔	`SIM`		Accuracy

dling (AB) measures recognizing unknown or unanswerable cases, conflicts, or false premises and avoiding fabrication. For example, explicitly saying "I don't know" when the needed information was never stated.

A key pattern in Table 3 is that MR and UP are the most consistently covered abilities, reflecting the dominant framing of user memory as retrieving user-related facts from long dialogue and using them later. In contrast, operational abilities are much less uniformly defined. Early long dialogue benchmarks such as MSC (Xu et al., 2022a) and DuLeMon (Xu et al., 2022b) emphasize multi-session coherence and persona usage, but their evaluation relies heavily on similarity-based or generated-based metrics (e.g., BLEU, Distinct-n, Perplexity, ROUGE), which weakly isolates memory. Fluency and generic helpfulness can mask incorrect recall, making FE, UR, and FR difficult to attribute. Newer benchmarks increasingly move toward mechanism attributable evaluation by defining specific task types or assessments. MemoryBank (Zhong et al., 2024) and HaluMem (Chen et al., 2025a) introduce explicit memory records and operation-level searching, turning extraction, updating, and memory integrity failures into measurable targets. LongMemEval (Wu et al., 2025b) further strengthens attribution by explicitly categorizing question types according to memory abilities, enabling cleaner localization of failures to temporal reasoning, updating, or abstention behaviors. However, CS and FR remain comparatively under systematically evaluated in the current benchmark designs. Compression is explicit in only a few benchmarks (Hu et al., 2025c), and selective forgetting or retention (Jia et al., 2025b) is frequently partial or absent, despite being essential for long-horizon assistants operating under finite memory budgets and evolving user states. AB is also inconsistently required (Ai et al., 2025; Jiang et al., 2025a). Only a few benchmarks explicitly reward abstention under missing evidence, leaving a gap for evaluating safe memory behavior that prevents confident hallucinations.

Evaluation in Table 3 follows several main aspects that reflect how much evaluation the benchmark provides. First, many benchmarks reduce evaluation to answer-level correctness on ground truth questions, reporting accuracy and F1 (e.g., PerLTQA (Du et al., 2024), LoCoMo (Maharana et al., 2024), DialSim (Kim et al., 2024a), PersonaMem (Jiang et al., 2025a), MemoryAgentBench (Hu et al., 2025c)), and in some cases, using retrieval-style scoring such as Recall@K or ranked relevance metrics (MAP/NDCG@K) when the task is explicitly formed as selecting supporting memories rather than generating free-form text (e.g., PerLTQA (Du et al., 2024), LongMemEval (Wu et al., 2025b)). Second, several user-centric benchmarks introduce memory-specific assessment that goes beyond final answers and directly evaluates memory system behavior. HaluMem reports memory accuracy and integrity as well as false memory rate to quantify hallucinated or incorrect memory operations (Chen et al., 2025a), LOCCO (Jia et al., 2025b) reports memory retention (MRS), and MemBench (Tan et al., 2025a) explicitly measures capacity and efficiency to capture performance and cost trade-offs under fixed memory budgets. Third, when outputs are open-ended or reference answers are insufficient, benchmarks increasingly rely on LLM-as-a-judge to score response correctness or preference adherence (e.g., LongMemEval (Wu et al., 2025b), PrefEval (Zhao et al., 2025c), and MemoryBench (Ai et al., 2025)). Compared to exact-match scoring, judge-based evaluation expands coverage to realistic assistant behavior, but it is more sensitive to the evaluator model and rubric.

7.2.2 Agent-centric Evaluation Benchmark

Agent-centric benchmarks evaluate a foundation model as a *semi-autonomous agent* that must execute multi-step actions in an environment to reach a clearly specified goal, rather than only producing a response to a static prompt. These benchmarks typically provide a task specification (e.g., an initial state with constraints) together with an objective success checker, and summarize performance with end-to-end metrics such as Accuracy/F1 for text QA tasks or Success Rate (SR) for interactive environments. Because correctness is defined by the environment's state, the target outcome is largely **user-invariant**: under the same task setting, different users issuing the same request should obtain the same completion signal, even if the agent follows different trajectories (Trivedi et al., 2024; Zhou et al., 2024).

We summarize commonly used agent-centric benchmarks in Table 4. Beyond tagging each benchmark by task environment (`Env`) and interface (`Interact`), we further annotate (i) **resource type**, indicating whether the task world and evidence are constructed with real-world or simulated data , or a hybrid of both, and (ii) **core agent abilities** (`Abilities`) most directly exercised by the benchmark. We define a set of *core ability tags* commonly used to characterize agent and memory benchmarks: `TEMP` (temporal/sequence reasoning over event order and time dependencies), `STATE` (tracking and updating environment/task state across multi-step interaction), `GROUND` (grounding natural-language instructions into concrete environment targets/actions), `PLAN` (planning and re-planning multi-step actions toward a goal), `TOOL` (selecting and correctly invoking tools/APIs to solve subtasks), `MHOP` (multi-hop reasoning that composes multiple pieces of evidence), `DIAL` (goal-directed dialogue management such as clarification and consistency across turns), and `ACT` (executing correct environment actions such as click/type/select and producing a successful trajectory); optionally, `TTL` denotes test-time learning, where agents improve later performance by accumulating experience in memory *without* parameter updates. In these works, **memory is treated as a tool and module for cross-session knowledge transfer and experiment sharing**, enabling agents to preserve intermediate findings, tool outputs, and state updates across long horizons and limited context windows to improve downstream task completion ability.

Across environments, agent-centric evaluation covers `TEXT/WEB` information seeking (Yang et al., 2018; Trivedi et al., 2022; Deng et al., 2023; Yao et al., 2022; Mialon et al., 2024), `OS/APP` computer use (Xie et al., 2024; Trivedi et al., 2024; Yao et al., 2025; Barres et al., 2025), `CODE` software engineering (Chen et al., 2021; Qiu et al., 2025b), embodied `ROBOT/GAME` control (Shridhar et al., 2020; 2021; Fan et al., 2022), and long-form `VIDEO/PAPER` workflows (Wei et al., 2025a; Mangalam et al., 2023; Fu et al., 2025; Wu et al., 2024a; Starace et al., 2025). These environments induce distinct memory pressures. Text and multi-hop QA emphasize evidence tracking and state bookkeeping over intermediate facts (Ho et al., 2020; Trivedi et al., 2022). Web, desktop, and app settings stress episodic action memory. Foundation agents must remember visited pages, filled fields, downloaded files, prior tool outputs, constraints discovered during interaction and avoid redundant exploration (Deng et al., 2023; Zhou et al., 2024; Yao et al., 2022). Code and paper workflows require working memory, such as files, patches, hypotheses, and experiment logs, to support cross-stage continuity (Starace et al., 2025; Miao et al., 2025). These requirements make compression, selection, and

Table 4: **Overview of agent-centric task benchmarks.** **Env**—TEXT: Text/Document; WEB: Web; OS: Operating System/Desktop; APP: Application/API-centric; CODE: Code/Software Engineering; ROBOT: Embodied Robotics; GAME: Game/Simulation; VIDEO: Video (long-form); PAPER: Scientific Paper/Research. **Interact**—QA: Question Answering; MT: Multi-turn; GUI: Graphical UI; API: Tool/API Invocation; EXEC: Execution-based; MM: Multimodal; ACT: Action/Control. **Data**—REAL: real-world; SIM: simulated/synthetic; MIX: mixed real+sim. **Abilities**—MHOP: Multi-hop Reasoning; PLAN: Planning/Acting; STATE: State Tracking; GROUND: Grounding; TOOL: Tool Use; DEBUG: Debugging; CODEGEN: Code Generation; PATCH: Patch/Repair; TEMP: Temporal Reasoning; DIAL: Dialogue Management; TTL: Test-Time Learning. **Evaluation**—SR: Success Rate / Solve Rate; PR: Pass Rate; GC: Goal Success; RR: Resolved Rate; LCRR: Longest Consecutive Correct Sequence.

Name	#Data	Env	Interact	Resource	Core Abilities	Link	Evaluation
HotpotQA	113K	TEXT	QA	REAL	MHOP, STATE	Github	Accuracy, F1
2WikiMultiHopQA	193K	TEXT	QA	REAL	MHOP, STATE	Github	Accuracy, F1
MuSiQue	25K	TEXT	QA	REAL	MHOP, STATE	Github	F1
HLE	2.5K	TEXT	QA	REAL	MHOP, STATE	Website	Accuracy, RMSE
BrowseComp	1,266	WEB	QA, GUI	REAL	PLAN, TOOL, MHOP, STATE	Website	Accuracy, PR
Mind2Web	2.35K	WEB	GUI	REAL	GROUND, PLAN, STATE	Website	Accuracy, F1, SR
WebArena	812	WEB	GUI	SIM	GROUND, PLAN, STATE	Website	SR
WebShop	12.1K	WEB	GUI	MIX	GROUND, PLAN, STATE	Github	Task Score, SR
GAIA	466	WEB	QA, GUI	REAL	TOOL, MHOP, PLAN, STATE	Website	Accuracy, SR
OSWorld	369	OS	GUI, MM	REAL	GROUND, PLAN, STATE	Website	SR
AppWorld	750	APP	API, MT	SIM	TOOL, CODEGEN, PLAN, STATE	Website	SR
τ-Bench	165	APP	API, MT	SIM	TOOL, PLAN, STATE, DIAL	Github	Pass^1, Pass^k
τ-Bench2	2.3K	APP	API, MT	SIM	TOOL, PLAN, STATE, DIAL	Github	Pass^1, Pass^k
HumanEval	164	CODE	EXEC	REAL	CODEGEN, DEBUG	Github	Pass@1
SWE-Bench	2.3K	CODE	EXEC	REAL	PATCH, DEBUG, STATE	Website	RR
LoCoBench	8K	CODE	EXEC	SIM	CODEGEN, PATCH, DEBUG, STATE	Github	Multi-metric
LoCoBench-Agent	8K	CODE	EXEC	SIM	TOOL, PLAN, CODEGEN, PATCH, DEBUG, STATE	Github	Multi-metric
PaperBench	20	PAPER	EXEC, MT	REAL	TOOL, PLAN, CODEGEN, DEBUG, STATE	Website	Replication Score
RECODE-H	102	PAPER	EXEC, MT	REAL	CODEGEN, PATCH, DEBUG, DIAL, STATE	Github	Recall, PR
ALFRED	25.7K	ROBOT	ACT, MM	SIM	GROUND, PLAN, STATE	Website	SR, GC
ALFWorld	3.8K	ROBOT	ACT, MT	SIM	PLAN, STATE, GROUND	Website	SR
MineDojo	3.1K	GAME	ACT, MM	MIX	PLAN, STATE, GROUND	Website	SR
EgoSchema	5,031	VIDEO	QA, MM	REAL	TEMP	Website	Accuracy
Video-MME	2,700	VIDEO	QA, MM	REAL	TEMP	Website	Accuracy
LongVideoBench	6.7K	VIDEO	QA, MM	REAL	TEMP	Website	Accuracy
MT-Mind2Web	720	WEB	GUI, MT	REAL	GROUND, PLAN, STATE, DIAL	Github	Accuracy, F1, SR
MPR	10.8k	TEXT	QA	SIM	MHOP, STATE	Github	Accuracy
StoryBench	397	GAME	MT, ACT	SIM	PLAN, STATE, TEMP	–	Accuracy, LCRR
Evo-Memory	~3,700	TEXT	QA, MT	MIX	TTL, PLAN, STATE	–	Accuracy, SR
LifelongAgentBench	1,396	APP, OS	API, MT	SIM	TTL, TOOL, PLAN, STATE	Website	SR
OdysseyBench	602	APP	GUI, MT	MIX	PLAN, TOOL, STATE, TEMP	Github	PR

retrieval fidelity critical to complete tasks (Jimenez et al., 2025; Qiu et al., 2025a). Video benchmarks additionally test sensory memory over long clips, where critical cues may appear far before the question is asked (Fu et al., 2025; Wu et al., 2024a). As benchmarks become more realistic and long-horizon, foundation agents cannot retain all task relevant context in the prompt, and must instead externalize state via explicit memory operations to preserve and retrieve critical information over time.

Evaluation method for benchmarks in Table 4 mainly fall into a few categories. Answer-level *Accuracy/F1* for `TEXT` (Trivedi et al., 2022; Phan et al., 2025), goal-based `SR, GC` for interactive `WEB/OS/APP/ROBOT/GAME` (Xie et al., 2024; Trivedi et al., 2024; Shridhar et al., 2020; 2021; Fan et al., 2022), and execution-centric metrics (e.g., `Pass@1`, `RR`) for `CODE` (Jimenez et al., 2025; Qiu et al., 2025a). For memory-centric analysis, two additional dimensions are crucial: (1) **dependency distance**—how far apart the required information and its later use occur, such as within-turn, cross-turn, or cross-session), and (2) **memory correctness under interaction**—whether stored items remain faithful, non-contradictory, and policy-consistent as the environment evolves. This motivates complementing end-to-end success with memory-sensitive measurements such as: (i) retrieval faithfulness and coverage for required facts or tool outputs, (ii) error modes in state tracking (drift, omission, contradiction), (iii) persistence under interruptions (resume after long gaps), and (iv) efficiency trade-offs (memory size, update frequency, and retrieval cost). We expect next-generation benchmarks to go beyond a single end-to-end accuracy or success rate and instead incorporate memory-related metrics. so that evaluation can also be attributed to the memory mechanism rather than to short-horizon prompting or incidental heuristics.

8 Applications

Memory transforms LLMs into dynamic, persistent agents, representing a fundamental shift in recent research. When implemented in complex real-world scenarios, agentic memory has emerged not merely as a storage utility, but as the cognitive substrate that enables continuity, learning, and personalization, bridging an agent's past experiences with its future actions. Recent work has broadly investigated memory-enabled capabilities in LLM agents where the ways of storing, operating, and managing memory vary significantly. To provide insights into how improvements in memory design boost further abilities, this section discusses and summarizes recent representative works across education, scientific research, gaming and simulation, robotics, healthcare, dialogue systems, software engineering, and workflow automation. The summarization is shown in Table 5.

Education. Educational agents require sustained, personalized interactions spanning a long period of time, making memory essential for tracking learner progress, adapting instruction, and maintaining pedagogical coherence (Chu et al., 2025). Without memory, agents treat every interaction as an isolated event, unable to build on a student's prior knowledge or maintain pedagogical consistency. Recent models illustrate this shift toward more sophisticated memory modules. For instance, LOOM (Cui et al., 2025a) utilizes a learner memory graph mapping educational concepts with prerequisite dependencies to facilitate personalized curriculum generation. Agent4Edu (Gao et al., 2025b) explicitly replicates the Ebbinghaus Forgetting Curve to simulate knowledge decay for teacher training. WebCoach (Liu et al., 2025b) uses persistent cross-session memory to enable self-evolving instructional guidance. These systems reveals that in the educational domain, memory functions less as a historical log and more as a cognitive digital twin (Zheng et al., 2022) of the student. Future works should move toward interoperable memory protocols that allow a learner's cognitive profile to persist across different educational platforms, effectively creating an evolving record of their intellectual development.

Scientific Research. Scientific research represents a frontier where the process has been lengthy and costly, requiring agents to synthesize vast literature, manage provenance, and maintain reasoning continuity across multi-stage endeavors. In recent studies, General Agentic Memory (GAM) (Yan et al., 2025a) employs a specialized researcher agent for deep research over a universal page-store, enabling dynamic context reconstruction for complex multi-hop reasoning. IterResearch (Chen et al., 2025c) maintains a workspace preserving only the evolving report and immediate results to prevent context suffocation. MirrorMind (Zeng et al., 2025) simulates collective intelligence through hierarchical architecture retrieving specific cognitive styles and knowledge bases. AISAC (Bhattacharya & Som, 2025) implements hybrid memory combining semantic retrieval with structured SQLite logs for reproducibility. These research agents exemplify a paradigm shift where memory serves as a verification layer for the research process, maintaining a transparent lineage of how a conclusion was reached. Future systems will evolve from solitary research assistants into lab-scale

Table 5: **Summarization of representative agentic memory applications.**

Application	Memory Utilization	Works
Education	Tracks learner progress and simulates knowledge decay to provide personalized pedagogical guidance.	LOOM (Cui et al., 2025a), Agent4Edu (Gao et al., 2025b), WebCoach (Liu et al., 2025b), CAM (Li et al., 2025d), Classroom Simulacra (Xu et al., 2025b), TeachTune (Jin et al., 2025a), EduAgent (Xu et al., 2024a), EvaAI (Lagakis & Demetriadis, 2024), OATutor (Pardos et al., 2023), MEDCO (Wei et al., 2024)
Scientific Research	Synthesizes vast literature and maintains reasoning provenance across multi-stage discovery processes.	IterResearch (Chen et al., 2025c), GAM (Yan et al., 2025a), MirrorMind (Zeng et al., 2025), AISAC (Bhattacharya & Som, 2025), Lee et al. (2024), ChemDFM (Zhao et al., 2025e), AI-coscientist (Gottweis et al., 2025), SciAgents (Ghafarollahi & Buehler, 2025), Agent Laboratory (Schmidgall et al., 2025), NovelSeek (Team et al., 2025)
Gaming & Simulation	Enables bottom-up skill acquisition and the emergence of complex social dynamics through episodic memories.	Voyager (Wang et al., 2025c), GITM (Zhu et al., 2023), Generative Agents (Park et al., 2023), GameGPT (Chen et al., 2023), M2PA (Zhou et al., 2025b), Jiang et al. (2025d), WarAgent (Hua et al., 2023), S3 (Gao et al., 2023), AvalonBench (Light et al., 2023), Mosaic (Liu et al., 2025c)
Robotics	Bridges high-level reasoning with low-level control by maintaining spatial graphs and trajectory summaries.	Memo (Gupta et al., 2025), MG-Nav (Wang et al., 2025b), JARVIS-1 (Wang et al., 2024q), KARMA (Wang et al., 2025s), VIPeR (Ming et al., 2025), SAM 2 (Liu et al., 2025d), LRLL (Tziafas & Kasaei, 2024), VideoAgent (Fan et al., 2024), Kim et al. (2023a), RAP (Kagaya et al., 2024), GridMM (Wang et al., 2023d)
Healthcare	Maintains longitudinal records of physiological trends and emotional states to build user trust and adherence.	TheraMind (Hu et al., 2025a), DAM (Lu & Li, 2025), Mem-PAL (Huang et al., 2025d), ReSurgSAM2 (Liu et al., 2025d), CAR-AD (Li et al., 2025e), AgentMD (Jin et al., 2025b), MedConMA (Wang et al., 2025k), MDAgents (Kim et al., 2024b), MedAgents (Tang et al., 2024), ChatCAD (Tang et al., 2025a)
Dialogue Systems	Manages context window constraints and persona consistency to simulate persistent human-like relationships.	A-Mem (Xu et al., 2025e), MemGPT (Packer et al., 2023), O-Mem (Wang et al., 2025i), MemoChat (Lu et al., 2023), Mem0 (Chhikara et al., 2025), SEAL (Wang et al., 2025d), LiCoMemory (Huang et al., 2025e), Lu & Li (2025), Terranova et al. (2025), LightMem (Fang et al., 2025a), RGMem (Tian et al., 2025)
Workflow Automation	Induces reusable workflow templates and learns tool-usage patterns from successful execution histories.	AWM (Wang et al., 2025u), ToolMem (Xiao et al., 2025), Synapse (Zheng et al., 2024), WebArena (Zhou et al., 2024), Wheeler & Jeunen (2025), Wang et al. (2025t), WALT (Prabhu et al., 2025), Mobile-agent-v2 (Wang et al., 2024c), AutoAgents (Chen et al., 2024a), SIT-Graph (Li et al., 2025f)
Software Engineering	Maintains global code context and recalls failure trajectories to improve multi-file debugging and development.	MetaGPT (Hong et al., 2023), ChatDev (Qian et al., 2024a), SWE-bench (Jimenez et al., 2025), SWE-Effi (Fan et al., 2025c), TroVE (Wang et al., 2024p), Self-organized agents (Ishibashi & Nishimura, 2024), Openhands (Wang et al., 2024j), Masai (Arora et al., 2024), DeepCode (Li et al., 2025m)
Online Streaming & Recommendation	Distills high-throughput multimodal feeds into persistent representations to recognize long-range temporal patterns.	WorldMM (Yeo et al., 2025b), GCAgent (Yeo et al., 2025a), (Xiong et al., 2025b), XMem++ (Bekuzarov et al., 2023), VideoScan (Li et al., 2025c), Xiong et al. (2025b), Qian et al. (2024b), VideoLLM-Online (Chen et al., 2024c), VideoLLM-MoD (Wu et al., 2024c), Di et al. (2025)
Information Search	Transforms static retrieval into active workspaces for synthesizing conflicting reports and tracking search provenance.	AgentFold (Ye et al., 2025b), MemSearcher (Yuan et al., 2025a), MoM (Zhao et al., 2025a), ReSum (Wu et al., 2025d), Memento (Zhou et al., 2025a), MLP Memory (Wei et al., 2025c), MemAgent (Yu et al., 2025b), Wang et al. (2025o), MemoryLLM (Wang et al., 2024m)
Finance & Accounting	Maintains strategic consistency across volatile market cycles and balances quantitative signals with qualitative historical precedents.	FinCon (Yu et al., 2024a), FinMem (Yu et al., 2025e), QuantAgent (Wang et al., 2024g), FLAG-Trader (Xiong et al., 2025a), Investor-Bench (Li et al., 2025b), TradingAgents (Xiao et al., 2024), Trading-GPT (Li et al., 2023b), Open-FinLLMs(Huang et al., 2024a)
Legal & Consulting	Manages multi-document provenance and synthesizes conflicting statutes into coherent advice across long-term case histories.	MALR (Yuan et al., 2024b), StaffPro (Maritan, 2025), Blair-Stanek et al. (2025), LegalMind (Vara et al., 2025), CaseGPT (Yang, 2024), Dallma (Westermann, 2024), Agentcourt (Chen et al., 2025b), Legal-GPT (Shi et al., 2024), Feat (Shen et al., 2025)

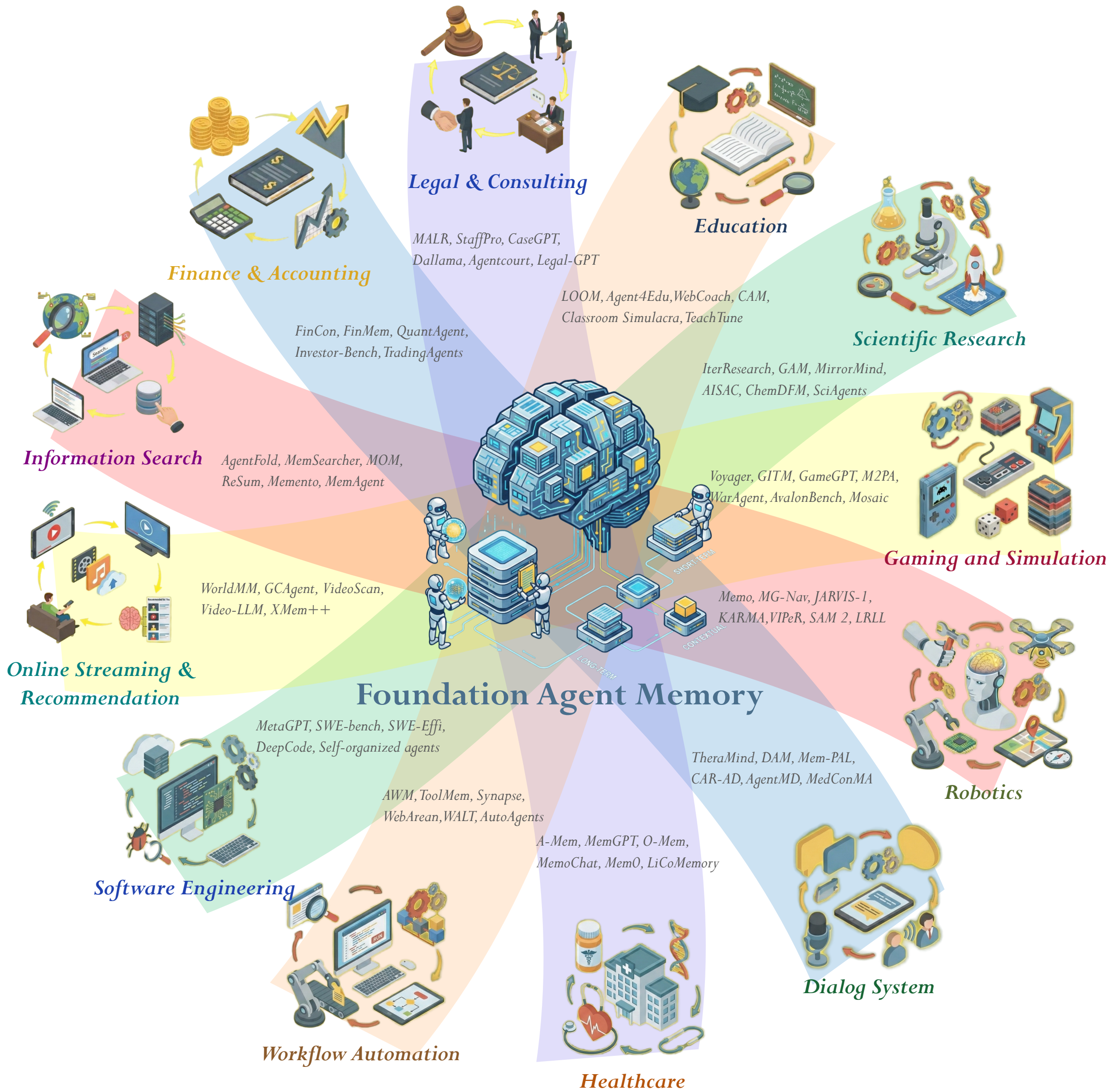

Figure 8: **Applications of the Foundation Agent Memory System**. The diagram introduce the general application domain of the foundation agent memory system, including *eduction*, *scientific research*, *gaming and simulation*, *robotics*, *dialog system*, *healthcare*, *workflow automation*, *software engineering*, *online stream and recommendation*, *information search*, *finance and accounting*, *legal and consulting*.

collective intelligences, where multiple agents share a unified, evolving knowledge graph of a specific scientific field, updating it in real-time as new papers are published and synthesized.

Gaming and Simulation. In open-ended gaming environments and simulations, memory enables skill acquisition, spatial exploration, and the emergence of complex social dynamics. Agents must utilize procedural memory to retain learned skills and episodic memory to maintain believable social histories. For example, Voyager (Wang et al., 2025c) stores successful actions as executable code in a skill library for compounding abilities. GITM (Zhu et al., 2023) employs hierarchical text-based memory where a planner records structured sub-goal summaries. Generative Agents (Park et al., 2023) uses a memory stream where agents reflect to synthesize high-level insights into relationships and plans. GameGPT (Chen et al., 2023) applies memory as shared state for multi-agent game development, managing versioning and conflict resolution. In these works, memory modules allow behavior to emerge bottom-up from the accumulation of experiences rather than top-down programming. The next frontier in this domain could be the development of socially alignment and forgetting mechanisms. As simulations run for extended periods, agents should mimic human-like memory decay, ensuring personalities evolve organically without being paralyzed by the noise of infinite, trivial historical data.

Robotics. Embodied agents operating in physical worlds face the challenge of partial observability (Fung et al., 2025), requiring memory to link visual inputs to semantic concepts and maintain spatial representations over time. Memory must be compressed yet sufficiently detailed to support navigation and manipulation in non-static environments. Memo (Gupta et al., 2025) introduces periodic summarization tokens compressing trajectories for long-horizon navigation. MG-Nav (Wang et al., 2025b) constructs spatial memory graphs with landmark regions rather than dense point clouds, mimicking human navigation. JARVIS-1 (Wang et al., 2024q) extends embodied agency with multimodal memory retrieving experiences based on visual and semantic similarity. These applications demonstrate that memory is the bridge between high-level reasoning and low-level control. Future research should focus on multimodal memory integration, enabling agents to simulate the physical affordances based on past successes and failures stored in their procedural memory.

Healthcare. In the domain of healthcare, memory enables agents to track longitudinal health trends, emotional trajectories, and the efficacy of interventions. TheraMind (Hu et al., 2025a) introduces a dual-loop architecture separating immediate responses from strategic cross-session memory updates for therapeutic strategy adjustment. DAM (Lu & Li, 2025) treats memory units as confidence distributions over sentiment polarities for stable probabilistic emotion modeling. Mem-PAL (Huang et al., 2025d) employs H^2Memory architecture distinguishing between objective physiological logs and subjective dialogue to infer health metric correlations. The deployment of agentic memory reveals that affective continuity is as critical as clinical accuracy, lead to a measurable increase in user trust and adherence. However, this domain faces challenges regarding privacy and ethics. Future architectures should implement privacy-preserving memory by design mechanisms that balance the utility of long-term memory with the imperative of patient confidentiality.

Dialogue Systems. For general-purpose assistants, memory creates the illusion of a continuous personalized relationship, managing context window while providing conversation history. This domain focuses on actively managing the trade-off between retention and context window constraints. MemGPT (Packer et al., 2023) introduces a context management system explicitly moving data between main and external context for longer conversations. O-Mem (Wang et al., 2025i) uses tri-component memory to extract and update holistic user personas for aligned responses. MemoChat (Lu et al., 2023) employs instructional tuning to train models on writing structured memos for improved long-range consistency. These dialogue architectures demonstrate a shift toward OS-level memory management, where the agent acts as a kernel managing its own resources. The future of dialogue systems lies in self-optimizing memory. Rather than relying on fixed heuristic rules for what to remember, next-generation agents will likely learn personalized memory policies.

Workflow Automation. LLM agents also work as assistants to boost productivity through workflow automation. Automation agents mainly orchestrate multi-step processes, coordinate tool usage, and adapt procedures based on execution feedback. Memory mechanisms enable agents to accumulate procedural knowledge, optimize workflow, and maintain task context across complex automation pipelines. AWM (Wang et al., 2025u) induces reusable workflow templates from successful trajectories as parameterized procedural memory. ToolMem (Xiao et al., 2025) implements semantic memory of tool usage patterns, learning effective tools for specific task types. WebArena (Zhou et al., 2024) benchmarking demonstrates that episodic memory of web interaction sequences substantially outperforms baselines. Synapse (Zheng et al., 2024) introduces trajectory-as-exemplar prompting, storing successful control sequences as episodic memory for analogical reasoning. In these works, the integration of memory facilitates the transition from rigid script execution to adaptive procedural learning which dramatic increase robustness and efficiency. Future systems will not just follow human-defined operation procedures but should actively rewrite their own instructions based on long-term execution logs, evolving from simple task executors into process architects that autonomously refine enterprise workflows for maximum efficiency.

Software Engineering. For software engineering, agents operate in complex codebases requiring long-horizon reasoning, multi-file coordination, and accumulated debugging experience. Memory enables these agents to maintain code context, learn from implementation attempts, and navigate large-scale repositories. MetaGPT (Hong et al., 2023) implements procedural memory for development workflows while maintaining shared semantic memory of project specifications. ChatDev (Qian et al., 2024a) extends this with episodic memory of development iterations for learning from debugging sessions. SWE-bench (Jimenez et al., 2025) evaluation reveals that memory mechanisms significantly improve issue resolution by maintaining context across multi-file edits and leveraging prior debugging experience. Critically, effective coding agents do not merely generate code, they recall the trajectory of previous failures and fixes to achieve higher success rates in passing unit tests. Future applications in this domain will likely move beyond local project memory toward shared, anonymized knowledge repositories where distributed coding agents contribute to and query a universal pool of algorithmic solutions and error patches, accelerating the global pace of automated software development.

Online Streaming and Recommendation. In the era of online streaming and recommendation systems, agents process high-throughput multimodal inputs where the relevance of information shifts dynamically over time. Memory allows agents to maintain temporal consistency and recognize long-range patterns across video frames and user interactions. For instance, WorldMM (Yeo et al., 2025b) utilizes dynamic multimodal memory storing visual-linguistic features for complex reasoning over long-duration video streams. GCA-gent (Yeo et al., 2025a) introduces dual-structured episodic memory separating schematic knowledge from narrative sequences for structured video understanding. Similarly, Xiong et al. (2025b) implements memory-enhanced knowledge buffers supporting multi-round interactions with retained context. These applications suggest that in streaming contexts, memory acts as a temporal filter that distillations transient data into persistent representations. Future research should focus on forgetting-aware recommendation memories that can distinguish between a user's fleeting interests and their long-term preferences, optimizing the balance between novelty and relevance in real-time feeds.

Information Search. Beyond simple retrieval, agents for information search must synthesize conflicting reports, track evolving stories, and manage vast document spaces without losing reasoning depth. Memory serves as the organizational framework that transforms static search results into an active workspace for knowledge synthesis. AgentFold (Ye et al., 2025b) addresses long-horizon web navigation through proactive context management, folding irrelevant trajectories to prevent overflow while preserving critical findings. MemSearcher (Yuan et al., 2025a) employs reinforcement learning for joint searching and memory management. MoM (Zhao et al., 2025a) utilizes scenario-aware memories, dynamically routing queries to specialized memory banks. Furthermore, Rajesh et al. (2025) bridges RAG with episodic memory, maintaining a repository of the search process itself. These systems demonstrate that effective search is not just about finding data, but about managing the cognitive load of the search trajectory. The next generation of information search agents will likely evolve toward collaborative memory structures, where multiple agents verify facts and update a shared belief graph among agents in response to breaking information cycles.

Finance and Accounting. The financial domain is characterized by high-frequency volatility, a mixture of quantitative data and qualitative news, and the critical need for long-term strategic consistency. Memory is crucial here because trading and accounting require agents to process real-time signals and recall historical market regimes and maintain a persistent trading character to avoid erratic decision-making. Recent works have introduced specialized architectures for this purpose. FinMem (Yu et al., 2025e) implements a layered memory system that separates immediate market observations from long-term investment experience, allowing the agent to refine its personality and risk profile over time. FinCon (Yu et al., 2024a) utilizes a multi-agent setup where memory serves as a repository for conceptual verbal reinforcement, enabling the system to learn from past financial decisions through reflective feedback loops. QuantAgent (Wang et al., 2024g) further pushes this by seeking investment grails through a self-improving mechanism where success trajectories are stored as procedural memory for future strategy refinement. Despite these gains, the primary challenge remains the signal-to-noise ratio in financial memory. Agents must learn to distinguish between transient market fluctuations and fundamental shifts. Future research should explore forgetting-aware financial agents that can prune outdated economic assumptions while retaining core risk management principles.

Legal and Consulting. In legal and consulting services, agents must navigate massive volumes of heterogeneous documents, where the precise provenance of every claim is mandatory. Memory is the cognitive substrate that allows these agents to perform multi-step reasoning over long-duration cases, ensuring that advice remains consistent with previously cited statutes or client history. For instance, MALR (Yuan et al., 2024b) utilizes a multi-agent framework to improve complex legal reasoning by maintaining an interaction history that simulates collaborative debate between legal experts. StaffPro (Maritan, 2025) focuses on the consulting side by using memory to profile workers and project requirements over time, enabling dynamic staffing through a feedback loop of past performance data. Blair-Stanek et al. (2025) demonstrates the power of memory in discovering novel tax-minimization strategies by synthesizing thousands of pages of evolving statutes and case law into a persistent reasoning graph. The core challenge in this domain is the high stakes of hallucinated memory. A single misremembered clause can lead to legal liability. Future directions will likely focus on verifiable memory architectures that link every retrieved insight back to a cryptographically signed source document, ensuring the highest levels of professional integrity and accountability.

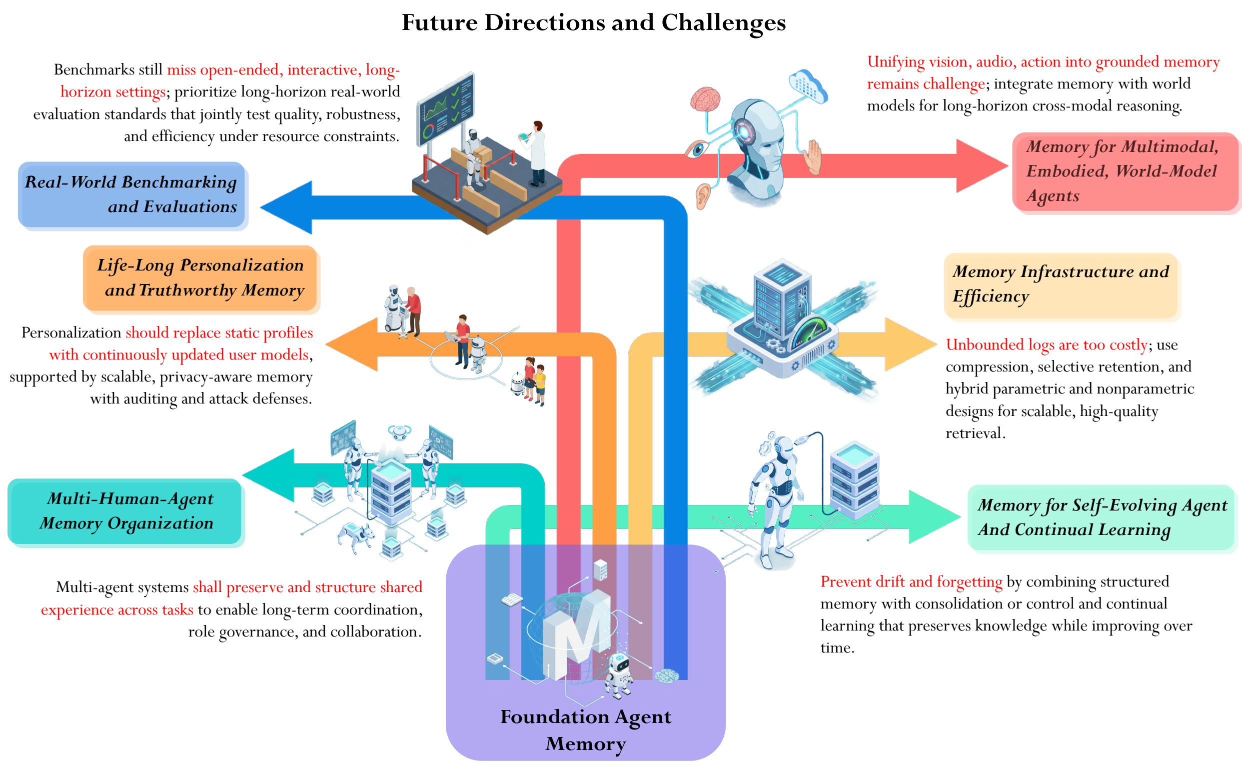

Figure 9: **Future Directions and Challenges in Foundation Agent Memory**. The diagram highlights key opportunities for the future argentic memory works, including *memory for self-evolving agent*, *multi-agent memory organization*, *human-agent collaborative memory*, *memory efficiency*, *memory for multi-modal and embodied agent*, *memory for continuous learning*, *life-long personalization*, *real-world benchmarking and evaluation*

9 Future Directions

9.1 Memory for Continual Learning and Self-Evolving Agents

A fundamental challenge in memory-enabled self-evolving agents lies in managing memory dynamics across both intra-task and cross-task timescales. At the intra-task level, agents must continuously decide what information to retain, compress, or discard from heterogeneous streams such as tool outputs, search results, feedback signals, and intermediate reasoning traces, all under strict context-window constraints (Gao et al., 2025a). Existing systems largely rely on heuristic memory controllers, leaving the internal coupling between memory evolution and reasoning behavior poorly understood. At the cross-task level, agents are expected to accumulate experience across episodes and task distributions (Wei et al., 2025d), yet current approaches primarily emphasize inference-time reuse rather than principled consolidation and generalization. In comparison, classical continual learning methods focus on preventing catastrophic forgetting through replay, regularization, or parameter isolation (Rebuffi et al., 2017; Lopez-Paz & Ranzato, 2017; Shin et al., 2017; Kemker & Kanan, 2018), but they typically treat memory as a static mechanism for knowledge retention. This framing is insufficient for agent-based systems, where memory must also track evolving interaction states, user-specific information, and procedural behaviors. Moreover, stable post-training adaptation from accumulated experience remains underexplored, with unresolved risks of negative transfer, uncontrolled drift, and semantic inconsistency (Ke et al., 2025b).

Future research should therefore reframe continual learning around a richer, agent-centric notion of memory that integrates semantic, episodic, and procedural components (Ke et al., 2025a; Lou, 2025). Beyond explicit textual logs, latent and structured memory representations offer a promising direction for scalable and efficient adaptation, enabling compact storage while preserving causal and behavioral abstractions. Progress will likely require moving beyond inference-time heuristics toward principled post-training paradigms that leverage accumulated agent experience for continual improvement. This includes designing consolidation mechanisms that selectively distill long-term knowledge, align evolving memory with model parameters, and mitigate forgetting without sacrificing plasticity. Correspondingly, new benchmarks are needed that evaluate not only task-level retention, but also sustained adaptation, relevance-aware memory management, and behavioral stability under non-stationary objectives and environments (Ke et al., 2024). Establishing a unified framework that connects classical continual learning objectives with structured memory design remains a key open direction for self-evolving foundation agents.

9.2 Multi-Human-Agent Memory Organization

Recent multi-agent LLM frameworks, such as AutoGen (Wu et al., 2024b) and AgentLite (Liu et al., 2024c), enable task decomposition and role-based coordination through structured message passing and prompt-driven control. In practice, such systems increasingly operate in human-agent collaborative settings, where artificial agents interact not only with other agents but also with human users or supervisors through iterative feedback, correction, and delegation (Lu et al., 2025c; Zou et al., 2025b;c). However, despite this growing complexity, coordination remains largely episodic and transient: interactions are scoped to a single task instance, and little experience is retained once the task is completed Suzgun et al. (2025). As a result, both agent–agent and human–agent collaborations are repeatedly re-established from scratch, limiting the system's ability to adapt interaction strategies, personalize behavior, or improve collaboration quality across repeated tasks or deployments.

Enabling persistent and adaptive collaboration among interacting entities (including both foundation agents and humans) would inspire long-term research questions (Han et al., 2024; Li et al., 2024c). One important direction is collaborative (social) memory, where agents retain experience about their collaborators, such as communication preferences, domain expertise, feedback patterns, or historical interaction outcomes, allowing them to adapt signaling strategies, calibrate trust, and reduce coordination overhead over time. At the same time, agents may benefit from role-specific flow and procedural memory, accumulating experience about their own recurring workflows (Wang et al., 2025u), including task decomposition patterns, execution strategies, and common failure modes, so that agents assuming stable functional roles can gradually refine their behavior through experience-driven specialization. Introducing persistent memory in such multi-entity settings also raises memory governance and coordination challenges, including questions of ownership, access, responsibility, and how divergent perspectives or human corrections should be handled. Addressing these issues is essential for preventing uncontrolled error propagation and for sustaining reliable collaboration as multi-agent systems scale in size, heterogeneity, and task complexity.

9.3 Memory Infrastructure and Efficiency

As foundation agents are increasingly deployed in long-horizon, interactive, and open-ended environments, memory infrastructure has emerged as a central efficiency bottleneck (Chhikara et al., 2025; Qiu et al., 2025b). Most existing agent memory designs remain text-centric, treating memory as an ever-growing collection of past interactions, summaries, or episodic logs that are retrieved and injected into the prompt (Xu et al., 2025e; Zhong et al., 2024). While this strategy can improve task performance, it induces substantial token overhead, with memory contexts routinely expanding to thousands of tokens and exhibiting diminishing marginal returns (Chhikara et al., 2025). This linear growth in memory cost (Kwon et al., 2023) directly translates into higher inference latency and reduced scalability, particularly in multi-turn or lifelong settings. More fundamentally, current approaches conflate memory capacity with prompt length, implicitly assuming that more context implies better reasoning. This assumption overlooks the need for selective retention, structured access, and consolidation of experience. Moreover, memory is often managed externally through heuristics such as summarization or truncation, rather than being integrated into the agent's reasoning and learning process. These limitations highlight a core challenge: how to design memory systems that enable agents to retain, abstract, and reuse experience efficiently under strict resource constraints, without relying on unbounded context expansion.

Future research on memory infrastructure and efficiency can be viewed as a progression toward increasingly abstract and integrated representations of experience. In the near term, a promising direction lies in organized text-based memory, where textual memories are explicitly structured for efficient access rather than maximal coverage. Recent work has explored schema-based or graph-structured memory representations (Edge et al., 2024; Chhikara et al., 2025), but these efforts primarily target reasoning accuracy rather than efficiency. An open opportunity is to design structure-aware storage and precision-oriented retrieval mechanisms that expose only reasoning-critical spans, minimizing unnecessary context injection. Beyond textual organization, efficiency gains may be achieved through compressed latent memory, where episodic, semantic, or procedural experiences are encoded into compact vector representations that function as persistent memory units rather than mere similarity indices. At a deeper level of integration, internalized or parametric memory offers a path toward constant-sized memory, where long-term experience is absorbed into internal states or model parameters. Frameworks such as MEM1 (Zhou et al., 2025c) and Mem-α (Wang et al., 2025p) exemplify this shift by training agents, via reinforcement learning, to consolidate, update, and discard memory as part of the reasoning process itself, enabling bounded memory even in long-horizon tasks. Realizing these directions will also require robust environment infrastructure capable of supporting controlled, multi-step interactions and scalable evaluation. Platforms such as NeMo Gym (NVIDIA, 2025), which decouple environment logic from training and provide modular reward and verification services, represent an essential component of this ecosystem. Together, these advances suggest a future in which memory is no longer an external prompt-management artifact but a core, learned subsystem co-evolving with agent reasoning and decision-making, realized through integrated memory architectures that combine structured latent representations (e.g., hierarchical vector tables with differentiable read/write interfaces), joint optimization of memory and policy via end-to-end reinforcement learning or meta-learning objectives, and adaptive memory controllers that dynamically allocate, compress, and retire memory units based on task relevance, uncertainty estimates, and long-term utility.

9.4 Life-Long Personalization and Trustworthy Memory

Life-long personalization seeks to equip foundation agents with the ability to continuously adapt to individual users across sessions, tasks, and extended time horizons (Wang et al., 2024i). Unlike conventional personalization approaches that rely on static user profiles or transient contextual signals, this setting requires agents to maintain evolving user representations that capture gradual preference shifts, long-term goals, and behavioral regularities. While recent efforts on persistent memory and dynamic user modeling have made initial progress (Zhong et al., 2024; Tan et al., 2025c; Zhang et al., 2025m), existing systems largely depend on heuristic aggregation of interaction histories or unstructured memory retrieval, which limits their ability to distill reliable, interpretable, and causally grounded user knowledge (Pink et al., 2025). Moreover, long-horizon personalization introduces non-trivial challenges in memory staleness, concept drift, and credit assignment: agents must decide which past interactions remain relevant, how to reconcile conflicting signals over time, and how to prevent outdated preferences from dominating current behavior. These issues are further exacerbated by scalability constraints, as naively retaining or replaying long interaction histories leads to prohibitive storage, retrieval, and inference costs, especially when deployed in real-world, always-on assistant settings.

A key research direction is the design of scalable and dynamic memory systems that can incrementally update user modeling while bridging fine-grained episodic traces with higher-level abstractions such as preferences, habits, or long-term intents. Promising approaches include hierarchical memory architectures that separate short-term episodic buffers from distilled semantic user profiles (Tan et al., 2025c), learned memory controllers that regulate when to write, compress, or overwrite user information (Zhang et al., 2025m), and continual representation learning techniques that mitigate forgetting under distribution shift (De Lange et al., 2021; Parisi et al., 2019). In parallel, the field requires new evaluation benchmarks tailored to life-long personalization, moving beyond single-turn accuracy toward metrics that assess long-term consistency, adaptability to preference changes, and robustness under extended interactions (Xu et al., 2025e). Equally important is the development of trustworthy memory infrastructures. Persistent user memory raises substantial risks, including privacy leakage (Wang et al., 2025a), memory poisoning (Tan et al., 2024b), and adversarial manipulation (Dong et al., 2025), which can accumulate silently over time. Recent work on secure and auditable memory modules (Wei et al., 2025b; Wang et al., 2025a) highlights the need for user-controllable mechanisms that support inspection, editing, and revocation of stored memories, alongside defenses against unauthorized access and malicious writes. Ultimately, robustness, transparency, and security should be treated as first-class objectives, on par with adaptability, when designing and evaluating life-long personalized foundation agents (Yu et al., 2025c).

9.5 Memory for Multimodal, Embodied, and World-Model Agents

A central challenge for next-generation foundation agents lies in designing memory systems that can faithfully represent, align, and abstract heterogeneous sensory streams, including vision, audio, language, tactile feedback, and proprioceptive signals, into coherent internal states (Bei et al., 2026). While textual memory mechanisms have achieved notable success in long-horizon reasoning and personalization (Xu et al., 2025e), existing approaches largely assume unimodal or language-dominant representations. Early efforts in multimodal agent memory (Long et al., 2025; Bo et al., 2025; Liu et al., 2025g) reveal that naively extending text-based memory to high-dimensional perceptual inputs leads to severe inefficiencies, semantic misalignment across modalities, and brittle retrieval behaviors. These challenges are further amplified in embodied settings, where agents operate in closed-loop environments and must reason over temporally extended perception–action–outcome trajectories. In such scenarios, memory must go beyond storing episodic observations and instead encode grounded knowledge about dynamics, affordances, and physical constraints (Wang et al., 2025s). However, current systems lack principled mechanisms for action-conditioned memory updates, cross-modal abstraction, and consistency maintenance across episodic, semantic, and procedural memory layers. As a result, embodied agents often struggle with skill fragmentation, long-horizon planning failures, and compounding errors caused by misaligned or stale memories.

Looking forward, a promising research direction is to elevate agent memory into an explicit, predictive world model that treats memory not as a passive log, but as a controllable internal state evolving over time. World-model-based formulations (Hafner et al., 2023; Ha & Schmidhuber, 2018) provide a unifying perspective in which memory updates can be modeled as latent state transitions conditioned on perception and action. This opens the door to *proactive memory planning*, where agents simulate the long-term consequences of storing, compressing, or forgetting information before committing updates (Schrittwieser et al., 2020; Silver et al., 2017). Within this framework, memory operations become internal actions optimized jointly with external decision-making, enabling agents to balance immediate utility with long-term consistency and task performance. Moreover, integrating multimodal memory with structured world representations, such as spatial maps, object-centric graphs (Singh et al., 2023), or skill graphs (Wang et al., 2025c; Feng et al., 2025a), can support abstraction across time and modality while improving retrieval efficiency. Finally, memory and world models should be co-trained in a mutually reinforcing loop: stable, structured memory can provide long-term state cues that improve world-model prediction, while world models can regularize memory evolution to prevent identity drift, goal inconsistency, and behavioral instability (Savinov et al., 2019). Advancing this synergy is key to building scalable, reliable multimodal and embodied agents capable of long-horizon autonomy in complex real-world environments.

9.6 Real-World Benchmarking and Evaluations

A central challenge in real-world benchmarking for memory-enabled foundation agents lies in the persistent mismatch between research-level benchmark abstractions and real-world deployment complexities, for both user-centric and agent-centric memory. On the user side, most existing benchmarks reduce long-term personalization to synthetic factual recall, where agents retrieve static user attributes embedded in long contexts or scripted interaction histories (e.g., persona facts, preferences, or conversations). While such settings facilitate controlled evaluation, they fail to capture real user satisfaction, which depends on preference drift, conflicting signals, partial observability, and delayed feedback over weeks or months. Benchmarks such as LoCoMo (Maharana et al., 2024) and PersonaMem (Jiang et al., 2025a) emphasize long-context retrieval accuracy, yet implicitly assume stationary user intent and unambiguous ground truth, overlooking critical failure modes such as stale preference reuse, incorrect overwriting of long-term user state, or unsafe retention of sensitive information. On the agent-centric side, interactive benchmarks including WebArena (Zhou et al., 2024) and OSWorld (Xie et al., 2024) improve realism through execution-based evaluation, but remain bounded by curated environments, reset-centric task design, and short evaluation horizons. These constraints obscure whether agents can accumulate, revise, and safely exploit experience across episodes, especially under non-stationary tools, policies, or environments. As a result, agents may optimize for short-horizon success while silently failing at memory-critical competencies such as provenance tracking, contradiction resolution, and long-term policy consistency. This gap mirrors observations in general assistant benchmarks such as GAIA (Mialon et al., 2024), where failures often arise not from isolated reasoning errors but from brittle state transitions and incorrect memory updates across multimodal and tool-mediated interactions.

Future research should move toward closed-loop, longitudinal, and execution-grounded evaluation paradigms that explicitly stress persistent memory under realistic constraints, for both users and agents. For user-centric memory, benchmarks should incorporate recurring interactions with controlled preference drift, ambiguous feedback, and real-user rewards, enabling direct measurement of satisfaction-aligned memory behaviors such as compression, selective forgetting, and safe overwriting, rather than static recall accuracy (e.g., extending user history with multi-month preference evolution or counterfactual feedback). For agent-centric memory, evaluation should go beyond simulated resets toward partially open or continuously evolving environments, where experience accumulation has real consequences, such as financial trading sandboxes, long-running web services, or competitive control tasks with delayed payoffs, enabling comparison between memory-augmented agents and memory-free baselines under identical conditions. Execution-based frameworks like OSWorld (Xie et al., 2024) can be extended with memory-sensitive invariants, requiring agents to version, audit, and roll back persistent state, and to attach provenance metadata to stored knowledge. In parallel, standardized tool-mediation layers (e.g., MCP-style interfaces) can enable reproducible logging, permission enforcement, and replay, supporting fine-grained evaluation of memory–policy interactions under realistic constraints. Finally, benchmarks should explicitly quantify resource–utility trade-offs, measuring memory quality as a function of token budget, storage cost, and latency, reflecting the bounded-memory conditions of real deployments. Collectively, these directions reframe benchmarking from episodic task completion toward systems-level evaluation of memory as a first-class capability, jointly shaping user trust, agent autonomy, and long-term utility across evolving environments.

10 Conclusions

Memory is becoming the key component for foundation agents operating in long-horizon, context-exploded, and user-dependent environments. In this survey, we unify the design along three dimensions, including memory substrates (internal and external), cognitive mechanisms (sensory, working, episodic, semantic, and procedural), and memory subjects (user- and agent-centric), and analyze how memory is operated under single- and multi-agent systems, as well as how it is increasingly shaped by prompting-, fine-tuning-, and RL-based learning policies. In addition, we also summarize the metrics and benchmarks used to assess foundation agent performance, and categorize current works into representative application domains. To foster future memory research, we list out six key challenges to collectively point to a reliable, scalable, self-evolving, and trustworthy memory infrastructures for real-world human-agent memory system design.

References

Sara Abdali, Richard Anarfi, CJ Barberan, Jia He, and Erfan Shayegani. Securing large language models: Threats, vulnerabilities and responsible practices. *arXiv preprint arXiv:2403.12503*, 2024.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Qingyao Ai, Yichen Tang, Changyue Wang, Jianming Long, Weihang Su, and Yiqun Liu. Memorybench: A benchmark for memory and continual learning in llm systems. *arXiv preprint arXiv:2510.17281*, 2025.

Chandra Vamsi Krishna Alla, Harish Naidu Gaddam, and Manohar Kommi. Budgetmem: Learning selective memory policies for cost-efficient long-context processing in language models. *arXiv preprint arXiv:2511.04919*, 2025.

Saad Alqithami. Forgetful but faithful: A cognitive memory architecture and benchmark for privacy-aware generative agents. *arXiv preprint arXiv:2512.12856*, 2025.

Benjamin Ampel, Chi-Heng Yang, James Hu, and Hsinchun Chen. Large language models for conducting advanced text analytics information systems research. *ACM Transactions on Management Information Systems*, 16(1):1–27, 2025.

Petr Anokhin, Nikita Semenov, Artyom Sorokin, Dmitry Evseev, Andrey Kravchenko, Mikhail Burtsev, and Evgeny Burnaev. Arigraph: Learning knowledge graph world models with episodic memory for llm agents. *arXiv preprint arXiv:2407.04363*, 2024.

Anthropic. Model context protocol. `https://github.com/modelcontextprotocol`, 2024. GitHub repository, an open protocol for connecting LLMs with external data sources and tools.

Anthropic. Claude agent skills. `https://github.com/anthropics/skills`, 2025. GitHub repository.

Rahul Aralikatte, Shashi Narayan, Joshua Maynez, Sascha Rothe, and Ryan McDonald. Focus attention: Promoting faithfulness and diversity in summarization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 6078–6095, 2021.

RM Aratchige and WMKS Ilmini. Llms working in harmony: A survey on the technological aspects of building effective llm-based multi agent systems. *arXiv preprint arXiv:2504.01963*, 2025.

Daman Arora, Atharv Sonwane, Nalin Wadhwa, Abhav Mehrotra, Saiteja Utpala, Ramakrishna Bairi, Aditya Kanade, and Nagarajan Natarajan. Masai: Modular architecture for software-engineering ai agents. *arXiv preprint arXiv:2406.11638*, 2024.

Richard C Atkinson and Richard M Shiffrin. Human memory: A proposed system and its control processes. In *Psychology of learning and motivation*, volume 2, pp. 89–195. Elsevier, 1968.

Alan Baddeley. The episodic buffer: a new component of working memory? *Trends in cognitive sciences*, 4 (11):417–423, 2000.

Alan Baddeley. Working memory. *Memory*, pp. 71–111, 2020.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Victor Barres, Honghua Dong, Soham Ray, Xujie Si, and Karthik Narasimhan. τ^2-bench: Evaluating conversational agents in a dual-control environment. *arXiv preprint arXiv:2506.07982*, 2025.

Ali Behrouz, Peilin Zhong, and Vahab Mirrokni. Titans: Learning to memorize at test time. *arXiv preprint arXiv:2501.00663*, 2024.

Yuanchen Bei, Weizhi Zhang, Siwen Wang, Weizhi Chen, Sheng Zhou, Hao Chen, Yong Li, Jiajun Bu, Shirui Pan, Yizhou Yu, et al. Graphs meet ai agents: Taxonomy, progress, and future opportunities. *arXiv preprint arXiv:2506.18019*, 2025.

Yuanchen Bei, Tianxin Wei, Xuying Ning, Yanjun Zhao, Zhining Liu, Xiao Lin, Yada Zhu, Hendrik Hamann, Jingrui He, and Hanghang Tong. Mem-gallery: Benchmarking multimodal long-term conversational memory for mllm agents. *arXiv preprint arXiv:2601.03515*, 2026.

Maksym Bekuzarov, Ariana Bermudez, Joon-Young Lee, and Hao Li. Xmem++: Production-level video segmentation from few annotated frames. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 635–644, 2023.

Jack Bell, Luigi Quarantiello, Eric Nuertey Coleman, Lanpei Li, Malio Li, Mauro Madeddu, Elia Piccoli, and Vincenzo Lomonaco. The future of continual learning in the era of foundation models: Three key directions. *arXiv preprint arXiv:2506.03320*, 2025.

Chandrachur Bhattacharya and Sibendu Som. Aisac: An integrated multi-agent system for transparent, retrieval-grounded scientific assistance. *arXiv preprint arXiv:2511.14043*, 2025.

Johan Bjorck, Fernando Castañeda, Nikita Cherniadev, Xingye Da, Runyu Ding, Linxi Fan, Yu Fang, Dieter Fox, Fengyuan Hu, Spencer Huang, et al. Gr00t n1: An open foundation model for generalist humanoid robots. *arXiv preprint arXiv:2503.14734*, 2025.

Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Robert Equi, Chelsea Finn, Niccolo Fusai, Manuel Y Galliker, et al. $\pi_{0.5}$: a vision-language-action model with open-world generalization. In *Proceedings of The 9th Conference on Robot Learning*, volume 305, pp. 17–40. PMLR, 2025.

Andrew Blair-Stanek, Nils Holzenberger, and Benjamin Van Durme. Can llms identify tax abuse? *arXiv preprint arXiv:2508.20097*, 2025.

Tim VP Bliss and Terje Lømo. Long-lasting potentiation of synaptic transmission in the dentate area of the anaesthetized rabbit following stimulation of the perforant path. *The Journal of physiology*, 232(2): 331–356, 1973.

Weihao Bo, Shan Zhang, Yanpeng Sun, Jingjing Wu, Qunyi Xie, Xiao Tan, Kunbin Chen, Wei He, Xiaofan Li, Na Zhao, et al. Agentic learner with grow-and-refine multimodal semantic memory. *arXiv preprint arXiv:2511.21678*, 2025.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gasic. Multiwoz-a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 5016–5026, 2018.

Mikhail S Burtsev, Yuri Kuratov, Anton Peganov, and Grigory V Sapunov. Memory transformer. *arXiv preprint arXiv:2006.11527*, 2020.

Hongru Cai, Yongqi Li, Wenjie Wang, Fengbin Zhu, Xiaoyu Shen, Wenjie Li, and Tat-Seng Chua. Large language models empowered personalized web agents. In *Proceedings of the ACM on Web Conference 2025*, pp. 198–215, 2025.

Zefan Cai, Yichi Zhang, Bofei Gao, Yuliang Liu, Yucheng Li, Tianyu Liu, Keming Lu, Wayne Xiong, Yue Dong, Junjie Hu, et al. Pyramidkv: Dynamic kv cache compression based on pyramidal information funneling. *arXiv preprint arXiv:2406.02069*, 2024.

Bowen Cao, Deng Cai, and Wai Lam. Infiniteicl: Breaking the limit of context window size via long short-term memory transformation. *arXiv preprint arXiv:2504.01707*, 2025a.

Jiaqi Cao, Jiarui Wang, Rubin Wei, Qipeng Guo, Kai Chen, Bowen Zhou, and Zhouhan Lin. Memory decoder: A pretrained, plug-and-play memory for large language models. *arXiv preprint arXiv:2508.09874*, 2025b.

Jack David Carson and Amir Reisizadeh. A statistical physics of language model reasoning. *arXiv preprint arXiv:2506.04374*, 2025.

Victor de Lamo Castrillo, Habtom Kahsay Gidey, Alexander Lenz, and Alois Knoll. Fundamentals of building autonomous llm agents. *arXiv preprint arXiv:2510.09244*, 2025.

cauri. Memory in multi-agent systems: Technical implementations. `https://artium.ai/insights/memory-in-multi-agent-systems-technical-implementations`, 2025. Artium.ai Insights article.

Dake Chen, Hanbin Wang, Yunhao Huo, Yuzhao Li, and Haoyang Zhang. Gamegpt: Multi-agent collaborative framework for game development. *arXiv preprint arXiv:2310.08067*, 2023.

Ding Chen, Simin Niu, Kehang Li, Peng Liu, Xiangping Zheng, Bo Tang, Xinchi Li, Feiyu Xiong, and Zhiyu Li. Halumem: Evaluating hallucinations in memory systems of agents. *arXiv preprint arXiv:2511.03506*, 2025a.

Guangyao Chen, Siwei Dong, Yu Shu, Ge Zhang, Jaward Sesay, Börje Karlsson, Jie Fu, and Yemin Shi. Autoagents: a framework for automatic agent generation. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pp. 22–30, 2024a.

Guhong Chen, Liyang Fan, Zihan Gong, Nan Xie, Zixuan Li, Ziqiang Liu, Chengming Li, Qiang Qu, Hamid Alinejad-Rokny, Shiwen Ni, et al. Agentcourt: Simulating court with adversarial evolvable lawyer agents. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 5850–5865, 2025b.

Guoxin Chen, Zile Qiao, Xuanzhong Chen, Donglei Yu, Haotian Xu, Wayne Xin Zhao, Ruihua Song, Wenbiao Yin, Huifeng Yin, Liwen Zhang, et al. Iterresearch: Rethinking long-horizon agents via markovian state reconstruction. *arXiv preprint arXiv:2511.07327*, 2025c.

Hongshen Chen, Xiaorui Liu, Dawei Yin, and Jiliang Tang. A survey on dialogue systems: Recent advances and new frontiers. *Acm Sigkdd Explorations Newsletter*, 19(2):25–35, 2017.

Jiangjie Chen, Xintao Wang, Rui Xu, Siyu Yuan, Yikai Zhang, Wei Shi, Jian Xie, Shuang Li, Ruihan Yang, Tinghui Zhu, et al. From persona to personalization: A survey on role-playing language agents. *arXiv preprint arXiv:2404.18231*, 2024b.

Joya Chen, Zhaoyang Lv, Shiwei Wu, Kevin Qinghong Lin, Chenan Song, Difei Gao, Jia-Wei Liu, Ziteng Gao, Dongxing Mao, and Mike Zheng Shou. Videollm-online: Online video large language model for streaming video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18407–18418, 2024c.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating large language models trained on code. *arXiv preprint arXiv:2110.14168*, 2021.

Nuo Chen, Hongguang Li, Jianhui Chang, Juhua Huang, Baoyuan Wang, and Jia Li. Compress to impress: Unleashing the potential of compressive memory in real-world long-term conversations. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 755–773, 2025d.

Xu Chen, Hongteng Xu, Yongfeng Zhang, Jiaxi Tang, Yixin Cao, Zheng Qin, and Hongyuan Zha. Sequential recommendation with user memory networks. In *Proceedings of the eleventh ACM international conference on web search and data mining*, pp. 108–116, 2018.

Yaoqi Chen, Jinkai Zhang, Baotong Lu, Qianxi Zhang, Chengruidong Zhang, Jingjia Luo, Di Liu, Huiqiang Jiang, Qi Chen, Jing Liu, et al. Retroinfer: A vector-storage approach for scalable long-context llm inference. *arXiv preprint arXiv:2505.02922*, 2025e.

Yi-Pei Chen, Noriki Nishida, Hideki Nakayama, and Yuji Matsumoto. Recent trends in personalized dialogue generation: A review of datasets, methodologies, and evaluations. *arXiv preprint arXiv:2405.17974*, 2024d.

Yixing Chen, Yiding Wang, Siqi Zhu, Haofei Yu, Tao Feng, Muhan Zhang, Mostofa Patwary, and Jiaxuan You. Multi-agent evolve: Llm self-improve through co-evolution. *arXiv preprint arXiv:2510.23595*, 2025f.

Weihua Cheng, Ersheng Ni, Wenlong Wang, Yifei Sun, Junming Liu, Wangyu Shen, Yirong Chen, Botian Shi, and Ding Wang. Mga: Memory-driven gui agent for observation-centric interaction. *arXiv preprint arXiv:2510.24168*, 2025.

Vinaik Chhetri, AB Siddique, and Umar Farooq. Understanding robustness of model editing in code llms: An empirical study. *arXiv preprint arXiv:2511.03182*, 2025.

Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. Mem0: Building production-ready ai agents with scalable long-term memory. *arXiv preprint arXiv:2504.19413*, 2025.

Konstantina Christakopoulou, Shibl Mourad, and Maja Matarić. Agents thinking fast and slow: A talker-reasoner architecture. *arXiv preprint arXiv:2410.08328*, 2024.

Zhendong Chu, Shen Wang, Jian Xie, Tinghui Zhu, Yibo Yan, Jingheng Ye, Aoxiao Zhong, Xuming Hu, Jing Liang, Philip S. Yu, and Qingsong Wen. LLM agents for education: Advances and applications. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 13782–13810, Suzhou, China, November 2025. Association for Computational Linguistics.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Neal J Cohen and Larry R Squire. Preserved learning and retention of pattern-analyzing skill in amnesia: Dissociation of knowing how and knowing that. *Science*, 210(4466):207–210, 1980.

Martin A. Conway and Christopher W. Pleydell-Pearce. The construction of autobiographical memories in the self-memory system. *Psychological Review*, 107(2):261–288, 2000.

Marc-Alexandre Côté, Akos Kádár, Xingdi Yuan, Ben Kybartas, Tavian Barnes, Emery Fine, James Moore, Matthew Hausknecht, Layla El Asri, Mahmoud Adada, et al. Textworld: A learning environment for text-based games. In *Workshop on Computer Games*, pp. 41–75. Springer, 2018.

Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, and Jimmy Lin. Ms marco: Benchmarking ranking models in the large-data regime. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pp. 1566–1576, 2021.

Justin Cui, Kevin Pu, and Tovi Grossman. Loom: Personalized learning informed by daily llm conversations toward long-term mastery via a dynamic learner memory graph. *arXiv preprint arXiv:2511.21037*, 2025a.

Yu Cui, Hang Fu, Haibin Zhang, Licheng Wang, and Cong Zuo. Free-mad: Consensus-free multi-agent debate. *arXiv preprint arXiv:2509.11035*, 2025b.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime G Carbonell, Quoc Le, and Ruslan Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 2978–2988, 2019.

Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Aleš Leonardis, Gregory Slabaugh, and Tinne Tuytelaars. A continual learning survey: Defying forgetting in classification tasks. *IEEE transactions on pattern analysis and machine intelligence*, 44(7):3366–3385, 2021.

Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Sam Stevens, Boshi Wang, Huan Sun, and Yu Su. Mind2web: Towards a generalist agent for the web. *Advances in Neural Information Processing Systems*, 36:28091–28114, 2023.

Xiang Deng, Jeff Da, Edwin Pan, Yannis Yiming He, Charles Ide, Kanak Garg, Niklas Lauffer, Andrew Park, Nitin Pasari, Chetan Rane, et al. Swe-bench pro: Can ai agents solve long-horizon software engineering tasks? *arXiv preprint arXiv:2509.16941*, 2025.

Darshan Deshpande, Varun Gangal, Hersh Mehta, Anand Kannappan, Rebecca Qian, and Peng Wang. Memtrack: Evaluating long-term memory and state tracking in multi-platform dynamic agent environments. *arXiv preprint arXiv:2510.01353*, 2025.

Shangzhe Di, Zhelun Yu, Guanghao Zhang, Haoyuan Li, Tao Zhong, Hao Cheng, Bolin Li, Wanggui He, Fangxun Shu, and Hao Jiang. Streaming video question-answering with in-context video kv-cache retrieval. *arXiv preprint arXiv:2503.00540*, 2025.

Frederick Dillon, Gregor Halvorsen, Simon Tattershall, Magnus Rowntree, and Gareth Vanderpool. Contextual memory reweaving in large language models using layered latent state reconstruction. *arXiv preprint arXiv:2502.02046*, 2025.

Shen Dong, Shaochen Xu, Pengfei He, Yige Li, Jiliang Tang, Tianming Liu, Hui Liu, and Zhen Xiang. A practical memory injection attack against llm agents. *arXiv preprint arXiv:2503.03704*, 2025.

Alexandre Drouin, Maxime Gasse, Massimo Caccia, Issam H Laradji, Manuel Del Verme, Tom Marty, David Vazquez, Nicolas Chapados, and Alexandre Lacoste. Workarena: how capable are web agents at solving common knowledge work tasks? In *Proceedings of the 41st International Conference on Machine Learning*, pp. 11642–11662, 2024.

Yiming Du, Hongru Wang, Zhengyi Zhao, Bin Liang, Baojun Wang, Wanjun Zhong, Zezhong Wang, and Kam-Fai Wong. Perltqa: A personal long-term memory dataset for memory classification, retrieval, and fusion in question answering. In *Proceedings of the 10th SIGHAN Workshop on Chinese Language Processing (SIGHAN-10)*, pp. 152–164, 2024.

Yiming Du, Wenyu Huang, Danna Zheng, Zhaowei Wang, Sebastien Montella, Mirella Lapata, Kam-Fai Wong, and Jeff Z Pan. Rethinking memory in ai: Taxonomy, operations, topics, and future directions. *arXiv preprint arXiv:2505.00675*, 2025.

Matthew Dunn, Levent Sagun, Mike Higgins, V Ugur Guney, Volkan Cirik, and Kyunghyun Cho. Searchqa: A new q&a dataset augmented with context from a search engine. *arXiv preprint arXiv:1704.05179*, 2017.

Esin Durmus, Faisal Ladhak, and Tatsunori B Hashimoto. Spurious correlations in reference-free evaluation of text generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1443–1454, 2022.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*, 2024.

Gilles O Einstein and Mark A McDaniel. Normal aging and prospective memory. *Journal of Experimental Psychology: Learning, memory, and cognition*, 16(4):717, 1990.

Andrew Estornell, Jean-Francois Ton, Yuanshun Yao, and Yang Liu. ACC-collab: An actor-critic approach to multi-agent LLM collaboration. In *The Thirteenth International Conference on Learning Representations*, 2025.

Linxi Fan, Guanzhi Wang, Yunfan Jiang, Ajay Mandlekar, Yuncong Yang, Haoyi Zhu, Andrew Tang, De-An Huang, Yuke Zhu, and Anima Anandkumar. Minedojo: Building open-ended embodied agents with internet-scale knowledge. *Advances in Neural Information Processing Systems*, 35:18343–18362, 2022.

Wei Fan, JinYi Yoon, and Bo Ji. imad: Intelligent multi-agent debate for efficient and accurate llm inference. *arXiv preprint arXiv:2511.11306*, 2025a.

Wenzhe Fan, Ning Yan, and Masood Mortazavi. Evomem: Improving multi-agent planning with dual-evolving memory. *arXiv preprint arXiv:2511.01912*, 2025b.

Yue Fan, Xiaojian Ma, Rujie Wu, Yuntao Du, Jiaqi Li, Zhi Gao, and Qing Li. Videoagent: A memory-augmented multimodal agent for video understanding. In *European Conference on Computer Vision*, pp. 75–92. Springer, 2024.

Zhiyu Fan, Kirill Vasilevski, Dayi Lin, Boyuan Chen, Yihao Chen, Zhiqing Zhong, Jie M Zhang, Pinjia He, and Ahmed E Hassan. Swe-effi: Re-evaluating software ai agent system effectiveness under resource constraints. *arXiv preprint arXiv:2509.09853*, 2025c.

Jizhan Fang, Xinle Deng, Haoming Xu, Ziyan Jiang, Yuqi Tang, Ziwen Xu, Shumin Deng, Yunzhi Yao, Mengru Wang, Shuofei Qiao, et al. Lightmem: Lightweight and efficient memory-augmented generation. *arXiv preprint arXiv:2510.18866*, 2025a.

Runnan Fang, Yuan Liang, Xiaobin Wang, Jialong Wu, Shuofei Qiao, Pengjun Xie, Fei Huang, Huajun Chen, and Ningyu Zhang. Memp: Exploring agent procedural memory. *arXiv preprint arXiv:2508.06433*, 2025b.

Tao Feng, Yexin Wu, Guanyu Lin, and Jiaxuan You. Graph world model. *arXiv preprint arXiv:2507.10539*, 2025a.

Xueyang Feng, Zhi-Yuan Chen, Yujia Qin, Yankai Lin, Xu Chen, Zhiyuan Liu, and Ji-Rong Wen. Large language model-based human-agent collaboration for complex task solving. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 1336–1357, 2024.

Yuan Feng, Junlin Lv, Yukun Cao, Xike Xie, and S Kevin Zhou. Identify critical kv cache in llm inference from an output perturbation perspective. *arXiv preprint arXiv:2502.03805*, 2025b.

Sandro Rama Fiorini, Leonardo G Azevedo, Raphael M Thiago, Valesca M de Sousa, Anton B Labate, and Viviane Torres da Silva. Episodic memory in agentic frameworks: Suggesting next tasks. *arXiv preprint arXiv:2511.17775*, 2025.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 24108–24118, 2025.

Pascale Fung, Yoram Bachrach, Asli Celikyilmaz, Kamalika Chaudhuri, Delong Chen, Willy Chung, Emmanuel Dupoux, Hongyu Gong, Hervé Jégou, Alessandro Lazaric, et al. Embodied ai agents: Modeling the world. *arXiv preprint arXiv:2506.22355*, 2025.

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.

Huan-ang Gao, Jiayi Geng, Wenyue Hua, Mengkang Hu, Xinzhe Juan, Hongzhang Liu, Shilong Liu, Jiahao Qiu, Xuan Qi, Yiran Wu, et al. A survey of self-evolving agents: On path to artificial super intelligence. *arXiv preprint arXiv:2507.21046*, 2025a.

Weibo Gao, Qi Liu, Linan Yue, Fangzhou Yao, Rui Lv, Zheng Zhang, Hao Wang, and Zhenya Huang. Agent4edu: Generating learner response data by generative agents for intelligent education systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 23923–23932, 2025b.

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.

Sebastian Gehrmann, Elizabeth Clark, and Thibault Sellam. Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text. *Journal of Artificial Intelligence Research*, 77:103–166, 2023.

Alireza Ghafarollahi and Markus J Buehler. Sciagents: automating scientific discovery through bioinspired multi-agent intelligent graph reasoning. *Advanced Materials*, 37(22):2413523, 2025.

Marc Glocker, Peter Hönig, Matthias Hirschmanner, and Markus Vincze. Llm-empowered embodied agent for memory-augmented task planning in household robotics. *arXiv preprint arXiv:2504.21716*, 2025.

Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Anil Palepu, Petar Sirkovic, Artiom Myaskovsky, Felix Weissenberger, Keran Rong, Ryutaro Tanno, et al. Towards an ai co-scientist. *arXiv preprint arXiv:2502.18864*, 2025.

Zhengyao Gu, Henry Peng Zou, Yankai Chen, Aiwei Liu, Weizhi Zhang, and Philip S Yu. Scaling laws for many-shot in-context learning with self-generated annotations. *arXiv preprint arXiv:2503.03062*, 2025.

Gunshi Gupta, Karmesh Yadav, Zsolt Kira, Yarin Gal, and Rahaf Aljundi. Memo: Training memory-efficient embodied agents with reinforcement learning. *arXiv preprint arXiv:2510.19732*, 2025.

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.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. Retrieval augmented language model pre-training. In *International conference on machine learning*, pp. 3929–3938. PMLR, 2020.

David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2(3), 2018.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023.

Dongge Han, Camille Couturier, Daniel Madrigal Diaz, Xuchao Zhang, Victor Rühle, and Saravan Rajmohan. Legomem: Modular procedural memory for multi-agent llm systems for workflow automation. *arXiv preprint arXiv:2510.04851*, 2025.

Shanshan Han, Qifan Zhang, Yuhang Yao, Weizhao Jin, and Zhaozhuo Xu. Llm multi-agent systems: Challenges and open problems. *arXiv preprint arXiv:2402.03578*, 2024.

Shibo Hao, Yi Gu, Haodi Ma, Joshua Hong, Zhen Wang, Daisy Wang, and Zhiting Hu. Reasoning with language model is planning with world model. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 8154–8173, 2023.

Kostas Hatalis, Despina Christou, Joshua Myers, Steven Jones, Keith Lambert, Adam Amos-Binks, Zohreh Dannenhauer, and Dustin Dannenhauer. Memory matters: The need to improve long-term memory in llm-agents. In *Proceedings of the AAAI Symposium Series*, volume 2, pp. 277–280, 2023.

Matthew Hausknecht, Prithviraj Ammanabrolu, Marc-Alexandre Côté, and Xingdi Yuan. Interactive fiction games: A colossal adventure. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 7903–7910, 2020.

Hongliang He, Wenlin Yao, Kaixin Ma, Wenhao Yu, Yong Dai, Hongming Zhang, Zhenzhong Lan, and Dong Yu. Webvoyager: Building an end-to-end web agent with large multimodal models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6864–6890, 2024a.

Junqing He, Liang Zhu, Rui Wang, Xi Wang, Gholamreza Haffari, and Jiaxing Zhang. Madial-bench: Towards real-world evaluation of memory-augmented dialogue generation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 9902–9921, 2025a.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Shiqi He, Yue Cui, Xinyu Ma, Yaliang Li, Bolin Ding, and Mosharaf Chowdhury. Branch-and-browse: Efficient and controllable web exploration with tree-structured reasoning and action memory. *arXiv preprint arXiv:2510.19838*, 2025b.

Zifan He, Yingqi Cao, Zongyue Qin, Neha Prakriya, Yizhou Sun, and Jason Cong. Hmt: Hierarchical memory transformer for efficient long context language processing. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 8068–8089, 2025c.

Zihong He, Weizhe Lin, Hao Zheng, Fan Zhang, Matt W Jones, Laurence Aitchison, Xuhai Xu, Miao Liu, Per Ola Kristensson, and Junxiao Shen. Human-inspired perspectives: A survey on ai long-term memory. *arXiv preprint arXiv:2411.00489*, 2024b.

Donald Olding Hebb. *The organization of behavior: A neuropsychological theory*. Psychology press, 2005.

Dan Hendrycks, Steven Basart, Saurav Kadavath, Mantas Mazeika, Akul Arora, Ethan Guo, Collin Burns, Samir Puranik, Horace He, Dawn Song, and Jacob Steinhardt. Measuring coding challenge competence with APPS. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021a.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021b.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2021c.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing a multi-hop qa dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 6609–6625, 2020.

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, et al. Metagpt: Meta programming for a multi-agent collaborative framework. In *The Twelfth International Conference on Learning Representations*, 2023.

Yining Hong, Rui Sun, Bingxuan Li, Xingcheng Yao, Maxine Wu, Alexander Chien, Da Yin, Ying Nian Wu, Zhecan James Wang, and Kai-Wei Chang. Embodied web agents: Bridging physical-digital realms for integrated agent intelligence. *arXiv preprint arXiv:2506.15677*, 2025.

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*, pp. 1–7, 2024.

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.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

He Hu, Yucheng Zhou, Chiyuan Ma, Qianning Wang, Zheng Zhang, Fei Ma, Laizhong Cui, and Qi Tian. Theramind: A strategic and adaptive agent for longitudinal psychological counseling. *arXiv preprint arXiv:2510.25758*, 2025a.

Mengkang Hu, Tianxing Chen, Qiguang Chen, Yao Mu, Wenqi Shao, and Ping Luo. Hiagent: Hierarchical working memory management for solving long-horizon agent tasks with large language model. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 32779–32798, 2025b.

Yuanzhe Hu, Yu Wang, and Julian McAuley. Evaluating memory in llm agents via incremental multi-turn interactions. *arXiv preprint arXiv:2507.05257*, 2025c.

Yuyang Hu, Shichun Liu, Yanwei Yue, Guibin Zhang, Boyang Liu, Fangyi Zhu, Jiahang Lin, Honglin Guo, Shihan Dou, Zhiheng Xi, et al. Memory in the age of ai agents. *arXiv preprint arXiv:2512.13564*, 2025d.

Wenyue Hua, Lizhou Fan, Lingyao Li, Kai Mei, Jianchao Ji, Yingqiang Ge, Libby Hemphill, and Yongfeng Zhang. War and peace (waragent): Large language model-based multi-agent simulation of world wars. *arXiv preprint arXiv:2311.17227*, 2023.

Wenyue Hua, Xianjun Yang, Mingyu Jin, Zelong Li, Wei Cheng, Ruixiang Tang, and Yongfeng Zhang. Trustagent: Towards safe and trustworthy llm-based agents through agent constitution. In *Trustworthy Multi-modal Foundation Models and AI Agents (TiFA)*, 2024.

Jimin Huang, Mengxi Xiao, Dong Li, Zihao Jiang, Yuzhe Yang, Yifei Zhang, Lingfei Qian, Yan Wang, Xueqing Peng, Yang Ren, et al. Open-finllms: Open multimodal large language models for financial applications. *arXiv preprint arXiv:2408.11878*, 2024a.

Wei-Chieh Huang, Henry Peng Zou, Yaozu Wu, Dongyuan Li, Yankai Chen, Weizhi Zhang, Yangning Li, Angelo Zangari, Jizhou Guo, Chunyu Miao, et al. Deepresearchguard: Deep research with open-domain evaluation and multi-stage guardrails for safety. *arXiv preprint arXiv:2510.10994*, 2025a.

Xu Huang, Weiwen Liu, Xiaolong Chen, Xingmei Wang, Hao Wang, Defu Lian, Yasheng Wang, Ruiming Tang, and Enhong Chen. Understanding the planning of llm agents: A survey. *arXiv preprint arXiv:2402.02716*, 2024b.

Xu Huang, Jianxun Lian, Yuxuan Lei, Jing Yao, Defu Lian, and Xing Xie. Recommender ai agent: Integrating large language models for interactive recommendations. *ACM Transactions on Information Systems*, 43(4):1–33, 2025b.

Yizheng Huang and Jimmy Huang. A survey on retrieval-augmented text generation for large language models. *arXiv preprint arXiv:2404.10981*, 2024.

Yuxuan Huang, Yihang Chen, Haozheng Zhang, Kang Li, Huichi Zhou, Meng Fang, Linyi Yang, Xiaoguang Li, Lifeng Shang, Songcen Xu, et al. Deep research agents: A systematic examination and roadmap. *arXiv preprint arXiv:2506.18096*, 2025c.

Zhaopei Huang, Qifeng Dai, Guozheng Wu, Xiaopeng Wu, Kehan Chen, Chuan Yu, Xubin Li, Tiezheng Ge, Wenxuan Wang, and Qin Jin. Mem-pal: Towards memory-based personalized dialogue assistants for long-term user-agent interaction. *arXiv preprint arXiv:2511.13410*, 2025d.

Zhengjun Huang, Zhoujin Tian, Qintian Guo, Fangyuan Zhang, Yingli Zhou, Di Jiang, and Xiaofang Zhou. Licomemory: Lightweight and cognitive agentic memory for efficient long-term reasoning. *arXiv preprint arXiv:2511.01448*, 2025e.

Luis Ibanez-Lissen, Lorena Gonzalez-Manzano, Jose Maria de Fuentes, Nicolas Anciaux, and Joaquin Garcia-Alfaro. Lumia: Linear probing for unimodal and multimodal membership inference attacks leveraging internal llm states. In *European Symposium on Research in Computer Security*, pp. 186–206. Springer, 2025.

Yoichi Ishibashi and Yoshimasa Nishimura. Self-organized agents: A llm multi-agent framework toward ultra large-scale code generation and optimization. *arXiv preprint arXiv:2404.02183*, 2024.

Md Ashraful Islam, Mohammed Eunus Ali, and Md Rizwan Parvez. Mapcoder: Multi-agent code generation for competitive problem solving. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4912–4944, 2024.

Samy Jelassi, David Brandfonbrener, Sham M Kakade, et al. Repeat after me: Transformers are better than state space models at copying. In *Forty-first International Conference on Machine Learning*, 2024.

Shian Jia, Ziyang Huang, Xinbo Wang, Haofei Zhang, and Mingli Song. Pisa: A pragmatic psych-inspired unified memory system for enhanced ai agency. *arXiv preprint arXiv:2510.15966*, 2025a.

Zixi Jia, Qinghua Liu, Hexiao Li, Yuyan Chen, and Jiqiang Liu. Evaluating the long-term memory of large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 19759–19777, 2025b.

Bowen Jiang, Zhuoqun Hao, Young-Min Cho, Bryan Li, Yuan Yuan, Sihao Chen, Lyle Ungar, Camillo J Taylor, and Dan Roth. Know me, respond to me: Benchmarking llms for dynamic user profiling and personalized responses at scale. *arXiv preprint arXiv:2504.14225*, 2025a.

Bowen Jiang, Yuan Yuan, Maohao Shen, Zhuoqun Hao, Zhangchen Xu, Zichen Chen, Ziyi Liu, Anvesh Rao Vijjini, Jiashu He, Hanchao Yu, et al. Personamem-v2: Towards personalized intelligence via learning implicit user personas and agentic memory. *arXiv preprint arXiv:2512.06688*, 2025b.

Jinhao Jiang, Kun Zhou, Wayne Xin Zhao, Yang Song, Chen Zhu, Hengshu Zhu, and Ji-Rong Wen. Kg-agent: An efficient autonomous agent framework for complex reasoning over knowledge graph. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9505–9523, 2025c.

Tao Jiang, Zichuan Lin, Lihe Li, Yi-Chen Li, Cong Guan, Lei Yuan, Zongzhang Zhang, Yang Yu, and Deheng Ye. Multi-agent in-context coordination via decentralized memory retrieval. *arXiv preprint arXiv:2511.10030*, 2025d.

Carlos E Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik R Narasimhan. Swe-bench: Can language models resolve real-world github issues? In *The Twelfth International Conference on Learning Representations*, 2025.

Bernal Jimenez Gutierrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. Hipporag: Neurobiologically inspired long-term memory for large language models. *Advances in Neural Information Processing Systems*, 37:59532–59569, 2024.

Hyoungwook Jin, Minju Yoo, Jeongeon Park, Yokyung Lee, Xu Wang, and Juho Kim. Teachtune: Reviewing pedagogical agents against diverse student profiles with simulated students. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pp. 1–28, 2025a.

Long Jin, Yang Chen, Tianyi Wang, Pan Hui, and Athanasios V Vasilakos. Understanding user behavior in online social networks: A survey. *IEEE communications magazine*, 51(9):144–150, 2013.

Qiao Jin, Zhizheng Wang, Yifan Yang, Qingqing Zhu, Donald Wright, Thomas Huang, Nikhil Khandekar, Nicholas Wan, Xuguang Ai, W John Wilbur, et al. Agentmd: Empowering language agents for risk prediction with large-scale clinical tool learning. *Nature Communications*, 16(1):9377, 2025b.

Zhi Jing, Yongye Su, and Yikun Han. When large language models meet vector databases: A survey. In *2025 Conference on Artificial Intelligence x Multimedia (AIxMM)*, pp. 7–13. IEEE, 2025.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1601–1611, 2017.

Sheena A Josselyn and Susumu Tonegawa. Memory engrams: Recalling the past and imagining the future. *Science*, 367(6473):eaaw4325, 2020.

Lars Benedikt Kaesberg, Jonas Becker, Jan Philip Wahle, Terry Ruas, and Bela Gipp. Voting or consensus? decision-making in multi-agent debate. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 11640–11671. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-acl.606.

Tomoyuki Kagaya, Thong Jing Yuan, Yuxuan Lou, Jayashree Karlekar, Sugiri Pranata, Akira Kinose, Koki Oguri, Felix Wick, and Yang You. Rap: Retrieval-augmented planning with contextual memory for multimodal llm agents. *arXiv preprint arXiv:2402.03610*, 2024.

Tomoyuki Kagaya, Subramanian Lakshmi, Anbang Ye, Thong Jing Yuan, Jayashree Karlekar, Sugiri Pranata, Natsuki Murakami, Akira Kinose, and Yang You. Vireskill: Vision-grounded replanning with skill memory for llm-based planning in lifelong robot learning. *arXiv preprint arXiv:2509.24219*, 2025.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. Large language models struggle to learn long-tail knowledge. In *International conference on machine learning*, pp. 15696–15707. PMLR, 2023.

Jiazheng Kang, Mingming Ji, Zhe Zhao, and Ting Bai. Memory os of ai agent. *arXiv preprint arXiv:2506.06326*, 2025a.

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*, 2025b.

Minki Kang, Wei-Ning Chen, Dongge Han, Huseyin A Inan, Lukas Wutschitz, Yanzhi Chen, Robert Sim, and Saravan Rajmohan. Acon: Optimizing context compression for long-horizon llm agents. *arXiv preprint arXiv:2510.00615*, 2025c.

Zixuan Ke, Weize Kong, Cheng Li, Mingyang Zhang, Qiaozhu Mei, and Michael Bendersky. Bridging the preference gap between retrievers and LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10438–10451, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.562.

Zixuan Ke, Fangkai Jiao, Yifei Ming, Xuan-Phi Nguyen, Austin Xu, Do Xuan Long, Minzhi Li, Chengwei Qin, PeiFeng Wang, silvio savarese, Caiming Xiong, and Shafiq Joty. A survey of frontiers in LLM reasoning: Inference scaling, learning to reason, and agentic systems. *Transactions on Machine Learning Research*, 2025a. ISSN 2835-8856. Survey Certification.

Zixuan Ke, Yifei Ming, Xuan-Phi Nguyen, Caiming Xiong, and Shafiq Joty. Demystifying domain-adaptive post-training for financial LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2025b. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1579.

Sam Keen. Implicit memory systems for llms. `https://deepengineering.substack.com/p/implicit-memory-systems-for-llms`, 2025. Deep Engineering (Substack) article.

Ronald Kemker and Christopher Kanan. FearNet: Brain-Inspired Model for Incremental Learning. In *ICLR*, 2018.

Byeonghwi Kim, Jinyeon Kim, Yuyeong Kim, Cheolhong Min, and Jonghyun Choi. Context-aware planning and environment-aware memory for instruction following embodied agents. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10936–10946, 2023a.

Donghyuk Kim, Sejeong Yang, Wonjin Shin, and Joo-Young Kim. V-rex: Real-time streaming video llm acceleration via dynamic kv cache retrieval. *arXiv preprint arXiv:2512.12284*, 2025a.

Jiho Kim, Woosog Chay, Hyeonji Hwang, Daeun Kyung, Hyunseung Chung, Eunbyeol Cho, Yohan Jo, and Edward Choi. Dialsim: A real-time simulator for evaluating long-term multi-party dialogue understanding of conversation systems. *arXiv preprint arXiv:2406.13144*, 2024a.

Minsoo Kim, Arnav Kundu, Han-Byul Kim, Richa Dixit, and Minsik Cho. Epicache: Episodic kv cache management for long conversational question answering. *arXiv preprint arXiv:2509.17396*, 2025b.

Taewoon Kim, Michael Cochez, Vincent François-Lavet, Mark Neerincx, and Piek Vossen. A machine with short-term, episodic, and semantic memory systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 48–56, 2023b.

Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik S Chan, Xuhai Xu, Daniel McDuff, Hyeonhoon Lee, Marzyeh Ghassemi, Cynthia Breazeal, and Hae W Park. Mdagents: An adaptive collaboration of llms for medical decision-making. *Advances in Neural Information Processing Systems*, 37:79410–79452, 2024b.

Shashank Kirtania, Param Biyani, Priyanshu Gupta, Yasharth Bajpai, Roshni Iyer, Sumit Gulwani, and Gustavo Soares. Improving language agents through brew. *arXiv preprint arXiv:2511.20297*, 2025.

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.

Ishant Kohar and Aswanth Krishnan. A benchmark for procedural memory retrieval in language agents. *arXiv preprint arXiv:2511.21730*, 2025.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466, 2019.

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*, pp. 611–626, 2023.

Anton Bulle Labate, Valesca Moura de Sousa, Sandro Rama Fiorini, Leonardo Guerreiro Azevedo, Raphael Melo Thiago, and Viviane Torres da Silva. Solving context window overflow in ai agents. *arXiv preprint arXiv:2511.22729*, 2025.

Paraskevas Lagakis and Stavros Demetriadis. Evaai: a multi-agent framework leveraging large language models for enhanced automated grading. In *International Conference on Intelligent Tutoring Systems*, pp. 378–385. Springer, 2024.

Yilong Lai, Yipin Yang, Jialong Wu, Fengran Mo, Zhenglin Wang, Ting Liang, Jianguo Lin, and Keping Yang. Crmweaver: Building powerful business agent via agentic rl and shared memories. *arXiv preprint arXiv:2510.25333*, 2025.

Tian Lan, Wenwei Zhang, Chengqi Lyu, Shuaibin Li, Chen Xu, Heyan Huang, Dahua Lin, Xian-Ling Mao, and Kai Chen. Training language models to critique with multi-agent feedback. *arXiv preprint arXiv:2410.15287*, 2024.

Chris Latimer, Nicoló Boschi, Andrew Neeser, Chris Bartholomew, Gaurav Srivastava, Xuan Wang, and Naren Ramakrishnan. Hindsight is 20/20: Building agent memory that retains, recalls, and reflects. *arXiv preprint arXiv:2512.12818*, 2025.

Gibbeum Lee, Volker Hartmann, Jongho Park, Dimitris Papailiopoulos, and Kangwook Lee. Prompted llms as chatbot modules for long open-domain conversation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 4536–4554, 2023.

Kuang-Huei Lee, Xinyun Chen, Hiroki Furuta, John Canny, and Ian Fischer. A human-inspired reading agent with gist memory of very long contexts. In *International Conference on Machine Learning*, pp. 26396–26415. PMLR, 2024.

Xiang Lei, Qin Li, and Min Zhang. D-smart: Enhancing llm dialogue consistency via dynamic structured memory and reasoning tree. *arXiv preprint arXiv:2510.13363*, 2025.

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.

Cheng Li, Ziang Leng, Chenxi Yan, Junyi Shen, Hao Wang, Weishi Mi, Yaying Fei, Xiaoyang Feng, Song Yan, HaoSheng Wang, et al. Chatharuhi: Reviving anime character in reality via large language model. *arXiv preprint arXiv:2308.09597*, 2023a.

Hao Li, Chenghao Yang, An Zhang, Yang Deng, Xiang Wang, and Tat-Seng Chua. Hello again! llm-powered personalized agent for long-term dialogue. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 5259–5276, 2025a.

Haohang Li, Yupeng Cao, Yangyang Yu, Shashidhar Reddy Javaji, Zhiyang Deng, Yueru He, Yuechen Jiang, Zining Zhu, Kp Subbalakshmi, Jimin Huang, et al. Investorbench: A benchmark for financial decision-making tasks with llm-based agent. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2509–2525, 2025b.

Huihan Li, Yuting Ning, Zeyi Liao, Siyuan Wang, Xiang Lorraine Li, Ximing Lu, Wenting Zhao, Faeze Brahman, Yejin Choi, and Xiang Ren. In search of the long-tail: Systematic generation of long-tail inferential knowledge via logical rule guided search. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 2348–2370, 2024a.

Jitang Li and Jinzheng Li. Memory, consciousness and large language model. *arXiv preprint arXiv:2401.02509*, 2024.

Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and William B Dolan. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 994–1003, 2016.

Manling Li, Shiyu Zhao, Qineng Wang, Kangrui Wang, Yu Zhou, Sanjana Srivastava, Cem Gokmen, Tony Lee, Erran Li Li, Ruohan Zhang, et al. Embodied agent interface: Benchmarking llms for embodied decision making. *Advances in Neural Information Processing Systems*, 37:100428–100534, 2024b.

Ruanjun Li, Yuedong Tan, Yuanming Shi, and Jiawei Shao. Videoscan: Enabling efficient streaming video understanding via frame-level semantic carriers. *arXiv preprint arXiv:2503.09387*, 2025c.

Rui Li, Zeyu Zhang, Xiaohe Bo, Zihang Tian, Xu Chen, Quanyu Dai, Zhenhua Dong, and Ruiming Tang. Cam: A constructivist view of agentic memory for llm-based reading comprehension. *arXiv preprint arXiv:2510.05520*, 2025d.

Rumeng Li, Xun Wang, Dan Berlowitz, Jesse Mez, Honghuang Lin, and Hong Yu. Care-ad: a multi-agent large language model framework for alzheimer's disease prediction using longitudinal clinical notes. *npj Digital Medicine*, 8(1):541, 2025e.

Sijia Li, Yuchen Huang, Zifan Liu, Zijian Li, Lei Song, Jiang Bian, Jun Zhang, Rui Wang, et al. Sit-graph: State integrated tool graph for multi-turn agents. *arXiv preprint arXiv:2512.07287*, 2025f.

Xiaoxi Li, Wenxiang Jiao, Jiarui Jin, Guanting Dong, Jiajie Jin, Yinuo Wang, Hao Wang, Yutao Zhu, Ji-Rong Wen, Yuan Lu, et al. Deepagent: A general reasoning agent with scalable toolsets. *arXiv preprint arXiv:2510.21618*, 2025g.

Xinyi Li, Sai Wang, Siqi Zeng, Yu Wu, and Yi Yang. A survey on llm-based multi-agent systems: workflow, infrastructure, and challenges. *Vicinagearth*, 1(1):9, 2024c.

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*, 2023b.

Yang Li, Zhiyuan He, Yuxuan Huang, Zhuhanling Xiao, Chao Yu, Meng Fang, Kun Shao, and Jun Wang. Adapting like humans: A metacognitive agent with test-time reasoning. *arXiv preprint arXiv:2511.23262*, 2025h.

Yangning Li, Weizhi Zhang, Yuyao Yang, Wei-Chieh Huang, Yaozu Wu, Junyu Luo, Yuanchen Bei, Henry Peng Zou, Xiao Luo, Yusheng Zhao, et al. Towards agentic rag with deep reasoning: A survey of rag-reasoning systems in llms. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 12120–12145, 2025i.

Yuan Li, Lichao Sun, and Yixuan Zhang. Metaagents: Large language model based agents for decision-making on teaming. *Proceedings of the ACM on Human-Computer Interaction*, 9(2):1–27, 2025j.

Yuhong Li, Yingbing Huang, Bowen Yang, Bharat Venkitesh, Acyr Locatelli, Hanchen Ye, Tianle Cai, Patrick Lewis, and Deming Chen. Snapkv: Llm knows what you are looking for before generation. *Advances in Neural Information Processing Systems*, 37:22947–22970, 2024d.

Zhiyu Li, Shichao Song, Chenyang Xi, Hanyu Wang, Chen Tang, Simin Niu, Ding Chen, Jiawei Yang, Chunyu Li, Qingchen Yu, et al. Memos: A memory os for ai system. *arXiv preprint arXiv:2507.03724*, 2025k.

Zhuofeng Li, Haoxiang Zhang, Seungju Han, Sheng Liu, Jianwen Xie, Yu Zhang, Yejin Choi, James Zou, and Pan Lu. In-the-flow agentic system optimization for effective planning and tool use. *arXiv preprint arXiv:2510.05592*, 2025l.

Zongwei Li, Zhonghang Li, Zirui Guo, Xubin Ren, and Chao Huang. Deepcode: Open agentic coding. *arXiv preprint arXiv:2512.07921*, 2025m.

Jiafeng Liang, Hao Li, Chang Li, Jiaqi Zhou, Shixin Jiang, Zekun Wang, Changkai Ji, Zhihao Zhu, Runxuan Liu, Tao Ren, et al. Ai meets brain: Memory systems from cognitive neuroscience to autonomous agents. *arXiv preprint arXiv:2512.23343*, 2025.

Kevin J Liang, Weituo Hao, Dinghan Shen, Yufan Zhou, Weizhu Chen, Changyou Chen, and Lawrence Carin. Mixkd: Towards efficient distillation of large-scale language models. *arXiv preprint arXiv:2011.00593*, 2020.

Jonathan Light, Min Cai, Sheng Shen, and Ziniu Hu. Avalonbench: Evaluating llms playing the game of avalon. *arXiv preprint arXiv:2310.05036*, 2023.

Jianzhe Lin, Zeyu Pan, Yun Zhu, Ruiqi Song, and Jining Yang. Towards continuous intelligence growth: Self-training, continual learning, and dual-scale memory in superintelliagent. *arXiv preprint arXiv:2511.23436*, 2025.

Chen Ling, Xujiang Zhao, Jiaying Lu, Chengyuan Deng, Can Zheng, Junxiang Wang, Tanmoy Chowdhury, Yun Li, Hejie Cui, Xuchao Zhang, et al. Domain specialization as the key to make large language models disruptive: A comprehensive survey. *ACM Computing Surveys*, 58(3):1–39, 2025.

Stacey Diane Arañez Litam, Clark D Ausloos, and John JS Harrichand. Stress and resilience among professional counselors during the covid-19 pandemic. *Journal of Counseling & Development*, 99(4):384–395, 2021.

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*, 2024a.

Bang Liu, Xinfeng Li, Jiayi Zhang, Jinlin Wang, Tanjin He, Sirui Hong, Hongzhang Liu, Shaokun Zhang, Kaitao Song, Kunlun Zhu, et al. Advances and challenges in foundation agents: From brain-inspired intelligence to evolutionary, collaborative, and safe systems. *arXiv preprint arXiv:2504.01990*, 2025a.

Genglin Liu, Shijie Geng, Sha Li, Hejie Cui, Sarah Zhang, Xin Liu, and Tianyi Liu. Webcoach: Self-evolving web agents with cross-session memory guidance. *arXiv preprint arXiv:2511.12997*, 2025b.

Genglin Liu, Vivian T Le, Salman Rahman, Elisa Kreiss, Marzyeh Ghassemi, and Saadia Gabriel. Mosaic: Modeling social ai for content dissemination and regulation in multi-agent simulations. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 6401–6428, 2025c.

Haofeng Liu, Mingqi Gao, Xuxiao Luo, Ziyue Wang, Guanyi Qin, Junde Wu, and Yueming Jin. Resurgsam2: Referring segment anything in surgical video via credible long-term tracking. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 435–445. Springer, 2025d.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023a.

Jiahong Liu, Zexuan Qiu, Zhongyang Li, Quanyu Dai, Wenhao Yu, Jieming Zhu, Minda Hu, Menglin Yang, Tat-Seng Chua, and Irwin King. A survey of personalized large language models: Progress and future directions. *arXiv preprint arXiv:2502.11528*, 2025e.

Jiarun Liu, Shiyue Xu, Yang Li, Shangkun Liu, Yongli Yu, and Peng Cao. Unifying dynamic tool creation and cross-task experience sharing through cognitive memory architecture. *arXiv preprint arXiv:2512.11303*, 2025f.

Junming Liu, Yifei Sun, Weihua Cheng, Haodong Lei, Yirong Chen, Licheng Wen, Xuemeng Yang, Daocheng Fu, Pinlong Cai, Nianchen Deng, et al. Memverse: Multimodal memory for lifelong learning agents. *arXiv preprint arXiv:2512.03627*, 2025g.

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*, 2023b.

Mugeng Liu, Siqi Zhong, Qi Yang, Yudong Han, Xuanzhe Liu, and Yun Ma. Webanns: Fast and efficient approximate nearest neighbor search in web browsers. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2483–2492, 2025h.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024b.

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*, 2025i.

Xiang Liu, Zhenheng Tang, Peijie Dong, Zeyu Li, Yue Liu, Bo Li, Xuming Hu, and Xiaowen Chu. Chunkkv: Semantic-preserving kv cache compression for efficient long-context llm inference. *arXiv preprint arXiv:2502.00299*, 2025j.

Zhiwei Liu, Weiran Yao, Jianguo Zhang, Liangwei Yang, Zuxin Liu, Juntao Tan, Prafulla K Choubey, Tian Lan, Jason Wu, Huan Wang, et al. Agentlite: A lightweight library for building and advancing task-oriented llm agent system. *arXiv preprint arXiv:2402.15538*, 2024c.

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, 2023c.

Lin Long, Yichen He, Wentao Ye, Yiyuan Pan, Yuan Lin, Hang Li, Junbo Zhao, and Wei Li. Seeing, listening, remembering, and reasoning: A multimodal agent with long-term memory. *arXiv preprint arXiv:2508.09736*, 2025.

David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. In *NIPS*, 2017.

Shangyu Lou. Urban-mas: Human-centered urban prediction with llm-based multi agent system. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Advances in Urban-AI*, pp. 37–40, 2025.

Jiarui Lu, Thomas Holleis, Yizhe Zhang, Bernhard Aumayer, Feng Nan, Haoping Bai, Shuang Ma, Shen Ma, Mengyu Li, Guoli Yin, et al. Toolsandbox: A stateful, conversational, interactive evaluation benchmark for llm tool use capabilities. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 1160–1183, 2025a.

Junfeng Lu and Yueyan Li. Dynamic affective memory management for personalized llm agents. *arXiv preprint arXiv:2510.27418*, 2025.

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.

Miao Lu, Weiwei Sun, Weihua Du, Zhan Ling, Xuesong Yao, Kang Liu, and Jiecao Chen. Scaling llm multi-turn rl with end-to-end summarization-based context management. *arXiv preprint arXiv:2510.06727*, 2025b.

Yaxi Lu, Shenzhi Yang, Cheng Qian, Guirong Chen, Qinyu Luo, Yesai Wu, Huadong Wang, Xin Cong, Zhong Zhang, Yankai Lin, Weiwen Liu, Yasheng Wang, Zhiyuan Liu, Fangming Liu, and Maosong Sun. Proactive agent: Shifting LLM agents from reactive responses to active assistance. In *The Thirteenth International Conference on Learning Representations*, 2025c.

Elias Lumer, Anmol Gulati, Vamse Kumar Subbiah, Pradeep Honaganahalli Basavaraju, and James A Burke. Memtool: Optimizing short-term memory management for dynamic tool calling in llm agent multi-turn conversations. *arXiv preprint arXiv:2507.21428*, 2025.

Haoyan Luo and Lucia Specia. From understanding to utilization: A survey on explainability for large language models. *arXiv preprint arXiv:2401.12874*, 2024.

Junyu Luo, Weizhi Zhang, Ye Yuan, Yusheng Zhao, Junwei Yang, Yiyang Gu, Bohan Wu, Binqi Chen, Ziyue Qiao, Qingqing Long, et al. Large language model agent: A survey on methodology, applications and challenges. *arXiv preprint arXiv:2503.21460*, 2025.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. Evaluating very long-term conversational memory of llm agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13851–13870, 2024.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9802–9822, 2023.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. *Advances in Neural Information Processing Systems*, 36: 46212–46244, 2023.

Reza Yousefi Maragheh and Yashar Deldjoo. The future is agentic: Definitions, perspectives, and open challenges of multi-agent recommender systems. *arXiv preprint arXiv:2507.02097*, 2025.

Alessio Maritan. Staffpro: an llm agent for joint staffing and profiling. *arXiv preprint arXiv:2507.21636*, 2025.

Vasilije Markovic, Lazar Obradovic, Laszlo Hajdu, and Jovan Pavlovic. Optimizing the interface between knowledge graphs and llms for complex reasoning. *arXiv preprint arXiv:2505.24478*, 2025.

Stephen J Martin, Paul D Grimwood, and Richard GM Morris. Synaptic plasticity and memory: an evaluation of the hypothesis. *Annual review of neuroscience*, 23(1):649–711, 2000.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372, 2022a.

Kevin Meng, Arnab Sen Sharma, Alex J Andonian, Yonatan Belinkov, and David Bau. Mass-editing memory in a transformer. In *The Eleventh International Conference on Learning Representations*, 2022b.

Grégoire Mialon, Clémentine Fourrier, Thomas Wolf, Yann LeCun, and Thomas Scialom. Gaia: a benchmark for general ai assistants. In *The Twelfth International Conference on Learning Representations*, 2024.

Chunyu Miao, Henry Peng Zou, Yangning Li, Yankai Chen, Yibo Wang, Fangxin Wang, Yifan Li, Wooseong Yang, Bowei He, Xinni Zhang, et al. Recode-h: A benchmark for research code development with interactive human feedback. *arXiv preprint arXiv:2510.06186*, 2025.

Julie Michelman, Nasrin Baratalipour, and Matthew Abueg. Enhancing reasoning with collaboration and memory. *arXiv preprint arXiv:2503.05944*, 2025.

Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. Large language models: A survey. *arXiv preprint arXiv:2402.06196*, 2024.

Yuhang Ming, Minyang Xu, Xingrui Yang, Weicai Ye, Weihan Wang, Yong Peng, Weichen Dai, and Wanzeng Kong. Viper: Visual incremental place recognition with adaptive mining and continual learning. *IEEE Robotics and Automation Letters*, 2025.

Niloofar Mireshghallah and Tianshi Li. Position: Privacy is not just memorization! *arXiv preprint arXiv:2510.01645*, 2025.

Samuel Miserendino, Michele Wang, Tejal Patwardhan, and Johannes Heidecke. SWE-lancer: Can frontier LLMs earn $1 million from real-world freelance software engineering? In *Forty-second International Conference on Machine Learning*, 2025.

Fengran Mo, Kelong Mao, Ziliang Zhao, Hongjin Qian, Haonan Chen, Yiruo Cheng, Xiaoxi Li, Yutao Zhu, Zhicheng Dou, and Jian-Yun Nie. A survey of conversational search. *ACM Transactions on Information Systems*, 43(6):1–50, 2025.

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.

Ali Modarressi, Abdullatif Köksal, Ayyoob Imani, Mohsen Fayyaz, and Hinrich Schütze. Memllm: Finetuning llms to use an explicit read-write memory. *arXiv preprint arXiv:2404.11672*, 2024.

Ruaridh Mon-Williams, Gen Li, Ran Long, Wenqian Du, and Christopher G Lucas. Embodied large language models enable robots to complete complex tasks in unpredictable environments. *Nature Machine Intelligence*, pp. 1–10, 2025.

Tsendsuren Munkhdalai, Manaal Faruqui, and Siddharth Gopal. Leave no context behind: Efficient infinite context transformers with infini-attention. *arXiv preprint arXiv:2404.07143*, 101, 2024.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*, 2021.

Jiayan Nan, Wenquan Ma, Wenlong Wu, and Yize Chen. Nemori: Self-organizing agent memory inspired by cognitive science. *arXiv preprint arXiv:2508.03341*, 2025.

Hongqiu Ni, Jiabao Zhang, Guopeng Li, Zilong Wang, Ruiqi Wu, Chi Zhang, and Haisheng Tan. Astraea: A state-aware scheduling engine for llm-powered agents. *arXiv preprint arXiv:2512.14142*, 2025.

NVIDIA. Nemo gym: An open source library for scaling reinforcement learning environments for llm. `https://github.com/NVIDIA-NeMo/Gym`, 2025. GitHub repository.

Parsa Omidi, Xingshuai Huang, Axel Laborieux, Bahareh Nikpour, Tianyu Shi, and Armaghan Eshaghi. Memory-augmented transformers: A systematic review from neuroscience principles to enhanced model architectures. *arXiv preprint arXiv:2508.10824*, 2025.

OpenAI. Memory and new controls for chatgpt, April 2025. URL `https://openai.com/index/memory-and-new-controls-for-chatgpt/`.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Siru Ouyang, Jun Yan, I Hsu, Yanfei Chen, Ke Jiang, Zifeng Wang, Rujun Han, Long T Le, Samira Daruki, Xiangru Tang, et al. Reasoningbank: Scaling agent self-evolving with reasoning memory. *arXiv preprint arXiv:2509.25140*, 2025.

Charles Packer, Vivian Fang, Shishir G. Patil, Kevin Lin, Sarah Wooders, and Joseph_E Gonzalez. Memgpt: Towards llms as operating systems. *arXiv preprint arXiv:2310.08560*, 2023.

Shankar Padmanabhan, Yasumasa Onoe, Michael Zhang, Greg Durrett, and Eunsol Choi. Propagating knowledge updates to lms through distillation. *Advances in Neural Information Processing Systems*, 36: 47124–47142, 2023.

Egor Pakhomov, Erik Nijkamp, and Caiming Xiong. Convomem benchmark: Why your first 150 conversations don't need rag. *arXiv preprint arXiv:2511.10523*, 2025.

Zhuoshi Pan, Qianhui Wu, Huiqiang Jiang, Xufang Luo, Hao Cheng, Dongsheng Li, Yuqing Yang, Chin-Yew Lin, H Vicky Zhao, Lili Qiu, et al. Secom: On memory construction and retrieval for personalized conversational agents. In *The Thirteenth International Conference on Learning Representations*, 2025.

Dmitrii Pantiukhin, Boris Shapkin, Antonia Anna Jost, Ivan Kuznetsov, and Nikolay V Koldunov. Accelerating earth science discovery via multi-agent llm systems. *Frontiers in Artificial Intelligence*, 8:1674927, 2025.

Zachary A Pardos, Matthew Tang, Ioannis Anastasopoulos, Shreya K Sheel, and Ethan Zhang. Oatutor: An open-source adaptive tutoring system and curated content library for learning sciences research. In *Proceedings of the 2023 chi conference on human factors in computing systems*, pp. 1–17, 2023.

German I Parisi, Ronald Kemker, Jose L Part, Christopher Kanan, and Stefan Wermter. Continual lifelong learning with neural networks: A review. *Neural networks*, 113:54–71, 2019.

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*, pp. 1–22, 2023.

Yoon-Joo Park and Alexander Tuzhilin. The long tail of recommender systems and how to leverage it. In *Proceedings of the 2008 ACM conference on Recommender systems*, pp. 11–18, 2008.

Atharv Singh Patlan, Peiyao Sheng, S Ashwin Hebbar, Prateek Mittal, and Pramod Viswanath. Real ai agents with fake memories: Fatal context manipulation attacks on web3 agents. *arXiv preprint arXiv:2503.16248*, 2025.

Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, et al. Kilt: a benchmark for knowledge intensive language tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2523–2544, 2021.

Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Chen Bo Calvin Zhang, Mohamed Shaaban, John Ling, Sean Shi, et al. Humanity's last exam. *arXiv preprint arXiv:2501.14249*, 2025.

Mathis Pink, Qinyuan Wu, Vy Ai Vo, Javier Turek, Jianing Mu, Alexander Huth, and Mariya Toneva. Position: Episodic memory is the missing piece for long-term llm agents. *arXiv preprint arXiv:2502.06975*, 2025.

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.

Hadi Pouransari, David Grangier, C Thomas, Michael Kirchhof, and Oncel Tuzel. Pretraining with hierarchical memories: separating long-tail and common knowledge. *arXiv preprint arXiv:2510.02375*, 2025.

Viraj Prabhu, Yutong Dai, Matthew Fernandez, Jing Gu, Krithika Ramakrishnan, Yanqi Luo, Silvio Savarese, Caiming Xiong, Junnan Li, Zeyuan Chen, et al. Walt: Web agents that learn tools. *arXiv preprint arXiv:2510.01524*, 2025.

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. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15174–15186, 2024a.

Chen Qian, Zihao Xie, YiFei Wang, Wei Liu, Kunlun Zhu, Hanchen Xia, Yufan Dang, Zhuoyun Du, Weize Chen, Cheng Yang, et al. Scaling large language model-based multi-agent collaboration. In *The Thirteenth International Conference on Learning Representations*, 2025a.

Hongjin Qian, Zheng Liu, Peitian Zhang, Kelong Mao, Defu Lian, Zhicheng Dou, and Tiejun Huang. Memorag: Boosting long context processing with global memory-enhanced retrieval augmentation. In *Proceedings of the ACM on Web Conference 2025*, pp. 2366–2377, 2025b.

Rui Qian, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Shuangrui Ding, Dahua Lin, and Jiaqi Wang. Streaming long video understanding with large language models. *Advances in Neural Information Processing Systems*, 37:119336–119360, 2024b.

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, et al. Toolllm: Facilitating large language models to master 16000+ real-world apis. In *The Twelfth International Conference on Learning Representations*, 2024.

Yujia Qin, Yining Ye, Junjie Fang, Haoming Wang, Shihao Liang, Shizuo Tian, Junda Zhang, Jiahao Li, Yunxin Li, Shijue Huang, et al. Ui-tars: Pioneering automated gui interaction with native agents. *arXiv preprint arXiv:2501.12326*, 2025.

Jielin Qiu, Zuxin Liu, Zhiwei Liu, Rithesh Murthy, Jianguo Zhang, Haolin Chen, Shiyu Wang, Ming Zhu, Liangwei Yang, Juntao Tan, et al. Locobench: A benchmark for long-context large language models in complex software engineering. *arXiv preprint arXiv:2509.09614*, 2025a.

Jielin Qiu, Zuxin Liu, Zhiwei Liu, Rithesh Murthy, Jianguo Zhang, Haolin Chen, Shiyu Wang, Ming Zhu, Liangwei Yang, Juntao Tan, et al. Locobench-agent: An interactive benchmark for llm agents in long-context software engineering. *arXiv preprint arXiv:2511.13998*, 2025b.

Derrick Quinn, Mohammad Nouri, Neel Patel, John Salihu, Alireza Salemi, Sukhan Lee, Hamed Zamani, and Mohammad Alian. Accelerating retrieval-augmented generation. In *Proceedings of the 30th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 1*, pp. 15–32, 2025.

Jack W Rae, Anna Potapenko, Siddhant M Jayakumar, and Timothy P Lillicrap. Compressive transformers for long-range sequence modelling. *arXiv preprint arXiv:1911.05507*, 2019.

Shreyas Rajesh, Pavan Holur, Chenda Duan, David Chong, and Vwani Roychowdhury. Beyond fact retrieval: Episodic memory for rag with generative semantic workspaces. *arXiv preprint arXiv:2511.07587*, 2025.

Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for squad. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 784–789, 2018.

Preston Rasmussen, Pavlo Paliychuk, Travis Beauvais, Jack Ryan, and Daniel Chalef. Zep: a temporal knowledge graph architecture for agent memory. *arXiv preprint arXiv:2501.13956*, 2025.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H. Lampert. icarl: Incremental classifier and representation learning. In *CVPR*, 2017.

Aniket Rege, Aditya Kusupati, Alan Fan, Qingqing Cao, Sham Kakade, Prateek Jain, Ali Farhadi, et al. Adanns: A framework for adaptive semantic search. *Advances in Neural Information Processing Systems*, 36:76311–76335, 2023.

Shuo Ren, Pu Jian, Zhenjiang Ren, Chunlin Leng, Can Xie, and Jiajun Zhang. Towards scientific intelligence: A survey of llm-based scientific agents. *arXiv preprint arXiv:2503.24047*, 2025.

Matthew Renze and Erhan Guven. Self-reflection in llm agents: Effects on problem-solving performance. *arXiv preprint arXiv:2405.06682*, 2024.

Alireza Rezazadeh, Zichao Li, Ange Lou, Yuying Zhao, Wei Wei, and Yujia Bao. Collaborative memory: Multi-user memory sharing in llm agents with dynamic access control. *arXiv preprint arXiv:2505.18279*, 2025a.

Alireza Rezazadeh, Zichao Li, Wei Wei, and Yujia Bao. From isolated conversations to hierarchical schemas: Dynamic tree memory representation for llms. In *The Thirteenth International Conference on Learning Representations*, 2025b.

Alsu Sagirova, Yuri Kuratov, and Mikhail Burtsev. Srmt: shared memory for multi-agent lifelong pathfinding. *arXiv preprint arXiv:2501.13200*, 2025.

Gobinda Saha, Isha Garg, and Kaushik Roy. Gradient projection memory for continual learning. In *International Conference on Learning Representations*, 2021.

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.

Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. Lamp: When large language models meet personalization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7370–7392, 2024.

Ankur Samanta, Akshayaa Magesh, Youliang Yu, Runzhe Wu, Ayush Jain, Daniel Jiang, Boris Vidolov, Paul Sajda, Yonathan Efroni, and Kaveh Hassani. Internalizing self-consistency in language models: Multi-agent consensus alignment. *arXiv preprint arXiv:2509.15172*, 2025.

Vinay Samuel, Henry Peng Zou, Yue Zhou, Shreyas Chaudhari, Ashwin Kalyan, Tanmay Rajpurohit, Ameet Deshpande, Karthik Narasimhan, and Vishvak Murahari. Personagym: Evaluating persona agents and llms. *arXiv preprint arXiv:2407.18416*, 2024.

Samarth Sarin, Lovepreet Singh, Bhaskarjit Sarmah, and Dhagash Mehta. Memoria: A scalable agentic memory framework for personalized conversational ai. *arXiv preprint arXiv:2512.12686*, 2025.

Parth Sarthi, Salman Abdullah, Aditi Tuli, Shubh Khanna, Anna Goldie, and Christopher D Manning. Raptor: Recursive abstractive processing for tree-organized retrieval. In *The Twelfth International Conference on Learning Representations*, 2024.

Nikolay Savinov, Anton Raichuk, Damien Vincent, Raphael Marinier, Marc Pollefeys, Timothy Lillicrap, and Sylvain Gelly. Episodic curiosity through reachability. In *International Conference on Learning Representations*, 2019.

Rohit Saxena, Hao Tang, and Frank Keller. End-to-end long document summarization using gradient caching. *Transactions of the Association for Computational Linguistics*, 13:1271–1297, 2025.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems*, 36:68539–68551, 2023.

Samuel Schmidgall, Yusheng Su, Ze Wang, Ximeng Sun, Jialian Wu, Xiaodong Yu, Jiang Liu, Zicheng Liu, and Emad Barsoum. Agent laboratory: Using llm agents as research assistants. *arXiv preprint arXiv:2501.04227*, 2025.

Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, et al. Mastering atari, go, chess and shogi by planning with a learned model. *Nature*, 588(7839):604–609, 2020.

Joan Serra, Didac Suris, Marius Miron, and Alexandros Karatzoglou. Overcoming catastrophic forgetting with hard attention to the task. In *International conference on machine learning*, pp. 4548–4557. PMLR, 2018.

Lior Shani, Aviv Rosenberg, Asaf Cassel, Oran Lang, Daniele Calandriello, Avital Zipori, Hila Noga, Orgad Keller, Bilal Piot, Idan Szpektor, et al. Multi-turn reinforcement learning with preference human feedback. *Advances in Neural Information Processing Systems*, 37:118953–118993, 2024.

Jiaqi Shao, Yuxiang Lin, Munish Prasad Lohani, Yufeng Miao, and Bing Luo. Do llm agents know how to ground, recover, and assess? a benchmark for epistemic competence in information-seeking agents. *arXiv preprint arXiv:2509.22391*, 2025.

Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. Character-llm: A trainable agent for role-playing. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 13153–13187, 2023.

Chen Shen, Wanqing Zhang, Kehan Li, Erwen Huang, Haitao Bi, Aiying Fan, Yiwen Shen, Hongmei Dong, Ji Zhang, Yuming Shao, et al. Feat: A multi-agent forensic ai system with domain-adapted large language model for automated cause-of-death analysis. *arXiv preprint arXiv:2508.07950*, 2025.

Juanming Shi, Qinglang Guo, Yong Liao, and Shenglin Liang. Legalgpt: legal chain of thought for the legal large language model multi-agent framework. In *International Conference on Intelligent Computing*, pp. 25–37. Springer, 2024.

Yaorui Shi, Yuxin Chen, Siyuan Wang, Sihang Li, Hengxing Cai, Qi Gu, Xiang Wang, and An Zhang. Look back to reason forward: Revisitable memory for long-context llm agents. *arXiv preprint arXiv:2509.23040*, 2025a.

Zitong Shi, Guancheng Wan, Wenke Huang, Guibin Zhang, Jiawei Shao, Mang Ye, and Carl Yang. Privacy-enhancing paradigms within federated multi-agent systems. In *ICML 2025 Workshop on Collaborative and Federated Agentic Workflows*, 2025b.

Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim. Continual learning with deep generative replay. In *NIPS*, 2017.

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: 8634–8652, 2023.

Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. Alfred: A benchmark for interpreting grounded instructions for everyday tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10740–10749, 2020.

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Cote, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. Alfworld: Aligning text and embodied environments for interactive learning. In *International Conference on Learning Representations*, 2021.

David Silver, Hado Hasselt, Matteo Hessel, Tom Schaul, Arthur Guez, Tim Harley, Gabriel Dulac-Arnold, David Reichert, Neil Rabinowitz, Andre Barreto, et al. The predictron: End-to-end learning and planning. In *International Conference on Machine Learning*, pp. 3191–3199. PMLR, 2017.

Kunal Pratap Singh, Jordi Salvador, Luca Weihs, and Aniruddha Kembhavi. Scene graph contrastive learning for embodied navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10884–10894, 2023.

George Sperling. The information available in brief visual presentations. *Psychological monographs: General and applied*, 74(11):1, 1960.

Larry R Squire. Memory and the hippocampus: a synthesis from findings with rats, monkeys, and humans. *Psychological review*, 99(2):195, 1992.

Giulio Starace, Oliver Jaffe, Dane Sherburn, James Aung, Jun Shern Chan, Leon Maksin, Rachel Dias, Evan Mays, Benjamin Kinsella, Wyatt Thompson, et al. Paperbench: Evaluating ai's ability to replicate ai research. In *Forty-second International Conference on Machine Learning*, 2025.

Theodore Sumers, Shunyu Yao, Karthik Narasimhan, and Thomas Griffiths. Cognitive architectures for language agents. *Transactions on Machine Learning Research*, 2023.

Haoran Sun and Shaoning Zeng. Hierarchical memory for high-efficiency long-term reasoning in llm agents. *arXiv preprint arXiv:2507.22925*, 2025.

Haoran Sun, Zekun Zhang, and Shaoning Zeng. Preference-aware memory update for long-term llm agents. *arXiv preprint arXiv:2510.09720*, 2025a.

Weiwei Sun, Miao Lu, Zhan Ling, Kang Liu, Xuesong Yao, Yiming Yang, and Jiecao Chen. Scaling long-horizon llm agent via context-folding. *arXiv preprint arXiv:2510.11967*, 2025b.

Mirac Suzgun, Mert Yuksekgonul, Federico Bianchi, Dan Jurafsky, and James Zou. Dynamic cheatsheet: Test-time learning with adaptive memory. *arXiv preprint arXiv:2504.07952*, 2025.

Valentin Tablan, Scott Taylor, Gabriel Hurtado, Kristoffer Bernhem, Anders Uhrenholt, Gabriele Farei, and Karo Moilanen. Smarter together: Creating agentic communities of practice through shared experiential learning. *arXiv preprint arXiv:2511.08301*, 2025.

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. *Advances in Neural Information Processing Systems*, 37:130109–130135, 2024.

Yashar Talebirad and Amirhossein Nadiri. Multi-agent collaboration: Harnessing the power of intelligent llm agents. *arXiv preprint arXiv:2306.03314*, 2023.

Haoran Tan, Zeyu Zhang, Chen Ma, Xu Chen, Quanyu Dai, and Zhenhua Dong. Membench: Towards more comprehensive evaluation on the memory of llm-based agents. *arXiv preprint arXiv:2506.21605*, 2025a.

Xingyu Tan, Xiaoyang Wang, Qing Liu, Xiwei Xu, Xin Yuan, Liming Zhu, and Wenjie Zhang. Memotime: Memory-augmented temporal knowledge graph enhanced large language model reasoning. *arXiv preprint arXiv:2510.13614*, 2025b.

Zhaoxuan Tan, Qingkai Zeng, Yijun Tian, Zheyuan Liu, Bing Yin, and Meng Jiang. Democratizing large language models via personalized parameter-efficient fine-tuning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 6476–6491, 2024a.

Zhen Tan, Chengshuai Zhao, Raha Moraffah, Yifan Li, Song Wang, Jundong Li, Tianlong Chen, and Huan Liu. Glue pizza and eat rocks-exploiting vulnerabilities in retrieval-augmented generative models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 1610–1626, 2024b.

Zhen Tan, Jun Yan, I-Hung Hsu, Rujun Han, Zifeng Wang, Long Le, Yiwen Song, Yanfei Chen, Hamid Palangi, George Lee, et al. In prospect and retrospect: Reflective memory management for long-term personalized dialogue agents. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8416–8439, 2025c.

Jing Tang, Hongru Xiao, Xiang Li, Wei Wang, and Zeyu Gong. Chatcad: An mllm-guided framework for zero-shot cad drawing restoration. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025a.

Xiangru Tang, Anni Zou, Zhuosheng Zhang, Ziming Li, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. Medagents: Large language models as collaborators for zero-shot medical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 599–621, 2024.

Xiangru Tang, Tianrui Qin, Tianhao Peng, Ziyang Zhou, Daniel Shao, Tingting Du, Xinming Wei, Peng Xia, Fang Wu, He Zhu, et al. Agent kb: Leveraging cross-domain experience for agentic problem solving. *arXiv preprint arXiv:2507.06229*, 2025b.

Mohammad Tavakoli, Alireza Salemi, Carrie Ye, Mohamed Abdalla, Hamed Zamani, and J Ross Mitchell. Beyond a million tokens: Benchmarking and enhancing long-term memory in llms. *arXiv preprint arXiv:2510.27246*, 2025.

NovelSeek Team, Bo Zhang, Shiyang Feng, Xiangchao Yan, Jiakang Yuan, Zhiyin Yu, Xiaohan He, Songtao Huang, Shaowei Hou, Zheng Nie, et al. Novelseek: When agent becomes the scientist–building closed-loop system from hypothesis to verification. *arXiv preprint arXiv:2505.16938*, 2025.

Alessandra Terranova, Björn Ross, and Alexandra Birch. Evaluating long-term memory for long-context question answering. *arXiv preprint arXiv:2510.23730*, 2025.

Ao Tian, Yunfeng Lu, Xinxin Fan, Changhao Wang, Lanzhi Zhou, Yeyao Zhang, and Yanfang Liu. Rgmem: Renormalization group-based memory evolution for language agent user profile. *arXiv preprint arXiv:2510.16392*, 2025.

Yu Tian, Xiao Yang, Jingyuan Zhang, Yinpeng Dong, and Hang Su. Evil geniuses: Delving into the safety of llm-based agents. *arXiv preprint arXiv:2311.11855*, 2023.

Susumu Tonegawa, Michele Pignatelli, Dheeraj S Roy, and Tomás J Ryan. Memory engram storage and retrieval. *Current opinion in neurobiology*, 35:101–109, 2015.

Linh Tran, Wei Sun, Stacy Patterson, and Ana Milanova. Privacy-preserving personalized federated prompt learning for multimodal large language models. In *International Conference on Learning Representations*, 2025.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. Musique: Multihop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554, 2022.

Harsh Trivedi, Tushar Khot, Mareike Hartmann, Ruskin Manku, Vinty Dong, Edward Li, Shashank Gupta, Ashish Sabharwal, and Niranjan Balasubramanian. Appworld: A controllable world of apps and people for benchmarking interactive coding agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16022–16076, 2024.

Endel Tulving. Episodic and semantic memory. In Endel Tulving and Wayne Donaldson (eds.), *Organization of Memory*, pp. 381–403. Academic Press, 1972.

Endel Tulving. Elements of episodic memory. *Oxford University Press*, 1983.

Endel Tulving. Memory and consciousness. *Canadian Psychology/Psychologie canadienne*, 26(1):1, 1985.

Endel Tulving. Episodic memory: From mind to brain. *Annual review of psychology*, 53(1):1–25, 2002.

Szymon Tworkowski, Konrad Staniszewski, Mikołaj Pacek, Yuhuai Wu, Henryk Michalewski, and Piotr Miłoś. Focused transformer: Contrastive training for context scaling. *Advances in neural information processing systems*, 36:42661–42688, 2023.

Georgios Tziafas and Hamidreza Kasaei. Lifelong robot library learning: Bootstrapping composable and generalizable skills for embodied control with language models. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 515–522. IEEE, 2024.

Nidadavolu Venkat Durga Sai Siva Vara, Prasad Raju, Nuruzzaman Faruqui, Nikhil Patel, Olivia-Roxana Alecsoiu, Priyabrata Thatoi, Salem A Alyami, and AKM Azad. Legalmind: Agentic ai-driven process optimization and cost reduction in legal services using deepseek. *IEEE Access*, 2025.

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.

S Velliangiri, V Anbarasu, P Karthikeyan, and SP Anandaraj. Intelligent personal health monitoring and guidance using long short-term memory. *Journal of Mobile Multimedia*, 18(2):349–371, 2022.

Ao Wang, Hui Chen, Jiaxin Li, Jianchao Tan, Kefeng Zhang, Xunliang Cai, Zijia Lin, Jungong Han, and Guiguang Ding. Prefixkv: Adaptive prefix kv cache is what vision instruction-following models need for efficient generation. *arXiv preprint arXiv:2412.03409*, 2024a.

Bing Wang, Xinnian Liang, Jian Yang, Hui Huang, Shuangzhi Wu, Peihao Wu, Lu Lu, Zejun Ma, and Zhoujun Li. Scm: Enhancing large language model with self-controlled memory framework. *arXiv e-prints*, pp. arXiv–2304, 2023a.

Bo Wang, Weiyi He, Shenglai Zeng, Zhen Xiang, Yue Xing, Jiliang Tang, and Pengfei He. Unveiling privacy risks in llm agent memory. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 25241–25260, 2025a.

Bo Wang, Jiehong Lin, Chenzhi Liu, Xinting Hu, Yifei Yu, Tianjia Liu, Zhongrui Wang, and Xiaojuan Qi. Mg-nav: Dual-scale visual navigation via sparse spatial memory. *arXiv preprint arXiv:2511.22609*, 2025b.

Danqing Wang, Kevin Yang, Hanlin Zhu, Xiaomeng Yang, Andrew Cohen, Lei Li, and Yuandong Tian. Learning personalized alignment for evaluating open-ended text generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 13274–13292, 2024b.

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. *Transactions on Machine Learning Research*, 2025c.

Hao Wang, Jialun Zhong, Changcheng Wang, Zhujun Nie, Zheng Li, Shunyu Yao, Yanzeng Li, and Xinchi Li. Seal: Self-evolving agentic learning for conversational question answering over knowledge graphs. *arXiv preprint arXiv:2512.04868*, 2025d.

Haoming Wang, Haoyang Zou, Huatong Song, Jiazhan Feng, Junjie Fang, Junting Lu, Longxiang Liu, Qinyu Luo, Shihao Liang, Shijue Huang, et al. Ui-tars-2 technical report: Advancing gui agent with multi-turn reinforcement learning. *arXiv preprint arXiv:2509.02544*, 2025e.

Jiayu Wang, Yifei Ming, Riya Dulepet, Qinglin Chen, Austin Xu, Zixuan Ke, Frederic Sala, Aws Albarghouthi, Caiming Xiong, and Shafiq Joty. Liveresearchbench: A live benchmark for user-centric deep research in the wild. *arXiv preprint arXiv:2510.14240*, 2025f.

Junjian Wang, Lidan Zhao, and Xi Sheryl Zhang. Madra: Multi-agent debate for risk-aware embodied planning. *arXiv preprint arXiv:2511.21460*, 2025g.

Junyang Wang, Haiyang Xu, Haitao Jia, Xi Zhang, Ming Yan, Weizhou Shen, Ji Zhang, Fei Huang, and Jitao Sang. Mobile-agent-v2: Mobile device operation assistant with effective navigation via multi-agent collaboration. *Advances in Neural Information Processing Systems*, 37:2686–2710, 2024c.

Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. *arXiv preprint arXiv:2305.04091*, 2023b.

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, 2024d.

Lei Wang, Jingsen Zhang, Hao Yang, Zhi-Yuan Chen, Jiakai Tang, Zeyu Zhang, Xu Chen, Yankai Lin, Hao Sun, Ruihua Song, et al. User behavior simulation with large language model-based agents. *ACM Transactions on Information Systems*, 43(2):1–37, 2025h.

Noah Wang, Zy Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, et al. Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 14743–14777, 2024e.

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. *Advances in Neural Information Processing Systems*, 37:53764–53797, 2024f.

Piaohong Wang, Motong Tian, Jiaxian Li, Yuan Liang, Yuqing Wang, Qianben Chen, Tiannan Wang, Zhicong Lu, Jiawei Ma, Yuchen Eleanor Jiang, et al. O-mem: Omni memory system for personalized, long horizon, self-evolving agents. *arXiv e-prints*, pp. arXiv–2511, 2025i.

Qingyue Wang, Yanhe Fu, Yanan Cao, Shuai Wang, Zhiliang Tian, and Liang Ding. Recursively summarizing enables long-term dialogue memory in large language models. *Neurocomputing*, 639:130193, 2025j.

Rui Wang, Yonghe Chen, Weiyu Zhang, Jiasheng Si, Hongjiao Guan, Xueping Peng, and Wenpeng Lu. Medconma: A confidence-driven multi-agent framework for medical q&a. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 421–433. Springer, 2025k.

Ruoyao Wang, Peter Jansen, Marc-Alexandre Côté, and Prithviraj Ammanabrolu. Scienceworld: Is your agent smarter than a 5th grader? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 11279–11298, 2022.

Saizhuo Wang, Hang Yuan, Lionel M Ni, and Jian Guo. Quantagent: Seeking holy grail in trading by self-improving large language model. *arXiv preprint arXiv:2402.03755*, 2024g.

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, 2024h.

Tianfu Wang, Yi Zhan, Jianxun Lian, Zhengyu Hu, Nicholas Jing Yuan, Qi Zhang, Xing Xie, and Hui Xiong. Llm-powered multi-agent framework for goal-oriented learning in intelligent tutoring system. In *Companion Proceedings of the ACM on Web Conference 2025*, pp. 510–519, 2025l.

Tiannan Wang, Meiling Tao, Ruoyu Fang, Huilin Wang, Shuai Wang, Yuchen Eleanor Jiang, and Wangchunshu Zhou. Ai persona: Towards life-long personalization of llms. *arXiv preprint arXiv:2412.13103*, 2024i.

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, 2023c.

Xiaoqiang Wang, Suyuchen Wang, and Yun Zhu. R^3Mem: Bridging memory retention and retrieval via reversible compression. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 4541–4557, Vienna, Austria, July 2025m. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.235.

Xingyao Wang, Boxuan Li, Yufan Song, Frank F Xu, Xiangru Tang, Mingchen Zhuge, Jiayi Pan, Yueqi Song, Bowen Li, Jaskirat Singh, et al. Openhands: An open platform for ai software developers as generalist agents. In *The Thirteenth International Conference on Learning Representations*, 2024j.

Xingyao Wang, Zihan Wang, Jiateng Liu, Yangyi Chen, Lifan Yuan, Hao Peng, and Heng Ji. MINT: Evaluating LLMs in multi-turn interaction with tools and language feedback. In *The Twelfth International Conference on Learning Representations*, 2024k.

Yancheng Wang, Ziyan Jiang, Zheng Chen, Fan Yang, Yingxue Zhou, Eunah Cho, Xing Fan, Yanbin Lu, Xiaojiang Huang, and Yingzhen Yang. Recmind: Large language model powered agent for recommendation. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 4351–4364, 2024l.

Yu Wang and Xi Chen. Mirix: Multi-agent memory system for llm-based agents. *arXiv preprint arXiv:2507.07957*, 2025.

Yu Wang, Yifan Gao, Xiusi Chen, Haoming Jiang, Shiyang Li, Jingfeng Yang, Qingyu Yin, Zheng Li, Xian Li, Bing Yin, et al. Memoryllm: towards self-updatable large language models. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 50453–50466, 2024m.

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*, 2025n.

Yu Wang, Xinshuang Liu, Xiusi Chen, Sean O'Brien, Junda Wu, and Julian McAuley. Self-updatable large language models by integrating context into model parameters. In *The Thirteenth International Conference on Learning Representations*, 2025o.

Yu Wang, Ryuichi Takanobu, Zhiqi Liang, Yuzhen Mao, Yuanzhe Hu, Julian McAuley, and Xiaojian Wu. Mem-α: Learning memory construction via reinforcement learning. *arXiv preprint arXiv:2509.25911*, 2025p.

Yu Wang, Ruihan Wu, Zexue He, Xiusi Chen, and Julian McAuley. Large scale knowledge washing. In *The Thirteenth International Conference on Learning Representations*, 2025q.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37:95266–95290, 2024n.

Zhao Wang, Sota Moriyama, Wei-Yao Wang, Briti Gangopadhyay, and Shingo Takamatsu. Talk structurally, act hierarchically: A collaborative framework for llm multi-agent systems. *arXiv preprint arXiv:2502.11098*, 2025r.

Zheng Wang, Zhongyang Li, Zeren Jiang, Dandan Tu, and Wei Shi. Crafting personalized agents through retrieval-augmented generation on editable memory graphs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 4891–4906, 2024o.

Zhiruo Wang, Daniel Fried, and Graham Neubig. Trove: Inducing verifiable and efficient toolboxes for solving programmatic tasks. *arXiv preprint arXiv:2401.12869*, 2024p.

Zihan Wang, Xiangyang Li, Jiahao Yang, Yeqi Liu, and Shuqiang Jiang. Gridmm: Grid memory map for vision-and-language navigation. In *Proceedings of the IEEE/CVF International conference on computer vision*, pp. 15625–15636, 2023d.

Zihao Wang, Shaofei Cai, Anji Liu, Yonggang Jin, Jinbing Hou, Bowei Zhang, Haowei Lin, Zhaofeng He, Zilong Zheng, Yaodong Yang, et al. Jarvis-1: Open-world multi-task agents with memory-augmented multimodal language models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024q.

Zixuan Wang, Bo Yu, Junzhe Zhao, Wenhao Sun, Sai Hou, Shuai Liang, Xing Hu, Yinhe Han, and Yiming Gan. Karma: Augmenting embodied ai agents with long-and-short term memory systems. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1–8. IEEE, 2025s.

Zora Zhiruo Wang, Apurva Gandhi, Graham Neubig, and Daniel Fried. Inducing programmatic skills for agentic tasks. *arXiv preprint arXiv:2504.06821*, 2025t.

Zora Zhiruo Wang, Jiayuan Mao, Daniel Fried, and Graham Neubig. Agent workflow memory. In *Forty-second International Conference on Machine Learning*, 2025u.

Hao Wei, Jianing Qiu, Haibao Yu, and Wu Yuan. Medco: Medical education copilots based on a multi-agent framework. In *European Conference on Computer Vision*, pp. 119–135. Springer, 2024.

Jason Wei, Zhiqing Sun, Spencer Papay, Scott McKinney, Jeffrey Han, Isa Fulford, Hyung Won Chung, Alex Tachard Passos, William Fedus, and Amelia Glaese. Browsecomp: A simple yet challenging benchmark for browsing agents. *arXiv preprint arXiv:2504.12516*, 2025a.

Qianshan Wei, Tengchao Yang, Yaochen Wang, Xinfeng Li, Lijun Li, Zhenfei Yin, Yi Zhan, Thorsten Holz, Zhiqiang Lin, and XiaoFeng Wang. A-memguard: A proactive defense framework for llm-based agent memory. *arXiv preprint arXiv:2510.02373*, 2025b.

Rubin Wei, Jiaqi Cao, Jiarui Wang, Jushi Kai, Qipeng Guo, Bowen Zhou, and Zhouhan Lin. Mlp memory: A retriever-pretrained memory for large language models. *arXiv preprint arXiv:2508.01832*, 2025c.

Tianxin Wei, Noveen Sachdeva, Benjamin Coleman, Zhankui He, Yuanchen Bei, Xuying Ning, Mengting Ai, Yunzhe Li, Jingrui He, Ed H Chi, et al. Evo-memory: Benchmarking llm agent test-time learning with self-evolving memory. *arXiv preprint arXiv:2511.20857*, 2025d.

Zhepei Wei, Wenlin Yao, Yao Liu, Weizhi Zhang, Qin Lu, Liang Qiu, Changlong Yu, Puyang Xu, Chao Zhang, Bing Yin, et al. Webagent-r1: Training web agents via end-to-end multi-turn reinforcement learning. *arXiv preprint arXiv:2505.16421*, 2025e.

Hannes Westermann. Dallma: Semi-structured legal reasoning and drafting with large language models. In *2nd Workshop on Generative AI and Law*, 2024.

Rebecca Westhäußer, Wolfgang Minker, and Sebatian Zepf. Enabling personalized long-term interactions in llm-based agents through persistent memory and user profiles. *arXiv preprint arXiv:2510.07925*, 2025.

Schaun Wheeler and Olivier Jeunen. Procedural memory is not all you need: Bridging cognitive gaps in llm-based agents. In *Adjunct Proceedings of the 33rd ACM Conference on User Modeling, Adaptation and Personalization*, pp. 360–364, 2025.

Bowen Wu, Wenqing Wang, Lihaoran Lihaoran, Yunhan Deng, Ying Li, Jingsong Yu, and Baoxun Wang. Interpersonal memory matters: A new task for proactive dialogue utilizing conversational history. In *Proceedings of the 29th Conference on Computational Natural Language Learning*, pp. 47–67, 2025a.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. Longmemeval: Benchmarking chat assistants on long-term interactive memory. In *The Thirteenth International Conference on Learning Representations*, 2025b.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. *Advances in Neural Information Processing Systems*, 37:28828–28857, 2024a.

Junde Wu, Jiayuan Zhu, Yuyuan Liu, Min Xu, and Yueming Jin. Agentic reasoning: A streamlined framework for enhancing llm reasoning with agentic tools. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 28489–28503, 2025c.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, et al. Autogen: Enabling next-gen llm applications via multi-agent conversations. In *First Conference on Language Modeling*, 2024b.

Shanglin Wu and Kai Shu. Memory in llm-based multi-agent systems: Mechanisms, challenges, and collective intelligence. *Authorea Preprints*, 2025.

Shiwei Wu, Joya Chen, Kevin Qinghong Lin, Qimeng Wang, Yan Gao, Qianli Xu, Tong Xu, Yao Hu, Enhong Chen, and Mike Zheng Shou. Videollm-mod: Efficient video-language streaming with mixture-of-depths vision computation. *Advances in Neural Information Processing Systems*, 37:109922–109947, 2024c.

Xixi Wu, Kuan Li, Yida Zhao, Liwen Zhang, Litu Ou, Huifeng Yin, Zhongwang Zhang, Xinmiao Yu, Dingchu Zhang, Yong Jiang, et al. Resum: Unlocking long-horizon search intelligence via context summarization. *arXiv preprint arXiv:2509.13313*, 2025d.

Yaozu Wu, Jizhou Guo, Dongyuan Li, Henry Peng Zou, Wei-Chieh Huang, Yankai Chen, Zhen Wang, Weizhi Zhang, Yangning Li, Meng Zhang, et al. Psg-agent: Personality-aware safety guardrail for llm-based agents. *arXiv preprint arXiv:2509.23614*, 2025e.

Yaozu Wu, Dongyuan Li, Yankai Chen, Renhe Jiang, Henry Peng Zou, Wei-Chieh Huang, Yangning Li, Liancheng Fang, Zhen Wang, and Philip S Yu. Multi-agent autonomous driving systems with large language models: A survey of recent advances, resources, and future directions. *Findings of the Association for Computational Linguistics: EMNLP*, 2025, 2025f.

Yaxiong Wu, Sheng Liang, Chen Zhang, Yichao Wang, Yongyue Zhang, Huifeng Guo, Ruiming Tang, and Yong Liu. From human memory to ai memory: A survey on memory mechanisms in the era of llms. *arXiv preprint arXiv:2504.15965*, 2025g.

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*, pp. 2585–2595, 2024.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, et al. The rise and potential of large language model based agents: A survey. *Science China Information Sciences*, 68(2):121101, 2025a.

Zhiheng Xi, Jixuan Huang, Chenyang Liao, Baodai Huang, Honglin Guo, Jiaqi Liu, Rui Zheng, Junjie Ye, Jiazheng Zhang, Wenxiang Chen, et al. Agentgym-rl: Training llm agents for long-horizon decision making through multi-turn reinforcement learning. *arXiv preprint arXiv:2509.08755*, 2025b.

Siyu Xia, Zekun Xu, Jiajun Chai, Wentian Fan, Yan Song, Xiaohan Wang, Guojun Yin, Wei Lin, Haifeng Zhang, and Jun Wang. From experience to strategy: Empowering llm agents with trainable graph memory. *arXiv preprint arXiv:2511.07800*, 2025.

Yijia Xiao, Edward Sun, Di Luo, and Wei Wang. Tradingagents: Multi-agents llm financial trading framework. *arXiv preprint arXiv:2412.20138*, 2024.

Yunzhong Xiao, Yangmin Li, Hewei Wang, Yunlong Tang, and Zora Zhiruo Wang. Toolmem: Enhancing multimodal agents with learnable tool capability memory. *arXiv preprint arXiv:2510.06664*, 2025.

Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh J Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, et al. Osworld: Benchmarking multimodal agents for open-ended tasks in real computer environments. *Advances in Neural Information Processing Systems*, 37:52040–52094, 2024.

Guojun Xiong, Zhiyang Deng, Keyi Wang, Yupeng Cao, Haohang Li, Yangyang Yu, Xueqing Peng, Mingquan Lin, Kaleb E Smith, Xiao-Yang Liu, et al. Flag-trader: Fusion llm-agent with gradient-based reinforcement learning for financial trading. *arXiv preprint arXiv:2502.11433*, 2025a.

Haomiao Xiong, Zongxin Yang, Jiazuo Yu, Yunzhi Zhuge, Lu Zhang, Jiawen Zhu, and Huchuan Lu. Streaming video understanding and multi-round interaction with memory-enhanced knowledge. *arXiv preprint arXiv:2501.13468*, 2025b.

Yiming Xiong, Jian Wang, Bing Li, Yuhan Zhu, and Yuqi Zhao. Self-organizing agent network for llm-based workflow automation. *arXiv preprint arXiv:2508.13732*, 2025c.

Zidi Xiong, Yuping Lin, Wenya Xie, Pengfei He, Zirui Liu, Jiliang Tang, Himabindu Lakkaraju, and Zhen Xiang. How memory management impacts llm agents: An empirical study of experience-following behavior. *arXiv preprint arXiv:2505.16067*, 2025d.

Haoran Xu, Jiacong Hu, Ke Zhang, Lei Yu, Yuxin Tang, Xinyuan Song, Yiqun Duan, Lynn Ai, and Bill Shi. Sedm: Scalable self-evolving distributed memory for agents. *arXiv preprint arXiv:2509.09498*, 2025a.

Jing Xu, Arthur Szlam, and Jason Weston. Beyond goldfish memory: Long-term open-domain conversation. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)*, pp. 5180–5197, 2022a.

Songlin Xu, Xinyu Zhang, and Lianhui Qin. Eduagent: Generative student agents in learning. *arXiv preprint arXiv:2404.07963*, 2024a.

Songlin Xu, Hao-Ning Wen, Hongyi Pan, Dallas Dominguez, Dongyin Hu, and Xinyu Zhang. Classroom simulacra: Building contextual student generative agents in online education for learning behavioral simulation. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pp. 1–26, 2025b.

Tianqi Xu, Linyao Chen, Dai-Jie Wu, Yanjun Chen, Zecheng Zhang, Xiang Yao, Zhiqiang Xie, Yongchao Chen, Shilong Liu, Bochen Qian, et al. Crab: Cross-environment agent benchmark for multimodal language model agents. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 21607–21647, 2025c.

Tianyang Xu, Hongqiu Wu, Weiqi Wu, and Hai Zhao. Open-theatre: An open-source toolkit for llm-based interactive drama. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 453–460, 2025d.

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*, 2025e.

Xinchao Xu, Zhibin Gou, Wenquan Wu, Zheng-Yu Niu, Hua Wu, Haifeng Wang, and Shihang Wang. Long time no see! open-domain conversation with long-term persona memory. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2639–2650, 2022b.

Yunqi Xu, Tianchi Cai, Jiyan Jiang, and Xierui Song. Face4rag: Factual consistency evaluation for retrieval augmented generation in chinese. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 6083–6094, 2024b.

Karmesh Yadav, Yusuf Ali, Gunshi Gupta, Yarin Gal, and Zsolt Kira. Findingdory: A benchmark to evaluate memory in embodied agents. *arXiv preprint arXiv:2506.15635*, 2025.

Ofir Yakobi and Shir Sadon. Bringing memory to ai: Mcp, a2a & agent context protocols. `https://orca.security/resources/blog/bringing-memory-to-ai-mcp-a2a-agent-context-protocols/`, 2025. Orca Security blog.

BY Yan, Chaofan Li, Hongjin Qian, Shuqi Lu, and Zheng Liu. General agentic memory via deep research. *arXiv preprint arXiv:2511.18423*, 2025a.

Sikuan Yan, Xiufeng Yang, Zuchao Huang, Ercong Nie, Zifeng Ding, Zonggen Li, Xiaowen Ma, Kristian Kersting, Jeff Z Pan, Hinrich Schütze, et al. Memory-r1: Enhancing large language model agents to manage and utilize memories via reinforcement learning. *arXiv preprint arXiv:2508.19828*, 2025b.

Cheng Yang, Xuemeng Yang, Licheng Wen, Daocheng Fu, Jianbiao Mei, Rong Wu, Pinlong Cai, Yufan Shen, Nianchen Deng, Botian Shi, et al. Learning on the job: An experience-driven self-evolving agent for long-horizon tasks. *arXiv preprint arXiv:2510.08002*, 2025a.

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*, 2024a.

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. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 28885–28900, 2025b.

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. *Advances in Neural Information Processing Systems*, 37:113519–113544, 2024b.

Rui Yang. Casegpt: a case reasoning framework based on language models and retrieval-augmented generation. *arXiv preprint arXiv:2407.07913*, 2024.

Rui Yang, Hanyang Chen, Junyu Zhang, Mark Zhao, Cheng Qian, Kangrui Wang, Qineng Wang, Teja Venkat Koripella, Marziyeh Movahedi, Manling Li, et al. Embodiedbench: Comprehensive benchmarking multi-modal large language models for vision-driven embodied agents. *arXiv preprint arXiv:2502.09560*, 2025c.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pp. 2369–2380, 2018.

Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. Webshop: Towards scalable real-world web interaction with grounded language agents. *Advances in Neural Information Processing Systems*, 35: 20744–20757, 2022.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *The eleventh international conference on learning representations*, 2023.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik R. Narasimhan. τ-bench: A benchmark for Tool-Agent-User interaction in real-world domains. In *The Thirteenth International Conference on Learning Representations*, 2025.

Weiran Yao, Shelby Heinecke, Juan Carlos Niebles, Zhiwei Liu, Yihao Feng, Le Xue, Rithesh RN, Zeyuan Chen, Jianguo Zhang, Devansh Arpit, et al. Retroformer: Retrospective large language agents with policy gradient optimization. In *The Twelfth International Conference on Learning Representations*, 2024.

Jiabo Ye, Xi Zhang, Haiyang Xu, Haowei Liu, Junyang Wang, Zhaoqing Zhu, Ziwei Zheng, Feiyu Gao, Junjie Cao, Zhengxi Lu, et al. Mobile-agent-v3: Fundamental agents for gui automation. *arXiv preprint arXiv:2508.15144*, 2025a.

Rui Ye, Zhongwang Zhang, Kuan Li, Huifeng Yin, Zhengwei Tao, Yida Zhao, Liangcai Su, Liwen Zhang, Zile Qiao, Xinyu Wang, et al. Agentfold: Long-horizon web agents with proactive context management. *arXiv preprint arXiv:2510.24699*, 2025b.

Asaf Yehudai, Lilach Eden, Alan Li, Guy Uziel, Yilun Zhao, Roy Bar-Haim, Arman Cohan, and Michal Shmueli-Scheuer. Survey on evaluation of llm-based agents. *arXiv preprint arXiv:2503.16416*, 2025.

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*, pp. 1–22, 2024.

Jeong Hun Yeo, Sangyun Chung, Sungjune Park, Dae Hoe Kim, Jinyoung Moon, and Yong Man Ro. Gcagent: Long-video understanding via schematic and narrative episodic memory. *arXiv preprint arXiv:2511.12027*, 2025a.

Woongyeong Yeo, Kangsan Kim, Jaehong Yoon, and Sung Ju Hwang. Worldmm: Dynamic multimodal memory agent for long video reasoning. *arXiv preprint arXiv:2512.02425*, 2025b.

Fangyi Yu, Nabeel Seedat, Drahomira Herrmannova, Frank Schilder, and Jonathan Richard Schwarz. Beyond pointwise scores: Decomposed criteria-based evaluation of llm responses. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 1931–1954, 2025a.

Hongli Yu, Tinghong Chen, Jiangtao Feng, Jiangjie Chen, Weinan Dai, Qiying Yu, Ya-Qin Zhang, Wei-Ying Ma, Jingjing Liu, Mingxuan Wang, et al. Memagent: Reshaping long-context llm with multi-conv rl-based memory agent. *arXiv preprint arXiv:2507.02259*, 2025b.

Miao Yu, Fanci Meng, Xinyun Zhou, Shilong Wang, Junyuan Mao, Linsey Pan, Tianlong Chen, Kun Wang, Xinfeng Li, Yongfeng Zhang, et al. A survey on trustworthy llm agents: Threats and countermeasures. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 6216–6226, 2025c.

Shuo Yu, Mingyue Cheng, Daoyu Wang, Qi Liu, Zirui Liu, Ze Guo, and Xiaoyu Tao. Memweaver: A hierarchical memory from textual interactive behaviors for personalized generation. *arXiv preprint arXiv:2510.07713*, 2025d.

Yangyang Yu, Zhiyuan Yao, Haohang Li, Zhiyang Deng, Yuechen Jiang, Yupeng Cao, Zhi Chen, Jordan Suchow, Zhenyu Cui, Rong Liu, et al. Fincon: A synthesized llm multi-agent system with conceptual verbal reinforcement for enhanced financial decision making. *Advances in Neural Information Processing Systems*, 37:137010–137045, 2024a.

Yangyang Yu, Haohang Li, Zhi Chen, Yuechen Jiang, Yang Li, Jordan W Suchow, Denghui Zhang, and Khaldoun Khashanah. Finmem: A performance-enhanced llm trading agent with layered memory and character design. *IEEE Transactions on Big Data*, 2025e.

Zhuohao Yu, Chang Gao, Wenjin Yao, Yidong Wang, Wei Ye, Jindong Wang, Xing Xie, Yue Zhang, and Shikun Zhang. Kieval: A knowledge-grounded interactive evaluation framework for large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5967–5985, 2024b.

Qianhao Yuan, Jie Lou, Zichao Li, Jiawei Chen, Yaojie Lu, Hongyu Lin, Le Sun, Debing Zhang, and Xianpei Han. Memsearcher: Training llms to reason, search and manage memory via end-to-end reinforcement learning. *arXiv preprint arXiv:2511.02805*, 2025a.

Ruifeng Yuan, Shichao Sun, Yongqi Li, Zili Wang, Ziqiang Cao, and Wenjie Li. Personalized large language model assistant with evolving conditional memory. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 3764–3777, 2025b.

Siyu Yuan, Kaitao Song, Jiangjie Chen, Xu Tan, Yongliang Shen, Kan Ren, Dongsheng Li, and Deqing Yang. Easytool: Enhancing llm-based agents with concise tool instruction. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 951–972, 2025c.

Tongxin Yuan, Zhiwei He, Lingzhong Dong, Yiming Wang, Ruijie Zhao, Tian Xia, Lizhen Xu, Binglin Zhou, Fangqi Li, Zhuosheng Zhang, et al. R-judge: Benchmarking safety risk awareness for llm agents. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 1467–1490, 2024a.

Weikang Yuan, Junjie Cao, Zhuoren Jiang, Yangyang Kang, Jun Lin, Kaisong Song, Tianqianjin Lin, Pengwei Yan, Changlong Sun, and Xiaozhong Liu. Can large language models grasp legal theories? enhance legal reasoning with insights from multi-agent collaboration. In *Findings of the association for computational linguistics: EMNLP 2024*, pp. 7577–7597, 2024b.

Murong Yue, Zhiwei Liu, Liangwei Yang, Jianguo Zhang, Zuxin Liu, Haolin Chen, Ziyu Yao, Silvio Savarese, Caiming Xiong, Shelby Heinecke, et al. Toollibgen: Scalable automatic tool creation and aggregation for llm reasoning. *arXiv preprint arXiv:2510.07768*, 2025.

Sizhe Yuen, Francisco Gomez Medina, Ting Su, Yali Du, and Adam J Sobey. Intrinsic memory agents: Heterogeneous multi-agent llm systems through structured contextual memory. *arXiv preprint arXiv:2508.08997*, 2025.

Qingbin Zeng, Bingbing Fan, Zhiyu Chen, Sijian Ren, Zhilun Zhou, Xuhua Zhang, Yuanyi Zhen, Fengli Xu, Yong Li, and Tie-Yan Liu. Mirrormind: Empowering omniscientist with the expert perspectives and collective knowledge of human scientists. *arXiv preprint arXiv:2511.16997*, 2025.

Mingliang Zhai, Zhi Gao, Yuwei Wu, and Yunde Jia. Memory-centric embodied question answer. *arXiv preprint arXiv:2505.13948*, 2025.

Xiao Zhan, Noura Abdi, William Seymour, and Jose Such. Healthcare voice ai assistants: factors influencing trust and intention to use. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW1):1–37, 2024.

Dianxing Zhang, Wendong Li, Kani Song, Jiaye Lu, Gang Li, Liuchun Yang, and Sheng Li. Memory in large language models: Mechanisms, evaluation and evolution. *arXiv preprint arXiv:2509.18868*, 2025a.

Gaoke Zhang, Bo Wang, Yunlong Ma, Dongming Zhao, and Zifei Yu. Multiple memory systems for enhancing the long-term memory of agent. *arXiv preprint arXiv:2508.15294*, 2025b.

Guibin Zhang, Muxin Fu, Guancheng Wan, Miao Yu, Kun Wang, and Shuicheng Yan. G-memory: Tracing hierarchical memory for multi-agent systems. *arXiv preprint arXiv:2506.07398*, 2025c.

Guibin Zhang, Muxin Fu, and Shuicheng Yan. Memgen: Weaving generative latent memory for self-evolving agents. *arXiv preprint arXiv:2509.24704*, 2025d.

Guibin Zhang, Yanwei Yue, Zhixun Li, Sukwon Yun, Guancheng Wan, Kun Wang, Dawei Cheng, Jeffrey Xu Yu, and Tianlong Chen. Cut the crap: An economical communication pipeline for LLM-based multi-agent systems. In *The Thirteenth International Conference on Learning Representations*, 2025e.

Guibin Zhang, Yanwei Yue, Xiangguo Sun, Guancheng Wan, Miao Yu, Junfeng Fang, Kun Wang, Tianlong Chen, and Dawei Cheng. G-designer: Architecting multi-agent communication topologies via graph neural networks. In *Forty-second International Conference on Machine Learning*, 2025f.

Hongxin Zhang, Weihua Du, Jiaming Shan, Qinhong Zhou, Yilun Du, Joshua B Tenenbaum, Tianmin Shu, and Chuang Gan. Building cooperative embodied agents modularly with large language models. In *The Twelfth International Conference on Learning Representations*, 2023a.

Jiayi Zhang, Jinyu Xiang, Zhaoyang Yu, Fengwei Teng, Xiong-Hui Chen, Jiaqi Chen, Mingchen Zhuge, Xin Cheng, Sirui Hong, Jinlin Wang, et al. Aflow: Automating agentic workflow generation. In *The Thirteenth International Conference on Learning Representations*, 2025g.

Junjie Zhang, Yupeng Hou, Ruobing Xie, Wenqi Sun, Julian McAuley, Wayne Xin Zhao, Leyu Lin, and Ji-Rong Wen. Agentcf: Collaborative learning with autonomous language agents for recommender systems. In *Proceedings of the ACM Web Conference 2024*, pp. 3679–3689, 2024a.

Jusheng Zhang, Yijia Fan, Kaitong Cai, Xiaofei Sun, and Keze Wang. OSC: Cognitive orchestration through dynamic knowledge alignment in multi-agent LLM collaboration. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 6320–6337. Association for Computational Linguistics, November 2025h.

Kai Zhang, Xiangchao Chen, Bo Liu, Tianci Xue, Zeyi Liao, Zhihan Liu, Xiyao Wang, Yuting Ning, Zhaorun Chen, Xiaohan Fu, et al. Agent learning via early experience. *arXiv preprint arXiv:2510.08558*, 2025i.

Qizheng Zhang, Changran Hu, Shubhangi Upasani, Boyuan Ma, Fenglu Hong, Vamsidhar Kamanuru, Jay Rainton, Chen Wu, Mengmeng Ji, Hanchen Li, et al. Agentic context engineering: Evolving contexts for self-improving language models. *arXiv preprint arXiv:2510.04618*, 2025j.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. Personalizing dialogue agents: I have a dog, do you have pets too? *arXiv preprint arXiv:1801.07243*, 2018.

Songyang Zhang, Zeming Li, Shipeng Yan, Xuming He, and Jian Sun. Distribution alignment: A unified framework for long-tail visual recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2361–2370, 2021.

Weizhi Zhang, Yuanchen Bei, Liangwei Yang, Henry Peng Zou, Peilin Zhou, Aiwei Liu, Yinghui Li, Hao Chen, Jianling Wang, Yu Wang, et al. Cold-start recommendation towards the era of large language models (llms): A comprehensive survey and roadmap. *arXiv preprint arXiv:2501.01945*, 2025k.

Weizhi Zhang, Yangning Li, Yuanchen Bei, Junyu Luo, Guancheng Wan, Liangwei Yang, Chenxuan Xie, Yuyao Yang, Wei-Chieh Huang, Chunyu Miao, et al. From web search towards agentic deep research: Incentivizing search with reasoning agents. *arXiv preprint arXiv:2506.18959*, 2025l.

Weizhi Zhang, Xinyang Zhang, Chenwei Zhang, Liangwei Yang, Jingbo Shang, Zhepei Wei, Henry Peng Zou, Zijie Huang, Zhengyang Wang, Yifan Gao, et al. Personaagent: When large language model agents meet personalization at test time. *arXiv preprint arXiv:2506.06254*, 2025m.

Yi Zhang, Zhongyang Yu, Wanqi Jiang, Yufeng Shen, and Jin Li. Long-term memory for large language models through topic-based vector database. In *2023 International Conference on Asian Language Processing (IALP)*, pp. 258–264. IEEE, 2023b.

Yuxiang Zhang, Jiangming Shu, Ye Ma, Xueyuan Lin, Shangxi Wu, and Jitao Sang. Memory as action: Autonomous context curation for long-horizon agentic tasks. *arXiv preprint arXiv:2510.12635*, 2025n.

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*, 2024b.

Zeyu Zhang, Quanyu Dai, Xiaohe Bo, Chen Ma, Rui Li, Xu Chen, Jieming Zhu, Zhenhua Dong, and Ji-Rong Wen. A survey on the memory mechanism of large language model-based agents. *ACM Transactions on Information Systems*, 43(6):1–47, 2025o.

Zeyu Zhang, Quanyu Dai, Rui Li, Xiaohe Bo, Xu Chen, and Zhenhua Dong. Learn to memorize: Optimizing llm-based agents with adaptive memory framework. *arXiv preprint arXiv:2508.16629*, 2025p.

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, 2023c.

Zhexin Zhang, Shiyao Cui, Yida Lu, Jingzhuo Zhou, Junxiao Yang, Hongning Wang, and Minlie Huang. Agent-safetybench: Evaluating the safety of llm agents. *arXiv preprint arXiv:2412.14470*, 2024c.

Zijian Zhang, Shuchang Liu, Ziru Liu, Rui Zhong, Qingpeng Cai, Xiangyu Zhao, Chunxu Zhang, Qidong Liu, and Peng Jiang. Llm-powered user simulator for recommender system. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 13339–13347, 2025q.

Jihao Zhao, Zhiyuan Ji, Simin Niu, Hanyu Wang, Feiyu Xiong, and Zhiyu Li. Mom: Mixtures of scenario-aware document memories for retrieval-augmented generation systems. *arXiv preprint arXiv:2510.14252*, 2025a.

Linxi Zhao, Sofian Zalouk, Christian K Belardi, Justin Lovelace, Jin Peng Zhou, Kilian Q Weinberger, Yoav Artzi, and Jennifer J Sun. Pre-training large memory language models with internal and external knowledge. *arXiv preprint arXiv:2505.15962*, 2025b.

Siyan Zhao, Mingyi Hong, Yang Liu, Devamanyu Hazarika, and Kaixiang Lin. Do llms recognize your preferences? evaluating personalized preference following in llms. In *The Thirteenth International Conference on Learning Representations*, 2025c.

Zheng Zhao, Clara Vania, Subhradeep Kayal, Naila Khan, Shay B Cohen, and Emine Yilmaz. Personalens: A benchmark for personalization evaluation in conversational ai assistants. *arXiv preprint arXiv:2506.09902*, 2025d.

Zihan Zhao, Da Ma, Lu Chen, Liangtai Sun, Zihao Li, Yi Xia, Bo Chen, Hongshen Xu, Zichen Zhu, Su Zhu, et al. Developing chemdfm as a large language foundation model for chemistry. *Cell Reports Physical Science*, 6(4), 2025e.

Junhao Zheng, Xidi Cai, Qiuke Li, Duzhen Zhang, ZhongZhi Li, Yingying Zhang, Le Song, and Qianli Ma. Lifelongagentbench: Evaluating llm agents as lifelong learners. *arXiv preprint arXiv:2505.11942*, 2025a.

Junhao Zheng, Shengjie Qiu, Chengming Shi, and Qianli Ma. Towards lifelong learning of large language models: A survey. *ACM Computing Surveys*, 57(8):1–35, 2025b.

Junhao Zheng, Chengming Shi, Xidi Cai, Qiuke Li, Duzhen Zhang, Chenxing Li, Dong Yu, and Qianli Ma. Lifelong learning of large language model based agents: A roadmap. *arXiv preprint arXiv:2501.07278*, 2025c.

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*, 2024.

Tianshi Zheng, Kelvin Kiu-Wai Tam, Newt Hue-Nam K Nguyen, Baixuan Xu, Zhaowei Wang, Jiayang Cheng, Hong Ting Tsang, Weiqi Wang, Jiaxin Bai, Tianqing Fang, et al. Newtonbench: Benchmarking generalizable scientific law discovery in llm agents. *arXiv preprint arXiv:2510.07172*, 2025d.

Xiaochen Zheng, Jinzhi Lu, and Dimitris Kiritsis. The emergence of cognitive digital twin: vision, challenges and opportunities. *International Journal of Production Research*, 60(24):7610–7632, 2022.

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, pp. 19724–19731, 2024.

Huichi Zhou, Yihang Chen, Siyuan Guo, Xue Yan, Kin Hei Lee, Zihan Wang, Ka Yiu Lee, Guchun Zhang, Kun Shao, Linyi Yang, et al. Memento: Fine-tuning llm agents without fine-tuning llms. *arXiv preprint arXiv:2508.16153*, 2025a.

Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, et al. Webarena: A realistic web environment for building autonomous agents. In *The Twelfth International Conference on Learning Representations*, 2024.

Yanfang Zhou, Xiaodong Li, Yuntao Liu, Yongqiang Zhao, Xintong Wang, Zhenyu Li, Jinlong Tian, and Xinhai Xu. M2pa: A multi-memory planning agent for open worlds inspired by cognitive theory. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 23204–23220, 2025b.

Zijian Zhou, Ao Qu, Zhaoxuan Wu, Sunghwan Kim, Alok Prakash, Daniela Rus, Jinhua Zhao, Bryan Kian Hsiang Low, and Paul Pu Liang. Mem1: Learning to synergize memory and reasoning for efficient long-horizon agents. *arXiv preprint arXiv:2506.15841*, 2025c.

Lixi Zhu, Xiaowen Huang, and Jitao Sang. How reliable is your simulator? analysis on the limitations of current llm-based user simulators for conversational recommendation. In *Companion Proceedings of the ACM Web Conference 2024*, pp. 1726–1732, 2024.

Lixi Zhu, Xiaowen Huang, and Jitao Sang. A llm-based controllable, scalable, human-involved user simulator framework for conversational recommender systems. In *Proceedings of the ACM on Web Conference 2025*, pp. 4653–4661, 2025a.

Rui-Jie Zhu, Tianhao Peng, Tianhao Cheng, Xingwei Qu, Jinfa Huang, Dawei Zhu, Hao Wang, Kaiwen Xue, Xuanliang Zhang, Yong Shan, et al. A survey on latent reasoning. *arXiv preprint arXiv:2507.06203*, 2025b.

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.

Yuanjie Zhu, Liangwei Yang, Ke Xu, Weizhi Zhang, Zihe Song, Jindong Wang, and Philip S Yu. Llm-memcluster: Empowering large language models with dynamic memory for text clustering. *arXiv preprint arXiv:2511.15424*, 2025c.

Henry Peng Zou, Zhengyao Gu, Yue Zhou, Yankai Chen, Weizhi Zhang, Liancheng Fang, Yibo Wang, Yangning Li, Kay Liu, and Philip S Yu. Testnuc: Enhancing test-time computing approaches and scaling through neighboring unlabeled data consistency. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 30750–30762, 2025a.

Henry Peng Zou, Wei-Chieh Huang, Yaozu Wu, Yankai Chen, Chunyu Miao, Hoang Nguyen, Yue Zhou, Weizhi Zhang, Liancheng Fang, Langzhou He, et al. A survey on large language model based human-agent systems. *Authorea Preprints*, 2025b.

Henry Peng Zou, Wei-Chieh Huang, Yaozu Wu, Chunyu Miao, Dongyuan Li, Aiwei Liu, Yue Zhou, Yankai Chen, Weizhi Zhang, Yangning Li, et al. A call for collaborative intelligence: Why human-agent systems should precede ai autonomy. *arXiv preprint arXiv:2506.09420*, 2025c.